Connection Science, Vol. 12, No. 2, 2000, 143–162

From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories[1]

ANGELO CANGELOSI,* ALBERTO GRECO† and STEVAN HARNAD‡

**Centre for Neural and Adaptive Systems,*
University of Plymouth, Drake Circus,
Plymouth PL4 8AA, UK
email: angelo@soc.plym.ac.uk
tel: +44 1752 232559
fax: +44 1752 232540
†Department of Anthropological and Psychological Sciences,
University of Geona, Italy
‡Cognitive Science Centre,
University of Southampton, UK

Abstract. Neural network models of categorical perception (compression of within-category similarity and dilation of between-category differences) are applied to the symbol-grounding problem (of how to connect symbols with meanings) by connecting analogue sensorimotor projections to arbitrary symbolic representations via learned category-invariance detectors in a hybrid symbolic/non-symbolic system. Our nets are trained to categorize and name 50×50 pixel images (e.g. circles, ellipses, squares and rectangles) projected on to the receptive field of a 7×7 retina. They first learn to do prototype matching and then entry-level naming for the four kinds of stimuli, grounding their names directly in the input patterns via hidden-unit representations ('sensorimotor toil'). We show that a higher-level categorization (e.g. 'symmetric' versus 'asymmetric') can be learned in two very different ways: either (1) directly from the input, just as with the entry-level categories (i.e. by toil); or (2) indirectly, from Boolean combinations of the grounded category names in the form of propositions describing the higher-order category ('symbolic theft'). We analyse the architectures and input conditions that allow grounding (in the form of compression/separation in internal similarity space) to be 'transferred' in this second way from directly grounded entry-level category names to higher-order category names. Such hybrid models have implications for the evolution and learning of language.

Keywords: symbol grounding, categorical perception, neural networks, pattern recognition.

1. Introduction

The non-linguistic or prelinguistic part of us is purely robotic, which is to say purely sensorimotor (Harnad 1995). Or, to put it in a more ecumenical way, so as to make it clear that 'robotic' is anything but pejorative in this context: the pinnacle of our hierarchy of robotic capacities is a very special kind of sensorimotor skill, that of: (1)

collectively making unique, arbitrary responses that *name* objects, events and states of affairs; and (2) combining those responses to *describe* further objects, events and states (not necessarily present ones and not necessarily describing them truly). This ability of a robot community to share names, descriptions and the thinking and knowledge that underlie them is what it means to have and use language (Harnad 1996).

The classically sensorimotor component of this ability—the non-linguistic interaction with those objects, events and states—is the traditional domain of robotics: vision, locomotion, object recognition and manipulation. But even in modelling that domain, robotics has found it helpful, and perhaps necessary, to make use of internal structures and processes that are, if not linguistic, then at least symbolic.

1.1. *The symbol-grounding problem*

A computer program is a set of rules (algorithms) for manipulating meaningless symbols in a way that can be systematically interpreted as meaning something (e.g. payroll calculations, solutions to quadratic equations, chess moves, moon-landing simulations, or natural language text). But although the symbols are meaningfully interpretable by their users, they are meaningless in and of themselves, just as the symbols on the pages of this paper are. For this reason, symbol systems alone are not viable models of the mind—they cannot be the language of thought. This is the symbol-grounding problem (Harnad 1990). To embody thought, a cognitive system must be autonomous: the connections between its symbols and what they stand for must be direct and intrinsic to the system, rather than having to be mediated by an external user/interpreter. Some researchers have suggested that hybrid connectionist and symbol models can deal with the symbol-grounding problem (Harnad 1993). Others have suggested that connectionist systems can handle compositionality and systematicity on their own, without requiring any hybridization with symbol systems (van Gelder 1990, Hadley 1994, Hadley and Cardei 1999).

A symbol is a physical object that represents other objects. In the most important and powerful symbol systems, those of natural language, symbols can express thoughts by being combined and recombined to form propositions. All artificial symbol systems (such as those of mathematics and physics) are merely subsets of natural language. The 'shape' of a symbol in a symbol system is arbitrary. It neither resembles nor is causally connected in any way to the object it represents, except by its users. It is merely part of a formal notational convention that its users, explicitly or implicitly, agree to adopt, whether it is a word in a language, a numeral of arithmetic, or a binary digit (0/1) in a low-level computational code.

How do symbols come to mean something? One candidate answer is 'by definition', but a definition just consists of further symbols: Where do *those* symbols get their meaning? Consider someone who does not speak any Chinese trying to find the meaning of a Chinese symbol in a Chinese–Chinese dictionary: all this person can do is search endlessly from symbol to meaningless symbol. How can the meanings of the symbols in a symbol system be grounded in something other than just further ungrounded symbols?

According to 'computationalists', cognition is computation (Pylyshyn 1984), implemented in a purely symbolic 'language of thought' (Fodor 1975). The meanings of the symbols arise somehow from the system's being connected in 'the right way' to the things in the world that its symbols stand for. But what is this 'right way'? And will

the properly 'connected' system still be a pure symbol system linked to the world, or will the *connecting* system now be part of a hybrid symbolic/non-symbolic 'language of thought'? In other words, is thought really just symbolic, or is it sensorimotor too, which is to say, robotic?

