Developmental Science

Developmental Science 11:3 (2008), pp 371–389 DOI: 10.1111/j.1467-7687.2008.00682.x

PAPER

Critical periods and catastrophic interference effects in the development of self-organizing feature maps

Fiona M. Richardson and Michael S.C. Thomas

Developmental Neurocognition Lab, School of Psychology, Birkbeck College, University of London, UK

Abstract

The use of self-organizing feature maps (SOFM) in models of cognitive development has frequently been associated with explanations of critical *or* sensitive periods. *By contrast, error-driven connectionist models of development have been linked with* catastrophic interference *between new knowledge and old knowledge. We introduce a set of simulations that systematically evaluate the conditions under which SOFMs demonstrate critical/sensitive periods in development versus those under which they display interference effects. We explored the relative contribution of network parameters (for example, whether learning rate and neighbourhood reduce across training), the representational resources available to the network, and the similarity between old and new knowledge in determining the functional plasticity of the maps. The SOFMs that achieved the best discrimination and topographic organization also exhibited sensitive periods in development while showing lower plasticity and hence limited interference. However, fast developing, coarser SOFMs also produced topologically organized representations, while permanently retaining their plasticity. We argue that the impact of map organization on behaviour must be interpreted in terms of the cognitive processes that the map is driving.*

Introduction

Theories of how the brain acquires knowledge are required to address the *stability–plasticity problem*, that is, how new knowledge may be incorporated into an information processing system while preserving existing knowledge (Grossberg, 1987). The stability–plasticity problem has particular importance where the individual's environment is non-stationary – that is, where the information content of experience tends to change over time. If one assesses the individual in adulthood, one can ask whether earlier experiences or later experiences were more influential in determining adult behaviour. If the earlier experiences were more important, one might refer to this as evidence of a *critical* or *sensitive period* in development. If the later experiences were more important, one might refer to this as evidence of *catastrophic interference* of new knowledge overwriting old knowledge.

The stability–plasticity problem comes to the fore in attempts to construct computational models of learning and development. For example, at least one popular computational methodology for studying development – back-propagation connectionist networks – has indicated that *catastrophic interference* may be a serious problem for the cognitive system when it attempts to acquire conceptual knowledge. Indeed, it may be such a serious problem that special processing structures are needed to overcome it (e.g. McClelland, McNaughton & O'Reilly, 1995).

In this paper, we consider the effects of a non-stationary environment on learning within an alternative neurocomputational formulism, self-organizing feature maps (Kohonen, 1995). Such maps have been employed within a range of developmental models, capturing the formation of representations within visual, sensorimotor, and language development domains (e.g. Li, Farkas & MacWhinney, 2004; McClelland, Thomas, McCandliss & Fiez, 1999; O'Reilly & Johnson, 1994; Oliver, Johnson, Karmiloff-Smith & Pennington, 2000; Westermann & Miranda, 2002, 2004). To date, and in contrast to back-propagation networks, self-organizing feature maps have been more closely associated with *critical* or *sensitive period* effects in development. However, their potential vulnerability to catastrophic inference has not been systematically explored. If these maps are a key mechanism within cognitive development, how robust are they to variations in the environment? In the following sections, we compare critical/sensitive period and catastrophic interference effects in self-organizing feature maps under conditions of a non-stationary environment. We take into account three potentially important factors that may modulate these effects: the intrinsic conditions of plasticity within the maps, the representational resources available to the system, and the relative similarity between old and new knowledge. We begin with a brief review of the empirical and computational literature relevant to the two facets of the stability–plasticity problem.

Address for correspondence: Fiona Richardson, Developmental Neurocognition Lab, School of Psychology, Birkbeck College, University of London, Malet Street, Bloomsbury, London WC1E 7HX, UK; e-mail: f.richardson@bbk.ac.uk

Catastrophic interference

For the human cognitive system, it is rare to find a total disruption or loss of previously acquired long-term knowledge as a result of learning new information. We are able to acquire new memories without forgetting old information. For example, our somatosensory cortex is able to retain and assimilate new information during motor learning without compromising the stability of previous skills (Braun, Heinz, Schweizer, Wiech, Birbaumer & Topka, 2001; Buonomano & Merzenich, 1998). Nevertheless, under some circumstances, catastrophic interference can be observed. When Mareschal, Quinn and French (2002) examined sequential category learning in 3- to 4-month-old infants, they found an asymmetric interference effect. The infants were shown a series of pictures of either cats or dogs and were able to induce the CAT or DOG category sufficiently to distinguish a novel animal from a cat or dog in a subsequent preferential looking task. When the two categories were learned sequentially, knowledge of the DOG category was preserved when the CAT category was learned after it. However, if learning of the DOG category followed learning of the CAT category, the later learning interfered with the earlier learning and knowledge of the CAT category was lost. The authors interpreted this effect in terms of catastrophic interference in a connectionist memory system; the asymmetry was taken to reflect the relative perceptual similarity structure of the two categories.

Interference effects have also been observed for more robust, long-term knowledge. Pallier, Dehaene, Poline, LeBihan, Argenti, Dupoux and Mehler (2003) examined the word recognition abilities of adults born in Korea who were adopted between the ages of 3 and 8 by French families. For these individuals, the language environment changed completely from Korean to French at the point of adoption. Behavioural tests showed that these adults had no residual knowledge of the Korean vocabulary that they knew as children. Moreover, functional brain imaging data demonstrated that their response to hearing Korean was no different from that produced by listening to other foreign languages that they had never encountered, and was the same as that found in native French speakers who had never learned Korean. Together the behavioural and imaging data are suggestive that under some circumstances, previously acquired knowledge can indeed be overwritten. Interestingly, comparison of the brain activations produced when listening to French differed between the two groups, with the Korean-born adults producing weaker activations than the French monolinguals. Interpretation of this effect is not straightforward but it does indicate that the earlier phase of Korean learning appears to have left its mark on the brains of the adopted individuals. It is possible that residual traces of prior Korean knowledge may still exist such that, should these individuals be re-exposed to Korean, they may find it easier to re-acquire the language (Pallier *et al.*, 2003).

The use of connectionist networks as models of memory has led to the extensive consideration of catastrophic interference in these systems (see French, 1999, for a review). Catastrophic interference appears to be a central feature of architectures that employ distributed representations and it is closely tied to their ability to generalize. Via superposition of knowledge over a common representational resource (the matrix of connection weights), distributed systems offer generalization for free; that is, they can extract the central tendency of a series of exemplars and use this tendency to generate responses to novel inputs. Where new knowledge conforms to the central tendency extracted from previous knowledge, learning is facilitated and new knowledge is easily accommodated (Ratcliff, 1990). Problems of catastrophic interference arise when the new knowledge is different from the old knowledge. The later learning has to use the common representational resource and overwrites previous knowledge (McCloskey & Cohen, 1989; Ratcliff, 1990).

Numerous computational solutions have been proposed in order to alleviate the catastrophic interference problem and thereby redeem connectionist models as plausible models of human memory. These approaches include the modification of the back-propagation learning rule in order to produce semi-distributed representations (Kortge, 1990; French, 1991, 1992). Alternatively, noise or 'pseudo' patterns may be used in order to extract the function learned by the network in response to early training (French & Chater, 2002). This knowledge may then be interspersed with subsequent training (Robins 1995; Robins & McCallum, 1998; Ans, Rousset, French & Musca, 2004). Essentially, catastrophic interference can be avoided in three ways: (1) use new representational resources for new knowledge; (2) use non-overlapping representational codes on the same resource ('localist' coding); and/or (3) simultaneously refresh old knowledge as new knowledge is introduced, so that the old and new knowledge can be combined within distributed representations over the same resource (called 'interleaving').

The occurrence of catastrophic interference effects in connectionist models prompted a proposal that the human cognitive system may incorporate processing structures specifically to avoid it. McClelland *et al.* (1995) suggested that human memory is split into two systems – the neocortex and the hippocampal system. The hippocampal system allows for rapid learning of new information, which is then transferred and integrated into the previous long-term knowledge stored in the neocortex. Seidenberg and Zevin (2006) argue that humans do not exhibit catastrophic interference effects because our experiences are typically interleaved. It is when we are immersed in one particular type of experience that interference may occur (as was the case in the Korean children switched to a French language environment). Moreover, in many cases, the new knowledge we are trying to learn bears some resemblance to previously acquired knowledge, reducing the scope for interference effects.

To date, the majority of simulation work exploring catastrophic interference effects has focused on error-driven learning systems such as back-propagation networks (Ans *et al.*, 2004; French, 1991, 1992, 1999; French & Chater, 2002; Kortge, 1990; McCloskey & Cohen, 1989; Ratcliff, 1990; Robins, 1995; Robins & McCallum, 1998; Sharkey & Sharkey, 1995). There has been no comparable work for self-organizing learning systems, in spite of their increasing prevalence in models of cognitive development (Li *et al.*, 2004; O'Reilly & Johnson, 1994; Oliver *et al.*, 2000; Westermann & Miranda, 2002, 2004). Given that some authors view self-organizing systems and error-driven associative systems as the two principal experience-dependent architectures within the brain (O'Reilly, 1998), this is a notable omission.

Critical periods

The notion of a *critical period* was used in the context of language acquisition by Lenneberg (1967) to refer to a limited duration in development during which children are particularly sensitive to the effects of experience. Latterly, alternative terms have been employed such as *sensitive* or *optimal* period, which are more neutral as to whether the period of plasticity comes to a complete close (see Birdsong, 2005; Johnson, 2005; Knudsen, 2004). The idea that early experiences are particularly influential and that they may even have irreversible effects on behaviour has been invoked in many examples of animal and human development, including filial imprinting in ducks and chicks, early visual development in several species, song learning in birds, and language acquisition in humans (Brainard & Doupe, 2002; Doupe & Kuhl, 1999; Hubel & Weisel, 1963; Johnson & Newport, 1989, 1991; Lorenz, 1958; Senghas, Kita & Özyürek, 2004). To take a well-known example, in second language acquisition, children are better learners than adults in terms of their ultimate proficiency (Johnson & Newport, 1989, 1991). This effect appears to be related to the age at which second language learning commences rather than degree of exposure, implicating differential contributions of early and late experiences. The exact function linking age of acquisition and ultimate attainment is still debated (e.g. Birdsong, 2005; DeKeyser & Larson-Hall, 2005).

At the neurobiological level, *neuroplasticity* is central to critical period phenomena.[1] Present data suggest that the termination of critical periods for more basic functions occurs prior to the opening of critical periods in higher-level systems (Jones, 2000). In this way, the development of low-level systems can have a lasting impact upon the opportunities for subsequent higher-level development. Although the profiles of plasticity that regulate critical periods may vary across brain systems (Uylings, 2006), there is a general trend for plasticity to decrease with increasing age (Hensch, 2004). As plasticity reduces, the ability of the system to undergo large-scale, speeded change also diminishes, thereby safeguarding existing information.

The mechanistic basis of critical periods has been studied extensively through the use of both connectionist-style error-driven and self-organizing learning systems. These models have explored early visual development (e.g. Miller, Keller & Stryker, 1989), age-of-acquisition effects in language (Lambon Ralph & Ehsan, 2006; Ellis & Lambon Ralph, 2000; Li *et al.*, 2004; Zevin & Seidenberg, 2002), and recovery after brain damage (Marchman, 1993). In error-driven connectionist networks, the privileged status of early learning has been explained with reference to the idea of *entrenchment*, where large connection weights produced by early training then compromise the ability of the network to alter its structure to accommodate new information (Ellis & Lambon Ralph, 2000; Zevin & Seidenberg, 2002; Seidenberg & Zevin, 2006). However, the prominence of catastrophic interference effects for this type of network implies that other factors – such as the similarity between old and new knowledge, the resource levels of the network, and continued training on old knowledge while new knowledge is introduced – must all play a role for early training to exert a greater influence than later training on endstate performance (Lambon Ralph & Ehsan, 2006; Thomas & Johnson, 2006).

Self-organizing feature maps fall into two camps, depending on whether their implementation involves dynamic changes to the model's parameters. Kohonen's (1982) algorithm uses two phases of training to achieve a topographic organization across the network's output layer that reflects the similarity structure of the input domain. In the *organization* or *ordering phase*, the network is trained with a high learning rate (a parameter that modulates the size of weight changes) and a large neighbourhood size (a parameter that modulates the extent of weight changes across the map in response to each input pattern). These parameter settings allow the network to achieve an initial rough organization of the appropriate topology. In the second *convergence* or *tuning phase*, the learning rate and neighbourhood size parameters are reduced to fine-tune the feature map and capture more detailed distinctions in the input set. The two phases are sometimes implemented by continuously declining functions that asymptote to non-zero values. We will refer to this configuration as the *dynamic parameter* implementation of the self-organizing feature map. Most saliently for this implementation, the functional plasticity of the system reduces *by definition*. The models will necessarily exhibit a sensitive period because this is the mechanism by which they achieve good topographic organization (Kohonen, 1995; Li *et al.*, 2004; Miikkulainen, 1997; Thomas & Richardson, 2006).

[1] In what follows, for brevity we will sometimes refer simply to 'critical period' effects, by which we intend the combined phenomenon of critical/sensitive/optimal periods in development. Debates on the distinctions between these terms are not directly relevant here, other than to note that all imply a non-linear relationship between age and functional plasticity in which there is a reduction in plasticity over time. The terms differ in the exact shape of the function and the residual level of functional plasticity when the period has closed.

Some implementations of self-organizing feature maps keep their parameters fixed across training but still report evidence for critical periods (e.g. for imprinting in chicks: O'Reilly & Johnson, 1994; for adult Japanese speakers attempting to learn the English /l/–/r/ phoneme contrast: McClelland *et al.*, 1999). In the O'Reilly and Johnson model, the critical period effect appears to be a consequence of input similarity and limited computational resources, while in the McClelland *et al.* model, it is a consequence of input similarity and assimilation in the output layer (see Thomas & Johnson, 2006, for discussion). However, both these models are characterized by highly simplified training sets in which little is demanded of the network in terms of detailed topographic organization. It is not clear that their behaviour will generalize to more complex training sets.

In sum, much more work has addressed critical periods in self-organizing feature maps than catastrophic interference, but even for critical periods the relative importance of several factors remains unclear. These include whether parameters are *dynamic* or *fixed* across training, the similarity between old and new information, and the level of resources available in the model to accommodate new information. We therefore set out to address these issues in a set of computer simulations.

Simulations

Design

Literature reporting simulations of catastrophic interference and critical period effects is typically qualitative in nature (e.g. catastrophic interference effects: French & Chater, 2002; McCloskey & Cohen, 1989; Ratcliff, 1990; Robins, 1995; Robins & McCallum, 1998; Sharkey & Sharkey, 1995; critical period effects: McClelland *et al.*, 1999; Oliver *et al.*, 2000; O'Reilly & Johnson, 1994). This form of abstract modelling is useful in these early stages of theory development in that it serves to identify the range of computational mechanisms that can generate broadly characterized empirical phenomena (Thomas, 2004; Thomas & Karmiloff-Smith, 2002). Relatively fewer computational accounts of catastrophic interference and critical period effects have the goal of simulating specific human behavioural data and even in these cases, such simulations may employ artificial or abstract training patterns and so are not necessarily quantitative in the strictest sense (catastrophic interference effects: Ans *et al.*, 2004; Mirman & Spivey, 2001; critical period effects: Ellis & Lambon-Ralph, 2001; Li *et al.*, 2004). The simulations outlined below are in keeping with the qualitative use of modelling in initial theory development, since the relative influence of critical period and particularly catastrophic interference effects in SOFMs is poorly understood. Nevertheless, the following simulations employ a cognitively constrained training environment (semantic representations with prototype/exemplar structure) rather than the arbitrary training patterns used in some simulations. The aim of future work will be to extend the principles to the simulation of quantitative data (for example, using larger scale systems such as Li *et al.*'s DevLex model, 2004).

We began by selecting a reasonably complex cognitive domain drawn from neuropsychology to assess both catastrophic interference and critical period effects in self-organizing feature maps (henceforth SOFMs). The training set comprised feature-based representations of exemplars from eight semantic categories (vehicles, tools, utensils, fruit, vegetables, dairy produce, animals and humans). These were based on vectors constructed by Small, Hart, Nguyen and Gordon (1996) to simulate patient performance in neuropsychological tests of semantic deficits. We split the training set into two halves that would correspond to *early* and *late* training experiences. The split was made in two ways. We either: (1) split each category in half, thereby producing two similar subsets; or (2) assigned living categories to one half and non-living categories to the other half, thereby producing two different subsets.

Each network was first exposed to the early set and, at a variable point across its training, a switch was made to the late set. We avoided interleaving training sets to maximize the effects of variability in the environment. We then evaluated the quality of the SOFM at the end of training. To assess *catastrophic interference*, we focused on performance on the *early* training set – had the early acquired knowledge been overwritten by the later acquired knowledge? To assess *critical periods*, we focused on performance on the *late* training set – was the network's ability to learn the late set compromised for switches that occurred at increasingly greater 'ages' of the network? Based on our review of the literature, we explored whether three additional factors modulated these effects:

1. SOFMs with *dynamic* parameters versus *fixed* parameters: we employed the standard Kohonen (1982) method of reducing neighbourhood and learning rate across training and contrasted it with a condition in which these two parameters were fixed at intermediate, compromise values throughout training. Can topographically well-organized maps only be achieved by reducing plasticity across training? If so, the existence of such maps in the brain might necessitate critical periods.
2. Resource levels: the capacity of the SOFM may be important for determining its flexibility to changes in the training environment. Intuitively, if there is no space left in a system when the environment changes, the system must either be compromised in learning the new or it must sacrifice the old. This manipulation either gave the map sufficient resources to employ a separate output unit for each pattern in the (combined) training set (*resource rich*), or reduced this level to approximately 25% capacity (*limited resource*).
3. Similarity: depending on the way in which the original training set was split, there was either *high similarity*

Table 1 *Training set information (note: 'humans' only had one prototype)*

Category	*n* items	*n* prototypes	*M* angle between prototypes	*M* angle between exemplars	*M* features active
vehicles	30	7	51.62	24.14	22.03
tools	22	6	51.72	28.71	17.91
utensils	20	6	46.96	26.65	15.85
dairy produce	15	4	48.09	29.31	15.33
animals	29	7	52.85	24.17	21.85
humans	21	1	–	33.19	28.52
fruit	26	6	46.05	23.06	17.99
vegetables	22	4	35.47	21.74	17.52

Table 2 *Angles between mean exemplars*

	vehicles	tools	utensils	dairy	animals	humans	fruit	vegetables
vehicles	–	66.43	64.29	76.94	75.4	73.29	78.75	80.58
tools	66.43	–	56.59	66.61	78.95	73.79	68.78	68.73
utensils	64.29	56.59	–	49.99	85.07	81.55	60.18	62.65
dairy	76.94	66.61	49.99	–	78.54	76.46	49.74	52.03
animals	75.4	78.95	85.07	78.54	–	36.05	72.81	73.43
humans	73.29	73.79	81.55	76.46	36.05	–	73.1	75.23
fruit	78.75	68.78	60.18	49.74	72.81	73.1	–	31.03
vegetables	80.58	68.73	62.65	52.03	73.43	75.23	31.03	–

or *low similarity* between the early and late training environment. If early knowledge and late knowledge are similar, will interference be eliminated since old knowledge generalizes to new? Conversely, under conditions of a radical change in training environment, will the effects of catastrophic interference outweigh those of the critical period?

Training sets

The training patterns were 185 exemplars derived from 41 prototypical concepts that spanned eight semantic categories: vehicles, tools, utensils, fruit, vegetables, dairy produce, animals and humans (adapted from the set used by Small *et al.*, 1996). Each training pattern was encoded according to the presence or absence of 154 meaningful semantic features (such as 'is_green' and 'is_food'), where the presence or absence of a particular feature was indicated by an activation value of 1 or 0 respectively. Exemplars were generated as semantically meaningful variations upon each prototype. For example, the prototype of 'apple' was green, whereas exemplars varied in properties such as colour ('is_red' rather than 'is_green') size, and shape. Vector-based representations then permitted evaluation of the similarity structure. For any pairs of vectors of length *n*, the similarity between them can be measured by the angle between the vectors in *n*-dimensional space. Table 1 shows the mean angle between prototypes and between each prototype and its accompanying exemplars, where 0° indicates complete similarity and 90° indicates entirely dissimilar or 'orthogonal' representations. The maximum angle between categories was 85° and the minimum 31°, with an average of 67° (individual values are shown in Table 2). Within each category, prototypes differed by an overall mean angle of 46°. The distance between prototypes and their accompanying exemplars was smaller, with a mean angle of 25°.

Four training sets were constructed from these 185 exemplars, arranged as two pairs. There were no repetitions of exemplars across pairs and each set consisted of a similar quantity of items. *Similar* training sets A and B consisted of 92 and 93 exemplars, respectively, and comprised half the exemplars of each category. Average vectors were calculated for the two sets. The angle between these mean vectors was 10°, indicating that the items within these two category sets were indeed highly similar. *Different* training sets A and B consisted of 98 and 87 exemplars, respectively. Set A consisted of exemplars from living and natural categories (humans, animals, fruit, and vegetables) while set B consisted of exemplars mainly from non-living categories (tools, utensils, vehicles, and dairy produce). We classified fruit and vegetables as living in this case since it enabled us to create two broadly internally consistent training sets which were nevertheless fairly dissimilar to each other: the angle between the mean vectors for the two *different* sets was 83°.

Architecture and algorithm

We employed two-dimensional SOFMs with a hexagonally arranged topology and 154 input units. The input layer was fully connected to the output layer. The output layer for *resource-rich* maps consisted of 196 units arranged in a 14 × 14 array. The output layer for *limited-resource* maps consisted of 49 units arranged in a 7 × 7 array. In these networks, during training, each input pattern produces a most-activated or winning output unit on the map. The activation u_i of each unit on the output layer is calculated via the summed product of the activations a_i of the input units that are connected to this unit and the strengths w_i of those connections:

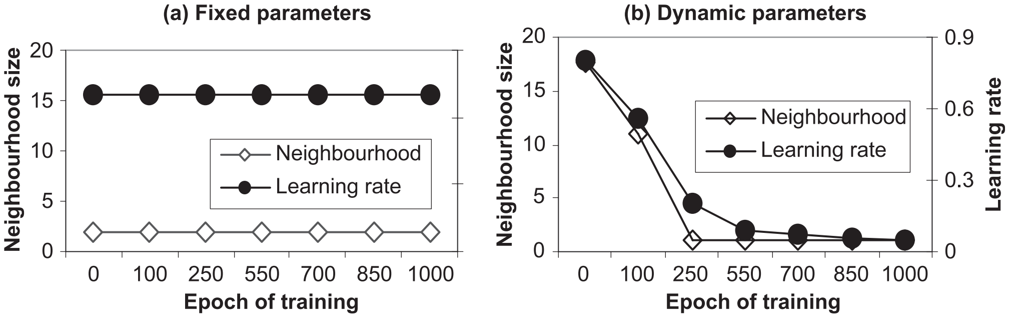

Figure 1 *Profile of parameter changes over learning that control functional plasticity in the SOFM for the (a)* fixed parameters *and (b)* dynamic parameters *conditions. It should be noted that the initial neighbourhood distance was set at the maximum distance value given the map size, and is shown in the figure for resource-rich (14 × 14) maps.*

$$u_i = \sum_i a_i w_i \quad [1]$$

The winning output unit for a given input pattern is the unit with the highest summed product. Some algorithms implement the selection of the winning output unit via a competitive process in the output layer, involving mutually excitatory short-range intra-layer connections, inhibitory long-range intra-layer connections, and cycling activation. In the current implementation, for simplicity the most active output unit is nominated as the winner. The winning unit updates its weights to the input layer, as do the units that surround the winner as a function of their distance from it. The distance is calculated using the Euclidean distance measure. Weights w_{iu} between input units i and winning unit u are updated via the following equation:

$$w_{iu}(t + 1) = w_{iu}(t) + \alpha(t)[a_i(t) - w_{iu}(t)] \quad [2]$$

where t denotes time, $\alpha(t)$ is the learning rate at time t (see below) and $[a_i(t) - w_{iu}(t)]$ is the difference between the activation of the input unit i and the current weight value (Kohonen, 1995). Output units n that fall within the neighbourhood of winning output unit u, as determined by the neighbourhood function (see below), also update their weights but now modified by a factor of 0.5 (Kohonen, 1995) so that:

$$w_{in}(t + 1) = w_{in}(t) + 0.5 \times \alpha(t)[a_i(t) - w_{in}(t)] \quad [3]$$

The learning rate and neighbourhood size were determined as follows. For the *fixed parameters* condition, the neighbourhood size was set to 2 and the learning rate to 0.7. This learning rate was selected following a parameter search, which found that maps of this type with a higher learning rate were more successful in developing representations, whereas maps with a lower learning rate produced very limited representations. These values were held constant throughout training. For the *dynamic parameters* condition, the two parameters decreased as a function of the number of training patterns presented. During the organizational phase of the map, which ran from the onset of training for 250 epochs (where 1 epoch corresponds to a single presentation of all the patterns in the training set), the learning rate was set at an initial value of 0.8 and decreased to a level of 0.2 by the start of the tuning phase (after 250 epochs). The exact formula for computing $\alpha(t)$ is shown in the Appendix. The neighbourhood distance was initially set to be the maximum neighbourhood distance for the map given its size, so that initially all units were neighbours (initial value of 18 for 14 × 14 maps, and 8 for 7 × 7 maps). This value then decreased to a level of 1 (immediate neighbours only) by the start of the tuning phase. The exact function determining the neighbourhood size is included in the Appendix. The parameter profiles for *fixed* and *dynamic* conditions are shown in Figure 1.

Map evaluation

Maps were evaluated using two methods. For visualization, colour-coded maps were created reflecting the category exemplars that activated each output unit. In order to generate these plots, we initially employed a cluster analysis upon all training patterns in order to gauge their relative similarity. The results of this analysis produced a vector indicating a sequence for the exemplars in the training set in terms of their relative similarity. Exemplars were then ordered according to this vector and an rgb (red green blue) colour value was then assigned to each exemplar from a series of graded colour values, with similar exemplars possessing a similar colour value. This colour value was then allocated to the winning output unit for that pattern. If the same unit was activated by more than one pattern, the colour values for that unit were averaged and the size of the unit plotted was increased (Thomas & Richardson, 2006).

Three quantitative metrics were used to assess map quality, based on those used by Oliver *et al.* (2000) in their simulations of typical and atypical SOFM development. The metrics were: *unit activity*, *discrimination*, and *organization*. First, *unit activity* was used as a basic indicator of the proportion of map space being used to represent the training set. This metric was calculated by summing the total number of winning units for the current training set and dividing by the map size. As a unit may categorize one or more patterns in map space, the *discrimination* metric was used to determine the granularity of categorization in map space. This metric calculated the mean proportion (between 0 and 1) of active units being used to represent each category in the current training set. Low values indicate coarse granularity and poor discrimination, with many different exemplars activating the same output unit. Higher values indicate fine-grained granularity and a good level of discrimination between exemplars. This measure was conceptually independent of the topographic layout of the clustering in map space. Topographic layout was evaluated using an *organization* metric. Under the hexagonal scheme, output units in map space were typically surrounded by a total of six immediate neighbours (units located in areas such as on the edge of map space had fewer immediate neighbours). For each active unit in map space representing one or more exemplars, this metric calculated the proportion of immediate neighbours that categorized exemplars of the same category. More specifically a neighbouring unit representing an exemplar from the same category generated a score of 1, whereas any inactive units or units representing another category contributed no score. Where a neighbouring unit responded to exemplars from more than one category, the unit was classified according to the category for which it was maximally active. The sum total of same neighbours was then calculated and then divided by the total number of immediate neighbours for the given unit. The output for this metric was the mean proportion between 0 and 1 of same neighbours over all active units in map space. Maps initialized with random weights typically have an organizational metric value of zero or near zero, indicating that no or very few neighbouring units represent exemplars from the same category. These maps therefore have none or almost no topographic organization. Conversely, a value near 1 indicates that the majority of neighbours classify exemplars from the same category and that there is good topographic organization.

Together, these three metrics provide the opportunity to identify map quality over several dimensions and they allow for the possibility that map characteristics may dissociate. Thus a map could, in principle, show good discrimination between exemplars but poor organization, or it could show good organization but poor discrimination between exemplars.

Training and testing regimes

Three sets of simulations were run. The first established the baseline development of maps for the split pattern sets when trained in isolation, against which the effects of interference or reduced plasticity could be assessed. The second set evaluated catastrophic interference effects and the third critical period effects. In each case, simulations followed a 2 × 2 × 2 design, with factors of parameters (*fixed* vs. *dynamic*), resources (*resource rich* vs. *limited resources*), and early–late training set similarity (*similar* vs. *different*). Simulations were counter-balanced across the split training sets, with A serving as the early set and B the late set or B as the early set and A as the late set. Illustrated data are collapsed and averaged over six replications with different random seeds determining initial weight randomization and random order of pattern presentation. All figures include standard errors of these means.

(i) Baseline development for single training sets

The developmental profile of fixed parameter and dynamic parameter maps was established by training maps on each of the four training sets (A and B similar; A and B different). Performance was assessed at 5, 50, 100, 250, 400, 550, 700, 850 and 1000 epochs. For the dynamic parameters condition, the organization phase ran from 0 to 250 epochs and the tuning phase from 250 to 1000 epochs.

(ii) Catastrophic interference effects

The network was initially trained on the early set. Training was then switched to the late set. Performance on the early set was assessed at the end of training. Switches took place at 5, 50, 250, 400, 550, 700, or 850 epochs of training. Note that two methods could be used to determine the 'end' of training. One could assess early set performance at 1000 epochs, so using a fixed total amount of training. However, this means that for switches occurring later in training, there is less opportunity for catastrophic interference to take place (i.e. only 150 epochs for a switch occurring at 850 epochs, compared to 995 epochs for a switch occurring after 5 epochs). Alternatively, one could assess early set performance following a fixed period of 1000 epochs following the switch, so for a switch at 850 epochs, network performance would be assessed at 1850 epochs. In practice, however, the effects of a switch stabilized relatively quickly, and therefore even the latest switch provided time for the effects of catastrophic interference to stabilize. Although we ran all simulations using both methods, we report here only the data for performance after 1000 epochs (first method), since the results are the same for both.

(iii) Critical period effects

The same method was used as in (ii) but performance was instead assessed at the end of training for the late set. Switches once more occurred after 5, 50, 250, 400, 550, 700 or 850 epochs of training on the early set.

Results

(i) Normal development of fixed and dynamic parameter maps

The typical developmental profiles of *fixed parameter* (FP) and *dynamic parameter* (DP) maps are displayed

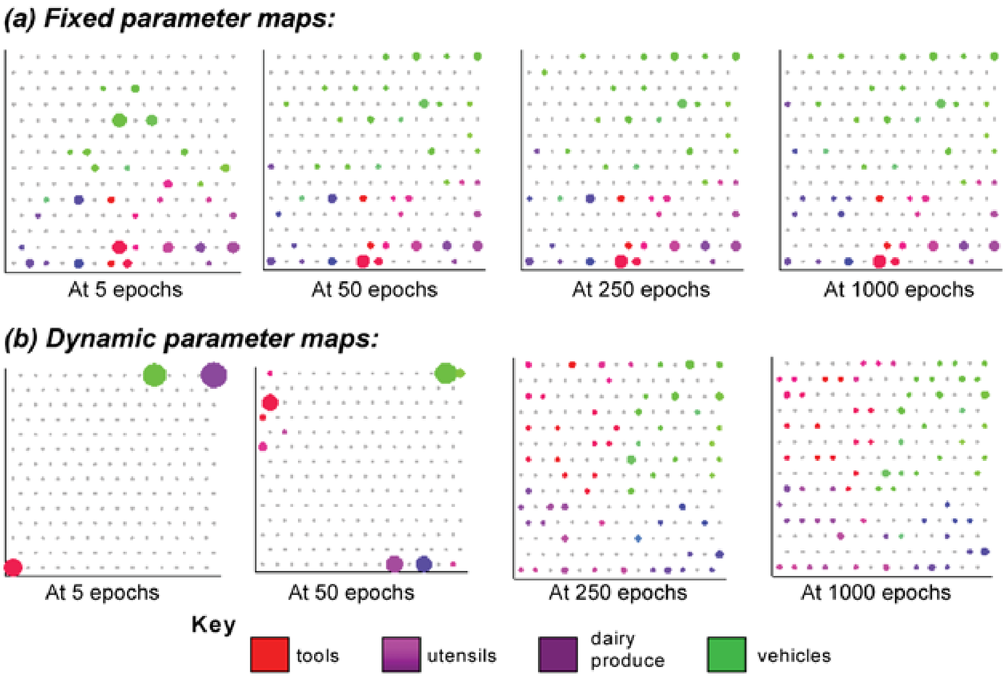

Figure 2 *SOFM plots illustrating the development of semantic categories for non-living categories for (a)* fixed *and (b)* dynamic parameter *maps. Maps with reducing learning rate and neighbourhood settings establish representations in map space more slowly than fixed parameter maps but produce maps with superior final organization and discrimination.*

with SOFM plots in Figure 2. These illustrate the emerging classification for one of the training sets (different set B: non-living categories). Both FP and DP maps developed topographically organized representations, marked by segregated areas of colour. Figure 2 indicates that the FP maps developed their representations more quickly but produced both fewer activated units and a lower level of exemplar discrimination in the endstate. By contrast, the DP maps developed more slowly but ultimately recruited more units and reached a higher level of discrimination. The quantitative metrics in Figure 3 confirm this impression. Note that the faster development of the FP map actually occurred when the DP map had higher plasticity (in terms of the learning rate and neighbourhood parameters). This is because the high plasticity of the DP map initially makes it unstable.

A reduction in map resources naturally resulted in fewer active units and therefore worse discrimination (see Thomas & Richardson, 2006). However, the relationship between FP and DP maps remained the same.

For the unit activity and discrimination metrics, the results were almost identical whether the *similar* and *different* subsets were used. This is despite the fact that the *different* subsets contained only four categories compared to the eight categories of *similar* subsets. In both cases, map resources were used to optimize discrimination between the exemplars present in the training set. The results were the same because for the *different* subsets, discrimination between exemplars increased, making use of the available resources. By contrast, the organization metric was affected by the choice of subset. This is because, by definition, the metric assesses how many neighbouring units represent exemplars from the same category. *Different* subsets possessed fewer categories than *similar* subsets, thus making any two units more likely to represent exemplars from the same category, resulting in a higher metric value for *different* subsets. However, the similarity effect was dependent both on parameter condition and resource level. For the DP network with plentiful resources, exemplar discrimination eventually became sufficiently fine-grained to reach the same level of organization for both *similar* and *different* subsets.

We now turn to consider the effects of a non-stationary training environment.

(ii) Catastrophic interference effects

Figure 4 depicts endstate performance on the early training set for conditions in which training switches to the late set after a certain number of epochs, compared with endstate performance when no switch took place (NS). Interference effects will be evidenced by poor endstate performance on the early set. The FP networks with rich resources demonstrated a drop in early set performance across all three metrics, irrespective of how late the shift occurred during training. These networks exhibited interference effects consistent with their continued level of plasticity. The interference effects were greater between *different* subsets than *similar* subsets, in line with equivalent findings from error-driven networks. In terms of unit activity and discrimination, the limited-resource FP networks showed interference effects only for switches

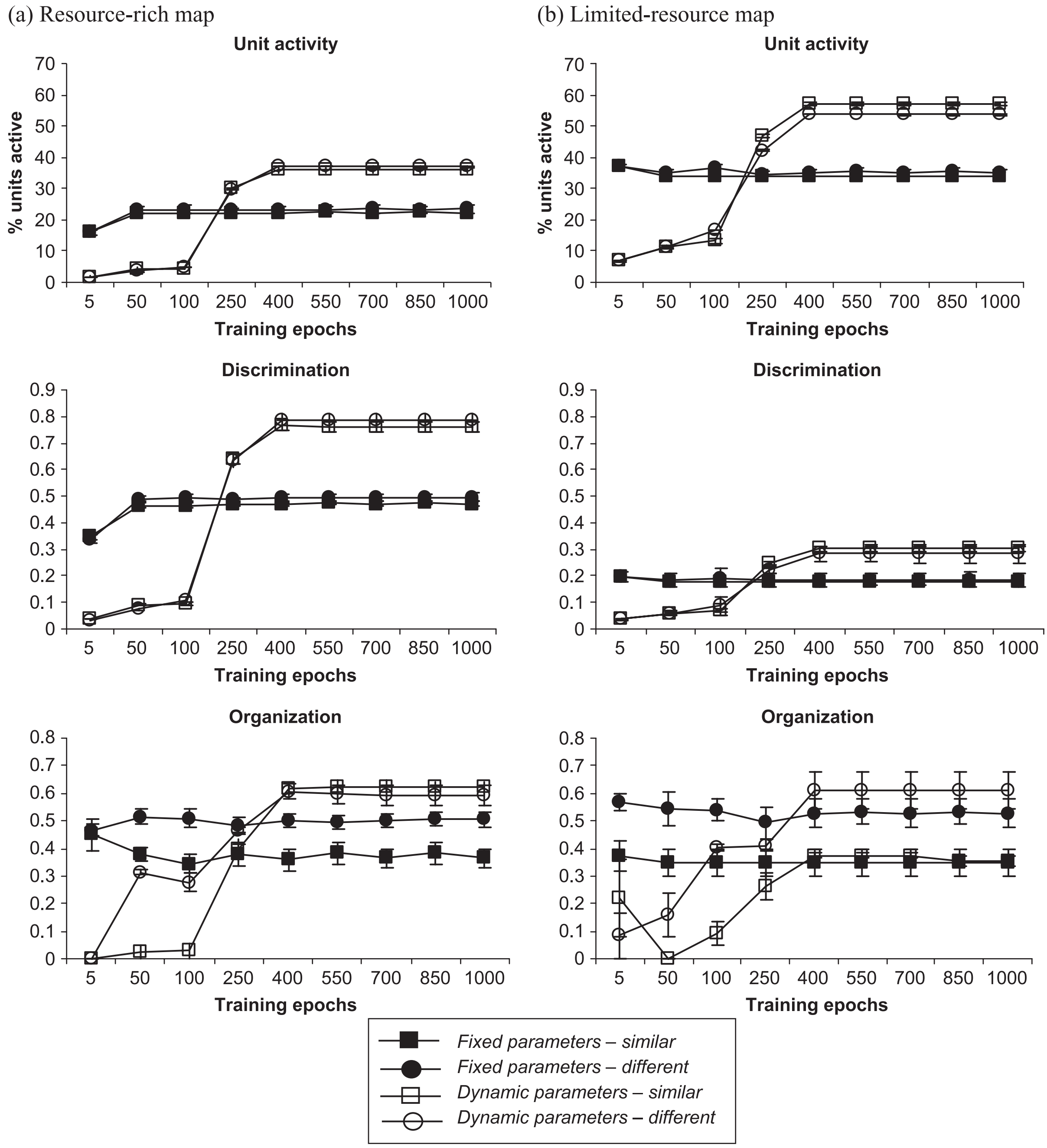

Figure 3 *Normal development: Metric results track changes in map quality over learning for (a)* resource-rich *maps and (b)* limited-resource *maps, for both* fixed *and* dynamic parameter *conditions.*

between *different* subsets. The map solution of the early set adequately generalized to the late set for the level of discrimination achievable and so no reorganization was necessary. Similarity effects were particularly evident in the organization metric for the limited-resources network, in which the competition for representational resources was maximized. The outcome of this competition depended to some extent on chance patterns of map organization, thereby increasing the variability of these data. In contrast, rich resources mitigated the effect of similarity on organization. Spare resources were now available to accommodate the new knowledge.

Consistent with their shift in emphasis from plasticity to stability across training, the DP networks exhibited interference predominantly for switches that occurred in the earlier parts of training. Once more the *different* subsets produced maximal interference. *Similar* subsets minimized the effects of the interference for early switches, since the organization fashioned by the new knowledge generalizes to the consistent old knowledge. Two further points are of note. First, the DP network's ability to preserve its old knowledge after a late occurring switch between *different* subsets was sensitive to map resources: the S-shaped curves in unit activity and discrimination

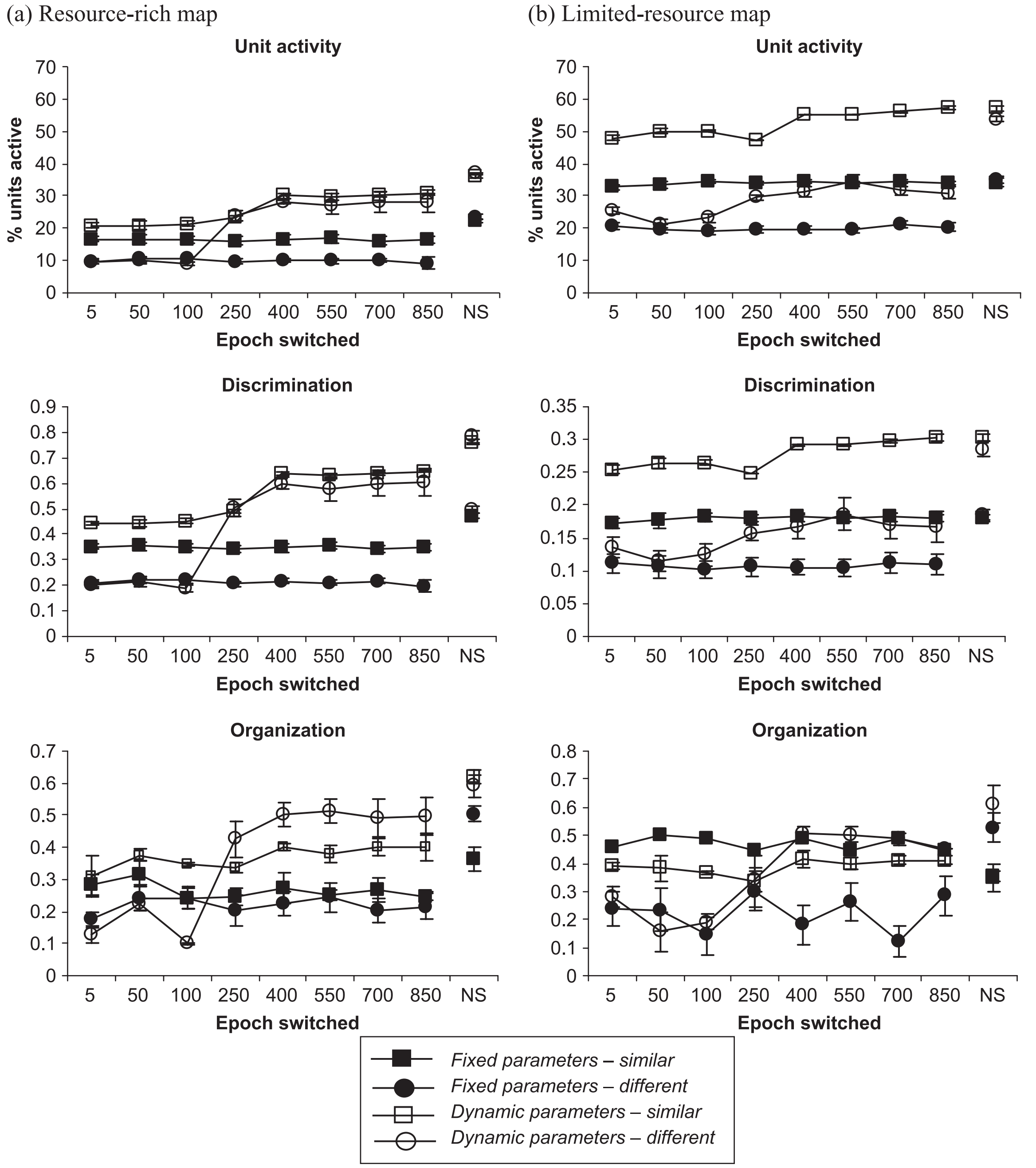

Figure 4 *Interference effects for the no-longer trained pattern set. Metric results show performance at the end of the normal period of training, for switches occurring at different points in training. Map quality was stable by 1000 epochs of training even for late occurring switches (NS = no-switch, i.e. for training on the early set in isolation).*

are present only in the resource-rich maps; for limited-resource maps, such a switch between *different* subsets always caused interference. Second, the rich-resource maps always experienced some interference irrespective of how late the switch occurred. Unless learning is deactivated, these systems cannot ensure complete stability in the face of a non-stationary environment.

In sum, conditions that maximized the necessity of change (a switch between *different* subsets), the opportunity for change (elevated intrinsic plasticity) or the impact of change (competition for limited resources) all led to interference effects in SOFMs. Where old knowledge generalized to new knowledge or where plasticity was sacrificed, stability prevailed. Could the interference be termed 'catastrophic'? To explore this question more fully we calculated an *index of interference* (IoI) based on the unit activity of normally developing maps minus the unit activity (tested using the early training set) of equivalent maps that underwent a switch. The scale of this index was –1/+1, where –1 indicates that the maps that underwent a switch during training out-performed those from during the normal condition, a value of zero indicated

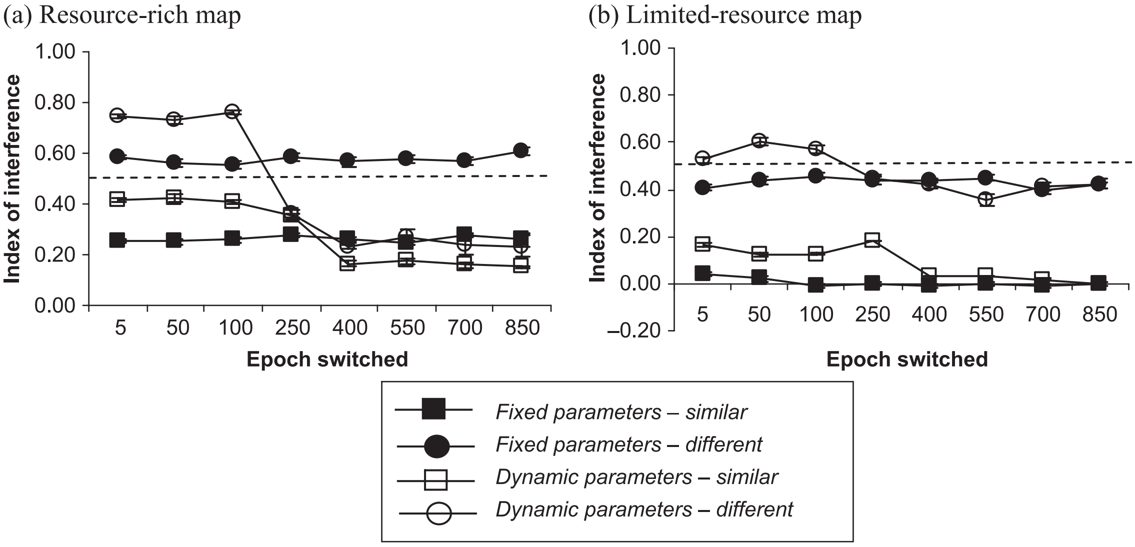

Figure 5 *Index of interference (IoI), which is the proportional change in unit activity compared to normally developing models. The figure shows IoI for switches occurring at different points in training for (a)* resource-rich *and (b)* limited-resource *maps. The level of interference is indicated by a positive value on the y-axis (with a maximum value of 1). No interference is represented by a value of zero, and negative values indicate where switched networks are more active than normally developing models. The dashed line indicates 50% interference, for which values above might be termed 'catastrophic'.*

no interference, and +1 indicates that prior knowledge has been completely overwritten. These results are shown in Figure 5 for both resource-rich and limited-resource conditions. Although interference is high for early switches in resource-rich dynamic parameter maps, if we define 'catastrophic' as losing more than 50% accuracy, in the majority of cases the level of interference is not catastrophic.

(iii) Critical period effects

Figure 6 depicts endstate performance on the late set for conditions in which training switched to this set increasingly further into the network's development. These data are compared with endstate performance when the network was trained on the same set from the beginning. A sensitive period would be demonstrated by increasingly poorer performance the later into the network's development that the training on this set begins; a critical period would be demonstrated by a point in the network's development after which the late set could not be learned at all.

For the FP networks, the point at which training commenced on the late set had no effect at all on endstate levels of unit activity or discrimination. Resource levels and similarity did not modulate this pattern. In contrast, the DP networks produced a sensitive period in line with the shift between organizational and tuning phases, that is, driven by the internal parameters of the system. Shifts to the late set occurring up until 100 epochs predicted an outcome similar to training on the late set from the beginning (e.g. around 75% discrimination in the resource-rich network). For shifts from 250 epochs onwards, the prognosis was much poorer, but importantly this pattern was strongly modulated by similarity. For the *similar* subsets, the latest switches only produced a decline to 62% discrimination. For the *different* subsets, the decline was much larger, to 25%.

Some degree of learning was always possible on the late set, suggesting that use of the strong sense of 'critical period' is not warranted for these networks. Nevertheless, the important finding is that the age-of-acquisition effects depended as much on similarity between old and new knowledge as intrinsic parameter settings. For unit activation and discrimination, the same kind of pattern was found in limited-resource networks. However, for both resource-rich and limited-resource networks, the sensitive period profile was not replicated in the organization metric, which was noisy but remained approximately level for switches at different epochs.

These results capture the outcome of a non-stationary environment at the end of the fixed training period, but they do not reveal the dynamics of change when a switch occurs. Figure 7 illustrates the process of reorganization triggered by a change in training set for two conditions. Figure 7(i) depicts a representative map for a late switch between *similar* subsets occurring at 700 epochs in the DP network with rich resources. By this point, the learning rate and neighbourhood parameters are at a level that limits subsequent change. Although the early and late subsets share no common training patterns, each nevertheless contains different exemplars from the same categories. The map produced by the early set is therefore likely to be useful for the late set. Figure 7(i) shows that following the switch there is a drop in discrimination (indicated by an increase in the size of the coloured dots). This is because distinctions between the exemplars of new knowledge are not captured. For example, taking the category 'vegetables', while the old knowledge may have included

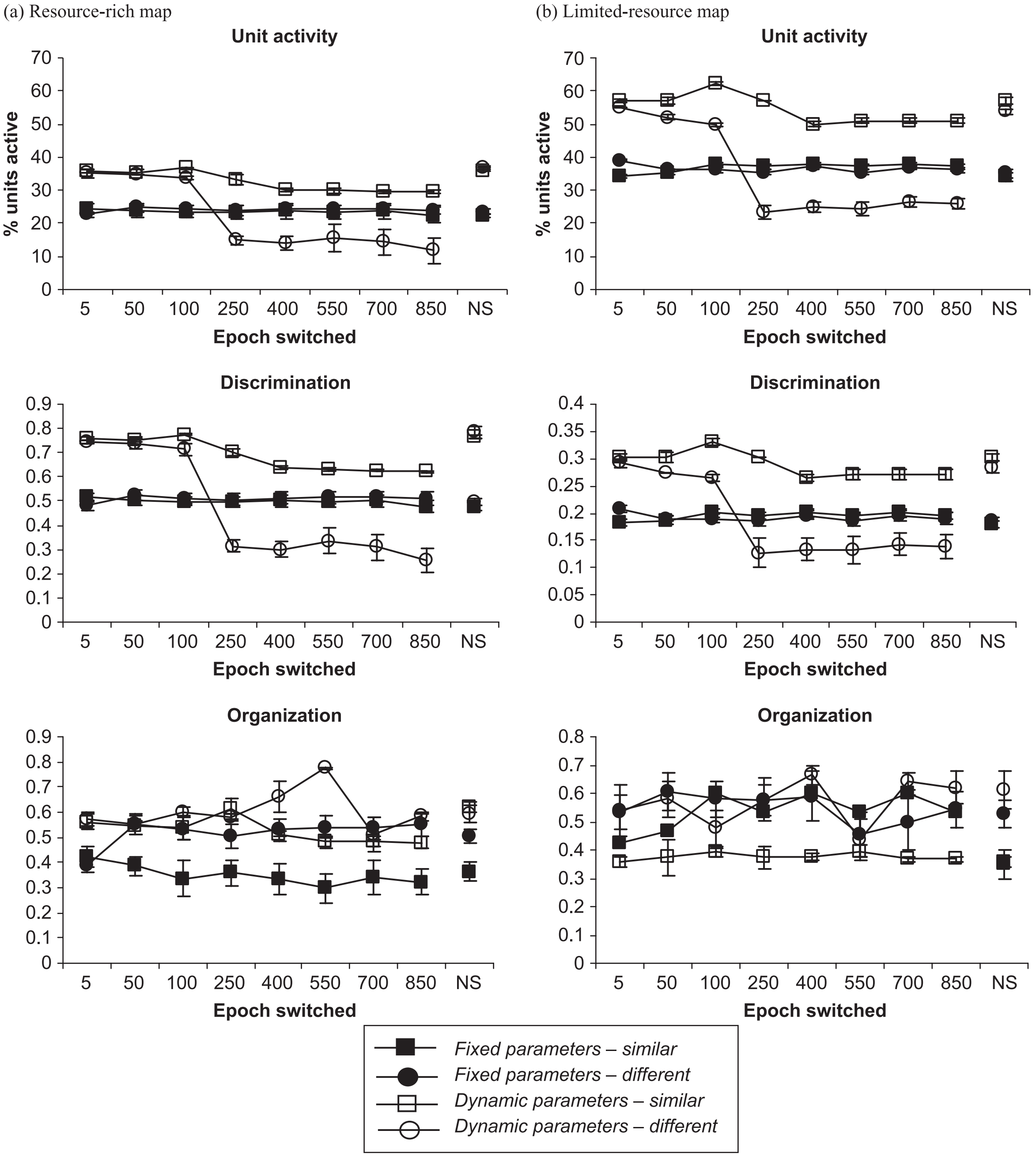

Figure 6 *Critical/sensitive period effects for the newly introduced pattern set. Metric results show performance at the end of the normal period of training, for switches occurring at different points in training. Map quality was stable by 1000 epochs of training (NS = no-switch, i.e. for training on the late set in isolation).*

the distinction between *lettuce*, *carrot*, and *potato* by activating separate output units for each, the new knowledge now contains *celery*, *parsnip*, and *turnip*, and these are initially conflated into a single output unit. However, it only takes fine-tuning over subsequent exemplars to learn these distinctions. Such cases are circled in Figure 7(i).

Figure 7(ii) illustrates the case of a late switch between *different* subsets (from non-living to living categories), again for a DP network with rich resources. Given that there is such limited overlap between old and new knowledge, one might question whether the representations developed by the early training set will be of any use in discriminating between patterns in the late training set. Figure 7(ii) shows that, without any further training, some discrimination is immediately available, albeit at a very coarse level. This is because the *different* subsets are not fully orthogonal, so that a single overlapping feature (such as *size*) used in discriminating the non-living categories of tools, utensils, dairy produce, and vehicles can then be employed to generate rough distinctions between the living categories of vegetables, fruit, animals, and humans. However, in line with the reduced plasticity of the DP network, few further distinctions can then be learned by the residual fine-tuning capacity of the system.

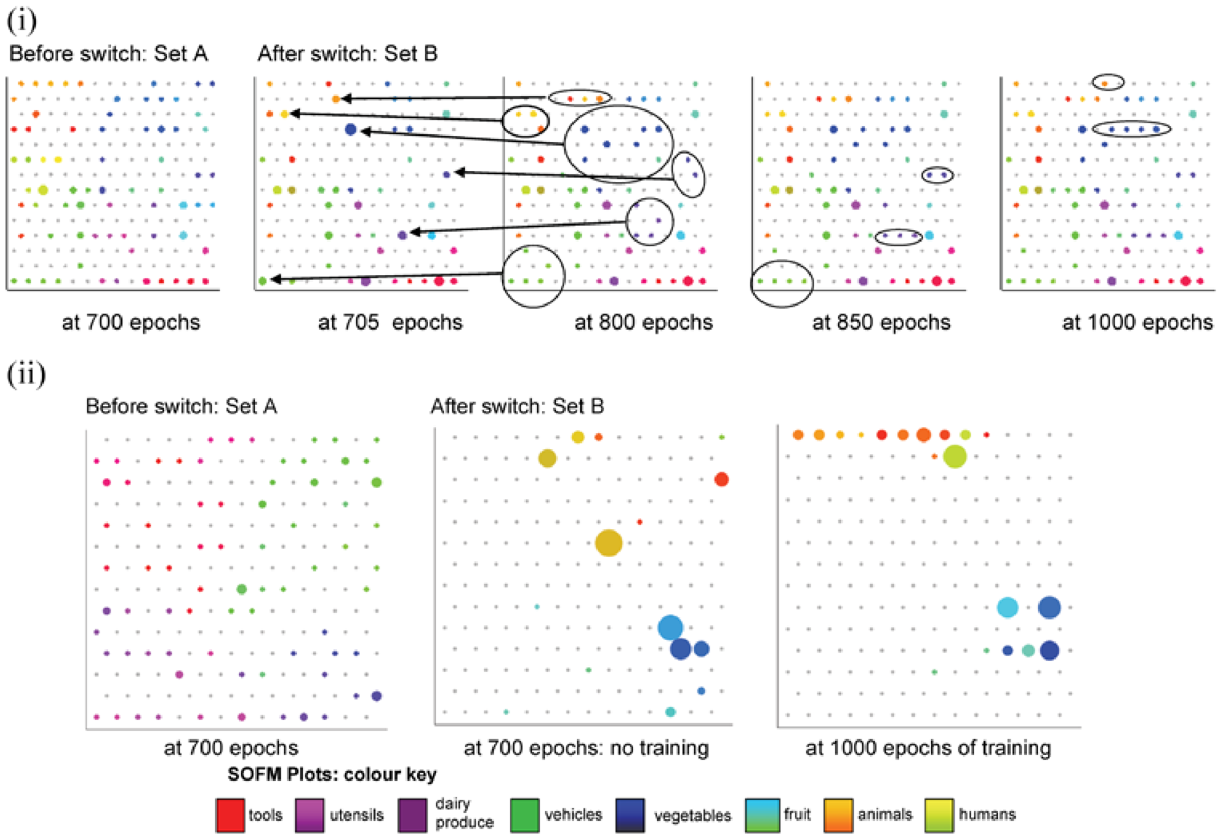

Figure 7 *Map reorganization immediately following a change in training set. Highlighted regions show the initial conflation following subsequent discrimination of exemplars in the new training set. (i) Resource-rich dynamic parameter map: late switch between* similar *subsets. (ii) Resource-rich dynamic parameter map: late switch between* different *subsets.*

Discussion

Both self-organizing and error-driven connectionist networks have been widely used to study mechanisms of cognitive development (Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett, 1996; Mareschal, Johnson, Sirios, Spratling, Thomas & Westermann, 2007). While self-organizing networks have been linked to explanations of critical or sensitive periods in development (Li *et al.*, 2004; McClelland *et al.*, 1999; O'Reilly & Johnson, 1994), error-driven networks have more often been associated with catastrophic interference effects where late-learned knowledge overwrites early-learned knowledge (French, 1999). In the current paper, we took a standard implementation of self-organized feature maps (Kohonen, 1995) and trained networks on a pattern set drawn from research into semantic deficits in cognitive neuropsychology (Small *et al.*, 1996), with the aim of evaluating the factors that mediate critical/sensitive period and interference effects in these systems. Our results demonstrated the following.

Two variations of the SOFM produced topographically organized representations of the categories in the training set. In the more traditional variation, the parameters of learning rate and neighbourhood size were reduced across training.[2] In this variation, learning was initially slow but eventually produced a high level of discrimination and organization. These networks demonstrated sensitive periods in development favouring the influence of early learning, with limited interference effects for changes in training occurring beyond this period. If a switch occurred within the sensitive period, the new knowledge was able to replace the old. However, after this replacement there was residual evidence that an early switch had taken place. This took the form of a lowering in the level of unit activity and discrimination that was ultimately attainable. These results are perhaps analogous to functional imaging data from Pallier *et al.* (2003) where Korean-born children who were adopted into French families at an early age and who showed loss of Korean when tested as adults nevertheless still exhibited depressed activation levels when listening to French when compared to native French speakers.

[2] This implementation does not necessarily imply a reduction in plasticity simply as a function of age (maturation). In the algorithm, the parameters reduce as a function of the number of training patterns encountered, that is, the level of experience (see Appendix). If the rate of experience can vary, the implementation is consistent with the idea that experience itself causes the closing of sensitive periods (Johnson, 2005). If experiences occur at a constant rate, the function is equivalent to a maturational reduction in plasticity.

In the second SOFM variation, the parameters of learning rate and neighbourhood size were fixed across training. The network learned very quickly but its final levels of discrimination and organization were poorer than the first variation. However, there was no indication of critical or sensitive periods in these networks; instead, interference effects were the salient characteristic. The clear inference is that the presence of topographic organization does not necessarily imply a system that will show critical/sensitive periods across development. The intrinsic properties of the learning device (i.e. its parameterization) are crucial in determining the trade-off between stability and plasticity. The simulations suggest a further trade-off: fast settling systems may retain plasticity at the expense of detail; higher performing systems may take longer to develop and involve sensitive periods. It is possible that different brain systems use the developing maps with different settings – fast, approximate and permanently plastic, versus slow, detailed and losing plasticity.

Research on catastrophic interference effects in error-driven connectionist networks pointed to the importance of the similarity between old and new knowledge (e.g. McCrae & Hetherington, 1993). The current simulations extended this work to self-organizing systems with comparable results: where a high degree of consistency existed between old and new knowledge, both the effects of critical periods and interference were attenuated; where the old and new knowledge were very different, critical period effects were maximized in the dynamic parameters network and interference effects were maximized in the fixed parameters network. The issue of similarity between old and new knowledge has been highlighted as one of the factors in the success of adults learning a second language beyond the sensitive period (see Hernandez, Li & MacWhinney, 2005, for a discussion of relevant literature in the context of SOFM models of bilingual acquisition).

Since both critical period effects and interference effects relate to a competition for representational resources, we also investigated whether such effects would be sensitive to the overall level of resources. A self-organizing network with fewer resources resulted in poorer discrimination between exemplars (see Thomas & Richardson, 2006, for further work). Reduced resources had no implications for critical period effects in discrimination. However, resources did exaggerate the effects of similarity on interference effects. With limited resources, a late switch to a different training set caused reduced unit activity and loss of discrimination for previously acquired knowledge – even in the dynamic parameters network where plasticity should have been attenuated. The implications of individual variation in neural resources for forming topographically organized systems are as yet unclear. For example, human studies have confirmed the presence of variation in the size of cortical areas without finding correlations in behavioural performance (Finlay, Cheung & Darlington, 2005). Studies of brain damage hint at a minimal level of resources necessary for cognitive development through the presence of 'crowding effects' after unilateral damage in childhood, in which there is a general lowering of IQ without marked specificity of behavioural deficits (e.g. Huttenlocher, 2002). And animal studies indicate that at the neural level, the result of reducing cortical resources prenatally without disrupting cortical input is the emergence of the same broad regions of functional specialization (visual, motor, somatosensory) but with reduced discrimination, i.e. more neurons responding to more than one modality (Huffman, Molnar, Van Dellen, Kahn, Blakemore & Krubitzer, 1999). The current simulations point to a further implication of resources for the stability of representations under conditions of a non-stationary environment.

We finish by briefly addressing two further issues. First, we consider the results from a more analytic perspective. Second, we consider why it should be important to develop good topographically organized representations, over and above representations that simply offer good discrimination.

Algorithmic determinants of plasticity

One limitation of the current findings is the extent to which they will generalize across different problem sets and self-organizing models. To address this question, we need to consider the details of the learning algorithms and how they contribute to changes in plasticity across training. The learning algorithm captures the opportunities for plasticity that a system will provide, while the similarity structure of the training set indicates whether these opportunities will be exploited. For example, in a review of computational models of sensitive periods in development, Thomas and Johnson (2006) pointed out that the Hebbian learning algorithm (from which most neural network learning algorithms are derived) includes a learning rate parameter ε that determines the size of the weight change Δw_{ij} between two neurons *i* and *j* when their activity (a_i, a_j) is correlated:

$$\Delta w_{ij} = \varepsilon a_i a_j \qquad [4]$$

The learning rate parameter is clearly a way to modulate the plasticity of the system. However, the weight change is *also* proportional to a_i and a_j. According to the algorithm, *simply by being more active*, a Hebbian system will become more plastic (see Thomas & Johnson, 2006, for discussion). When we consider the weight change algorithm for the Kohonen SOFM (equations 1 to 3), it is apparent that there are two modes by which a weight can change: either when it connects an input unit to the most active (winning) output unit (equation 2) or when it connects an input to an output unit within the neighbourhood of the winning output unit (equation 3).

From the perspective of a single input pattern, the size of the weight change in both these cases is proportional to the learning rate parameter and to the difference between the activation of the input unit and the current size of the weight. Weight change is small if both the input unit activation and existing weight are large, or if both the activation and weight are small. Conversely, maximum weight change occurs when the activation is large and the weight small (positive change) or the activation is small and the weight is large (negative weight change). The weight change that is triggered by a single pattern, then, will depend on whether the existing network is already behaving in the way that the pattern requires, conditioned by the two externally determined parameters of neighbourhood size and learning rate.

However, a pattern is part of a training set. The final factor influencing plasticity is whether, between presentations of a single pattern, other patterns have attempted to alter the value of a target weight. The nature of this influence depends on the relative similarity of a given pattern and the rest of the training set. Will other patterns use the same input and output units, and will they have altered the connecting weight in the same direction? The same output unit will be employed if it is the winner in both cases, implying some similarity between our single pattern and the patterns in the rest of the training set. It will also be employed if the output unit is in the neighbourhood of the winner. In this case, the single pattern and other training patterns may be dissimilar, which is more likely when the neighbourhood size is large and the map size small. The weight will be moved in the right direction by other patterns only if they have the same setting of the input unit as a target pattern, i.e. they share some degree of input similarity.

The parameters of the algorithm and the similarity structure of the training set therefore interact to determine plasticity. In the classical SOFM, the learning rate and neighbourhood size are initially set high. There is large weight change but much of it is between input units and neighbours rather than winners. This process acts as a smoothing function to produce widespread changes that reflect the broad similarity structure of the input set. In later learning, the smaller neighbourhood size focuses weight changes towards the properties of individual patterns, while a smaller learning rate decreases the instability caused by competition between patterns. In the reported simulations, we employed a relatively rich training set drawn from work on the modelling of neuropsychological deficits. Some self-organizing cognitive models have employed much simpler representations, such as a small number of bars or blobs falling across an input retina. These input sets place a weaker requirement on the algorithm to develop a richly structured topographic organization, but they do allow for more extreme manipulations of similarity, including completely orthogonal input patterns (e.g. Oliver *et al.*, 2000; O'Reilly & Johnson, 1994; McClelland *et al.*, 1999). For reasons of practicality, we took the exaggerated case of a sudden and absolute change in training set. Additional work would be necessary to assess the extent to which interleaving old and new knowledge might alleviate interference effects, in line with the findings from work on error-driven networks.

Implementations of self-organizing feature maps differ in the details of their algorithms. Some models employ weight decay or normalization in their learning algorithms to keep the total weight size constant; other models provide the units of the output layer with threshold functions; other models implement a competitive process to select the winning output unit for each input pattern via intra-layer connections, and include adaptive changes to the intra-layer weights as part of learning, thereby altering neighbourhood effects; other models allow the recruitment of new output units across training for very novel inputs and include bi-directional connections between input and output layers that can change the similarity structure of the input (see, e.g. Grossberg, 1987; Li *et al.*, 2004; Miller *et al.*, 1989; Oliver *et al.*, 2000; O'Reilly & Johnson, 1994). Notably, not all models use dynamic parameter changes across training, instead achieving their topographic organization with fixed parameters. However, these fixed parameter networks also tend to be the models with less richly structured training sets, thereby placing weaker demands on global organization. The motivation for reducing neighbourhood size and learning rate across training in the Kohonen net is specifically to achieve good map organization for a training set with rich similarity structure.

What significance would the difference have for learning algorithms in terms of the balance between critical period and interference effects? Further simulation work is required to answer this question definitively, but we can anticipate at least two differences that would increase the probability of critical period effects and attenuate interference effects. First, if the decay of unused weights between input and output layers ever permitted weight sizes to drop to zero (effectively pruning unused connections), then initially unused areas of the input space would lose the ability to activate the output layer. Relatedly, if output units had fixed thresholds and weights decayed (or were weakened by normalization as other weights strengthened), unused areas of the input space might no longer be able to propagate enough activation to push output units above threshold. Second, intra-layer competitive processes on the output layer might result in assimilation effects, whereby novel inputs that are highly similar to existing exemplars simply serve to activate the output unit for that existing exemplar and therefore fail to trigger adaptation in the network. That is, the system does not adapt because it has failed to perceive anything different in the world that might require the generation of new representations (see McClelland *et al.*, 1999, for simulation work related to adult Japanese learners of English and the /l/–/r/ contrast). These two cases are illustrated in Figure 8.

How important is it to have good topographic maps?

Turning to the second issue, Figures 4 and 6 indicated that switches in training set played a stronger role in modulating the unit activation and discrimination metrics

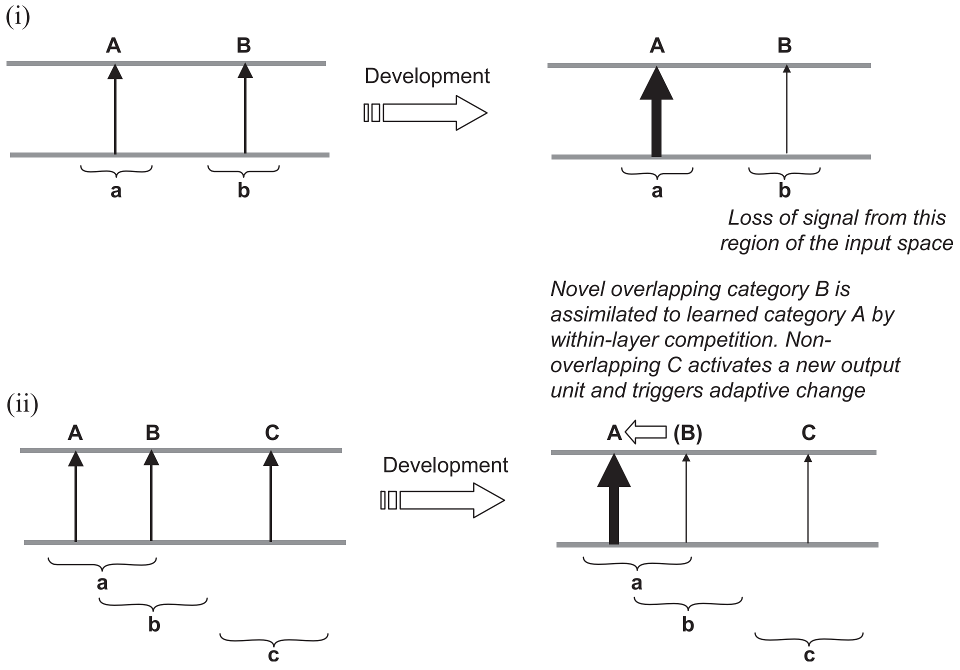
(i)
A
B
Development
a
b
A
B
a
b
Loss of signal from this region of the input space
Novel overlapping category B is assimilated to learned category A by within-layer competition. Non-overlapping C activates a new output unit and triggers adaptive change
(ii)
A
B
C
Development
a
b
c
A
(B)
C
a
b
c

Figure 8 *Additional algorithmic assumptions that could affect the ongoing functional plasticity of a SOFM: (i) Loss of signal via weight normalization/decay or via fixed output unit thresholds. After training on category A, there is loss of signal for novel category B. (ii) Assimilation of novel inputs into existing categories via intra-layer competition. After training on category A, novel category B (but not novel category C) is assimilated into category A and so does not trigger reorganization.*

than the organization metric, especially for critical periods. This led us to consider what might be the importance of good topographic organization for driving behaviour, over and above developing a highly activated map with good discrimination between exemplars in the training set. While there may be metabolic and signalling advantages of having units that represent the same information close together on a neural sheet, are there necessarily computational advantages of good topographic organization? It has certainly been argued in the literature that the development of self-organizing feature maps with *poor topology* may result in developmental disorders (Oliver *et al.*, 2000) and even that maps that are malformed in a certain way could lead to symptoms of autism (Gustaffson, 1997).

The impact of disruptions of topology (independent of discrimination) would seem to depend on certain assumptions about the downstream system that the map is driving. In particular, bad topology will disrupt behaviour if (a) the downstream system also has a topographic organization and (b) units in the downstream system have receptive fields that cover only a limited region of the map. Such an architecture would mean that a given downstream unit could not be driven by map units with widely disparate locations. The second assumption of receptive fields is problematic, however. Unless the map locations of relevant categories could be anticipated in advance, how would the downstream units know where to position their receptive fields on the map? In our simulations, while the relative organization of categories was predictable (e.g. animals would fall next to humans), the absolute location was not (e.g. whether animals were represented top-left or bottom-right).

The implication of such unpredictability if it is replicated in brain development is that the receptive fields of the downstream system could not be pre-specified but would have to be learned. Downstream systems must co-develop with upstream systems. Under a simple version of this process, the SOFM and the downstream system would begin by being fully connected. As the topology of each was established, receptive fields would emerge as the outcome of a regressive developmental process (illustrated in Figure 9). If this is correct, whether or not a map with non-optimal topology manifests in a disorder would depend on the severity of the disruption to the upstream map and the point during development at which the disruption took place. As importantly, it would also depend on the degree of compensation available in the downstream map and the connectivity between the two maps, requiring us also to consider the developmental trajectory and plasticity conditions of that downstream system.[3]

[3] Similar arguments could be made regarding the optimal setting of the discrimination metric. Coarse representations may be better for extracting broad categories, while exemplar-based representations offer better old–new discrimination. A downstream system with wide receptive fields would conflate neighbouring units into a single categorical response, while one with narrow receptive fields could be exemplar driven. (If the width of the receptive fields were modulated by attention, both responses would be available.) The discrimination metric therefore has to be considered both in the context of the downstream system and the demands of the task.

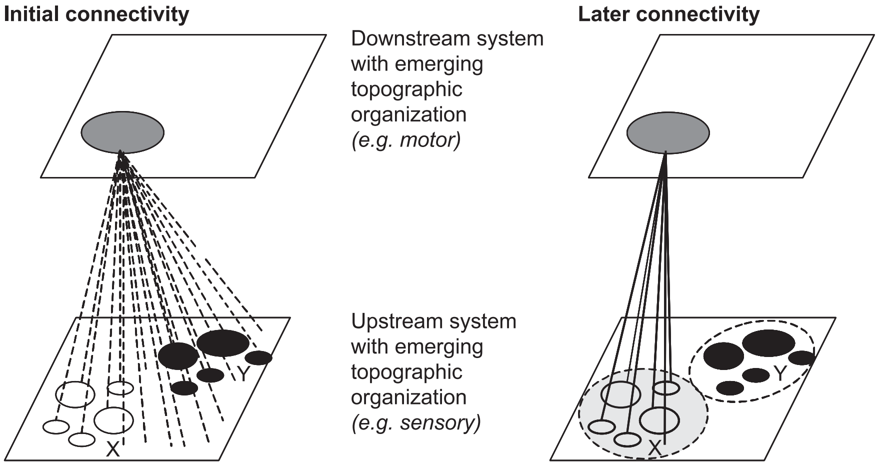
Initial connectivity
Later connectivity
Downstream system with emerging topographic organization (e.g. motor)
Upstream system with emerging topographic organization (e.g. sensory)
X
Y
X
Y

Figure 9 *Why should poor map organization impair function? Two developmental assumptions are necessary: (i) the downstream output system is also topographically organized; (ii) the output system has emergent receptive fields with restricted coverage of the input layer. Points X and Y can drive the same downstream unit before the emergence of receptive fields but not afterwards (see text for further details).*

In sum, although we can evaluate the quality of SOFMs in isolation, the relevance of the metrics is ultimately dependent on the systems to which the map is connected and the processes it is driving. We have demonstrated that the impact of a non-stationary environment on a SOFM is contingent on its plasticity conditions, as well as factors such as similarity and resources. But the impact on *behaviour* of a non-stationary environment is additionally contingent on the plasticity conditions that prevail in the other systems to which it is connected, as well as the nature of the connectivity between them.

Conclusion

For a self-organizing feature map in a non-stationary environment, internal parameter settings, available representational resources, and the similarity between old and new knowledge all influence the stability of acquired knowledge and the sensitivity of the system to change. Topographically organized systems are possible in networks that do not exhibit critical or sensitive periods, but maps optimized for high discrimination, and indeed those most widely used in models of cognitive development, do necessitate reducing sensitivity to change with increasing experience.

Appendix

Dynamic parameter changes over learning in the SOFM (Kohonen, 1982, 1995)

Organization phase

The learning rate is given by

$$lr = (proportion \times (\max\_lr - \min\_lr)) + \min\_lr \qquad [5]$$

where *lr* is the learning rate, max_*lr* is the highest learning rate at the start of the organization phase and min_*lr* is the tuning phase learning rate. *Proportion* is given by

$$proportion = 1 - \frac{(curptot - 1)}{orgpats} \qquad [6]$$

where *curptot* is the current total of pattern presentations and *orgpats* is the total number of pattern presentations in the organization phase.

The neighbourhood distance is given by

$$nd = (proportion \times (\max\_nd - 1)) + \min\_nd \qquad [7]$$

where min_*nd* is minimum neighbourhood distance and max_*nd* is maximum neighbourhood distance.

Tuning phase

The learning rate and neighbourhood distance in the tuning phase are given by

$$lr = \frac{(\min\_lr \times orgpats)}{(curptot - 1)} \qquad [8]$$

$$nd = \min\_nd \qquad [9]$$

Acknowledgement

This research was supported by MRC Career Establishment Grant G0300188 to Michael Thomas.

References

Ans, B., Rousset, S., French, R.M., & Musca, S. (2004). Self-refreshing memory in artificial neural networks: learning temporal sequences without catastrophic forgetting. *Connection Science*, **16** (2), 71–99.

Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J.F. Kroll & A.M.B. de Groot (Eds.), *Handbook of bilingualism* (pp. 109–127). Oxford: Oxford University Press.

Brainard, M.S., & Doupe, A.J. (2002). What songbirds teach us about learning. *Nature*, **417**, 351–358.

Braun, C., Heinz, U., Schweizer, R., Weich, K., Birbaumer, N., & Topka, H. (2001). Dynamic organization of the somatosensory cortex induced by motor activity. *Brain*, **124**, 2259–2267.

Buonomano, D.V., & Merzenich, M.M. (1998). Cortical plasticity: from synapses to maps. *Annual Review of Neuroscience*, **21**, 149–186.

DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean? In J.F. Kroll & A.M.B. de Groot (Eds.), *Handbook of bilingualism* (pp. 88–108). Oxford: Oxford University Press.

Doupe, A.J., & Kuhl, P.K. (1999). Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience*, **22**, 567–631.

Ellis, A.W., & Lambon-Ralph, M.A. (2000). Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **26**, 1103–1123.

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Finlay, B.L., Cheung, D.T.-M., & Darlington, R.B. (2005). Developmental constraints on developmental structure in brain evolution. In Y. Munakata & M.H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and Performance XXI* (pp. 131–162). Oxford: Oxford University Press.

French, R.M. (1991). Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. In *Proceedings of the 13th Annual Cognitive Science Society Conference* (pp. 173–178). Hillsdale, NJ: Lawrence Erlbaum.

French, R.M. (1992). Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, **4**, 365–377.

French, R.M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, **3** (4), 128–135.

French, R.M., & Chater, N. (2002). Using noise to compute error surfaces in connectionist networks: a novel means of reducing catastrophic forgetting. *Neural Computation*, **14** (7), 1755–1769.

French, R.M., Mareschal, D., Mermillod, M., & Quinn, P.C. (2004). The role of bottom-up processing in perceptual categorization by 3- to-4-month-old infants: simulations and data. *Journal of Experimental Psychology: General*, **133** (3), 382–397.

Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, **11**, 23–64.

Gustafsson, L. (1997). Inadequate cortical feature maps: a neural circuit theory of autism. *Biological Psychiatry*, **42**, 1138–1147.

Hensch, T.K. (2004). Critical period regulation. *Annual Review of Neuroscience*, **27**, 549–579.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, **9** (5), 220–225.

Hertz-Pannier, L., Chiron, C., Jambaqué, I., Renaux-Kieffer, V., Van de Moortele, P.-F., Delalande, O., Fohlen, M., Brunelle, F., & Le Bihan, D. (2002). Late plasticity for language in a child's non-dominant hemisphere: a pre- and post-surgery fMRI study. *Brain*, **125**, 361–372.

Hubel, D., & Weisel, T. (1963). Receptive fields of cells in the striate cortex of very young, visually inexperienced kittens. *Journal of Neurophysiology*, **26**, 944–1002.

Huffman, K.J., Molnar, Z., Van Dellen, A., Kahn, D.M., Blakemore, C., & Krubitzer, L. (1999). Formation of cortical fields on a reduced cortical sheet. *Journal of Neuroscience*, **19** (22), 9939–9952.

Huttenlocher, P.R. (2002). *Neural plasticity: The effects of the environment on the development of the cerebral cortex*. Cambridge, MA: Harvard University Press.

Johnson, J.S., & Newport, E.L. (1989). Critical period effects in second language learning – the influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, **21** (1), 60–99.

Johnson, J.S., & Newport, E.L. (1991). Critical period effects on universal properties of language – the status of subjacency in the acquisition of a second language. *Cognition*, **39** (3), 215–258.

Johnson, M.H. (2005). Sensitive periods in functional brain development: problems and prospects. *Developmental Psychobiology*, **46**, 287–292.

Jones, E.G. (2000). Cortical and subcortical contributions to activity-dependent plasticity in primate somatosensory cortex. *Annual Review of Neuroscience*, **23**, 1–37.

Knudsen, E.I. (2004). Sensitive periods in the development of the brain and behaviour. *Journal of Cognitive Neuroscience*, **16** (8), 1412–1425.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**, 59–69.

Kohonen, T. (1995). *Self-organizing maps*. New York: Springer.

Kortge, C. (1990). Episodic memory in connectionist networks. In *Proceedings of the 12th Annual Cognitive Science Society Conference* (pp. 764–771). Hillsdale, NJ: Lawrence Erlbaum.

Lambon Ralph, M.A., & Ehsan, S. (2006). Age of acquisition effects depend on the mapping between representations and the frequency of occurrence: empirical and computational evidence. *Visual Cognition*, **13** (7–8), 928–948.

Lenneberg, E. (1967). *Biological foundations of language*. New York: Wiley.

Li, P., Farkas, I., & McWhinney, B. (2004). Early lexical development in a self-organising neural network. *Neural Networks*, **17**, 1345–1362.

Lorenz, K.Z. (1958). The evolution of behaviour. *Scientific American*, **199**, 67–74.

McCandliss, B.D., Fiez, J.A., Protopapas, A., Conway, M., & McClelland, J.L. (2002). Success and failure in teaching the [r]–[l] contrast to Japanese: predictions of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective and Behavioral Neuroscience*, **2**, 89–108.

McClelland, J.L., McNaughton, B.L., & O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102** (3), 419–457.

McClelland, J.L., Thomas, A.G., McCandliss, B.D., & Fiez, J.A. (1999). Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data. In J. Reggia, E. Ruppin, & D. Glanzman (Eds.), *Progress in brain research. Volume 121: Disorders of brain, behavior and cognition: The neurocomputational perspective* (pp. 75–80). Amsterdam: Elsevier.

McCloskey, M., & Cohen, N. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. In G.H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 24, pp. 109–164). New York: Academic Press.

McCrae, K., & Hetherington, P.A. (1993). Catastrophic interference is eliminated in pre-trained networks. In *Proceedings of the 15th Annual Cognitive Science Society Conference* (pp. 723–728). Hillsdale, NJ: Lawrence Erlbaum.

Marchman, V.A. (1993). Constraints on plasticity in a connectionist model of English past tense. *Journal of Cognitive Neuroscience*, **5** (2), 215–234.

Mareschal, D., Johnson, M., Sirios, S., Spratling, M., Thomas, M.S.C., & Westermann, G. (2007). *Neuroconstructivism: How the brain constructs cognition*. Oxford: Oxford University Press.

Mareschal, D., & Quinn, P.C. (2001). Categorization in infancy. *Trends in Cognitive Sciences*, **5** (10), 443–450.

Mareschal, D., Quinn, P., & French, R.M., (2002). Asymmetric interference in 3- to 4-month-olds' sequential category learning. *Cognitive Science*, **79**, 1–13.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, **59**, 334–366.

Miller, K.D., Keller, J.B., & Stryker, M.P. (1989). Occular dominance column development: analysis and simulation. *Science*, **245**, 605–615.

Mirman, D., & Spivey, M. (2001). Retroactive interference in neural networks and in humans: the effect of pattern-based learning. *Connection Science*, **13** (3), 257–275.

Oliver, A., Johnson, M.H., Karmiloff-Smith, A., & Pennington, B. (2000). Deviations in the emergence of representations: a neuroconstructivist framework for analysing developmental disorders. *Developmental Science*, **3**, 1–23.

O'Reilly, R.C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, **2**, 455–462.

O'Reilly, R.C., & Johnson, M.H. (1994). Object recognition and sensitive periods: a computational analysis of visual imprinting. *Neural Computation*, **6**, 357–389.

Pallier, C., Dehaene, S., Poline, J.-B., LeBihan, D., Argenti, A.-M., Dupoux, E., & Mehler, J. (2003). Brain imaging of language plasticity in adopted adults: can a second language replace a first? *Cerebral Cortex*, **13**, 155–161.

Ratcliff, R. (1990). Connectionist models of recognition memory – constraints imposed by learning and forgetting functions. *Psychological Review*, **97** (2), 285–308.

Robins, A. (1995). Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science*, **7**, 123–146.

Robins, A., & McCallum, S. (1998). Catastrophic forgetting and the pseudo-rehearsal solution in Hopfield-type networks. *Connection Science*, **10** (2), 121–135.

Schaefer, M., Flor, H., Heinze, H.J., & Rotte, M. (2005). Dynamic shifts in the organization of the primary somatosensory cortex induced by bimanual spatial coupling of motor activity. *Neuroimage*, **25** (2), 359–400.

Seidenberg, M.S., & Zevin, J.D. (2006). Connectionist models in developmental cognitive neuroscience: critical periods and the paradox of success. In Y. Munakata & M.H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and Performance XXI* (pp. 315–347). Oxford: Oxford University Press.

Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: evidence from an emerging sign language in Nicaragua. *Science*, **305**, 1779–1782.

Sharkey, N., & Sharkey, A. (1995). An analysis of catastrophic interference. *Connection Science*, **7**, 301–329.

Small, S.L., Hart J., Nguyen, T., & Gordon, B. (1996). Distributed representations of semantic knowledge in the brain: computational experiments using feature based codes. In J. Reggia, E. Ruppin, & R.S. Berndt (Eds.), *Neural modelling of brain and cognitive disorders* (pp. 109–132). Singapore: World Scientific.

Thomas, M.S.C. (2004). State of connectionism 2004. *Parallaxis*, **8**, 43–61.

Thomas, M.S.C., & Johnson, M.H. (2006). The computational modelling of sensitive periods. *Developmental Psychobiology*, **48** (4), 337–344.

Thomas, M.S.C., & Karmiloff-Smith, A. (2002). Residual normality: friend or foe? *Behavioural and Brain Sciences*, **25**, 772–787.

Thomas, M.S.C., & Richardson, F. (2006). Atypical representational change: conditions for the emergence of atypical modularity. In Y. Munakata & M.H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and Performance XXI* (pp. 315–347). Oxford: Oxford University Press.

Uylings, H.B.M. (2006). Development of the human cortex and the concept of 'critical' or 'sensitive' periods. *Language Learning*, **56**, Suppl.1, 59–90.

Westermann, G., & Miranda, E.R. (2002). Modelling development of mirror neurons for auditory-motor integration. *Journal of New Music Research*, **31** (4), 367–375.

Westermann, G., & Miranda, E.R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language*, **89**, 393–400.

Zevin, J.D., & Seidenberg, M.S. (2002). Age of acquisition effects in reading and other tasks. *Journal of Memory and Language*, **47**, 1–29.

Received: 8 March 2007
Accepted: 3 May 2007