1.2. *Neural networks and categorical perception*

To 'discriminate' is to discern whether two patterns projected on to our sensory surfaces are the same or different. This does not require sophisticated symbolic operations, only a comparison between iconic representations, the internal analogue of the sensory patterns, perhaps by superimposing one on to the other. But, of course, to discriminate inputs is not yet to be able to say what those inputs are. To identify an object, one must somehow detect the invariant features in its iconic representations, the features that make them icons of that particular object (or kind of object) rather than another; the rest of the features must be ignored. The more abstract representations that this feature-filtering of the icons generates are categorical representations (Harnad 1987).

Categorical representations are still only sensory rather than symbolic, because they continue to preserve some of the 'shape' of the sensory projections, but this shape has been 'warped' in the service of categorization: the feature-filtering has compressed within-category differences and expanded between-category distances in similarity space so as to allow a reliable category boundary to separate members from non-members. This compression/expansion effect is called 'categorical perception' (Harnad 1987) and has been shown to occur in both human subjects (Goldstone 1994, Andrews *et al.* 1998, Pevtzow and Harnad 1997) and neural nets (Harnad *et al.* 1995, Tijsseling and Harnad 1997, Csato *et al.* submitted) during the course of category learning.

Categorical representations can be connected to labels, the names of the categories, but such labels still do not mean anything until they are combined to form propositions. Only at that stage do they become symbols, and the propositions of which they are components become symbolic representations (Harnad 1987).

One of the most natural capabilities of neural nets is category learning. Nets can be trained to detect the invariants in sensory input patterns that allow them to be sorted in a specified way. Once the patterns have been sorted, the category can be given a name. That name is then grounded in the system's autonomous capacity to pick out, from the 'shadow' it casts on its sensors, the thing (or kind of thing) in the world that the name refers to—without the mediation and interpretation of an external user.

The training of both neural nets and people to categorize through trial and error with corrective feedback has come to be called 'supervised learning', but we will refer to it here as the acquisition of categories through 'sensorimotor toil', to contrast it with a radically different way of acquiring categories, which we will refer to as 'symbolic theft'. Acquiring a category through 'toil' is based on learning through direct sensorimotor interaction with its members under the guidance of corrective feedback. The outcome is a new category and usually also a new name for it the name can then serve as a grounded elementary symbol. Acquiring a category through 'theft', in contrast, is based on symbols only, rather than on sensorimotor interaction with the things the symbols stand for: the category is merely *described* by a proposition composed of grounded symbols. (Why we refer to this as 'theft' will be explained in section 4 in the context of a hypothesis about the evolutionary role of language; for

now, just think of a 'stolen' category as one that is acquired without having to do any trial and error training with instances and feedback in order to get it; see Cangelosi and Harnad in press.)

Categories grounded directly through sensorimotor toil have iconic and categorical representations, whereas categories grounded indirectly through symbolic theft have symbolic representations consisting of their propositional descriptions in the form of symbol strings. The descriptions are Boolean or even more complex, quantified combinations of category names that are already grounded, either directly by toil, or indirectly by theft. In the simulations described later, we test what happens when nets that first acquire a set of categories through direct sensorimotor toil are then taught a higher-level category through symbolic theft (i.e. by being given a string of symbols that *tells* them what the higher-order category is). We shall show that sensorimotor grounding not only transfers to higher-order, symbol-based categories in a bottom-up fashion, but that the new, symbol-based categories also have some of the characteristic top-down effects of sensorimotor category learning, namely, that they deform or 'warp' internal similarity space in the service of categorization (for the warp effect on directly grounded categories see Tijsseling and Harnad 1997). This sensorimotor 'imprint' on symbolic thought may be what grounds it.

2. Method

2.1. *The stimulus set*

Our neural nets were trained to categorize and name 50×50 pixel images of circles, ellipses, squares and rectangles projected on to the receptive field of a 7×7 unit 'retina'. Once the net had grounded these four entry-level (E-Level) category names ('circle', 'ellipse', etc.) through direct trial and error experience supervised by corrective feedback ('toil'), it was taught the higher-level (H-Level) category 'symmetric/asymmetric' on the basis of strings of symbols alone ('theft').

A total of 292 stimuli was used (256 training, 32 test and four teaching input stimuli). The 256 stimuli consisted of four groups of circles, ellipses, squares and rectangles (figure 1). In each group there were 64 (8×8) stimuli that varied in size (eight sizes generated by reducing the diameter by 2 pixels) and retinal position (eight positions generated by shifting the centre of the figure by 1 pixel in the eight adjacent cells). The 32 test stimuli were also subdivided into four groups of eight stimuli each, one for each size. The position for each size was hence fixed, but it varied across sizes. The four teaching inputs were the largest instances of each shape (prototype).

2.2. *Neural networks*

Ten three-layered feed-forward nets differing in their random initial weights were exposed to the 256 training stimuli during the three learning stages. The input layers consisted of two groups of units: the retina, with 49 units (7×7) and the six linguistic/symbolic units (one each for the six category names: 'circle', 'ellipse', 'square', 'rectangle', 'symmetric' and 'asymmetric'). The hidden layer had five units receiving connections from both groups of input units. The output had the same organization as the 49 retinal units plus six linguistic/symbolic units (figure 2).

The coding of our linguistic/symbolic input and output units is localist. These are treated as having already been 'pre-categorized' in our simulations, and as taking the form of single unit activity. The units should be interpreted as being based on the

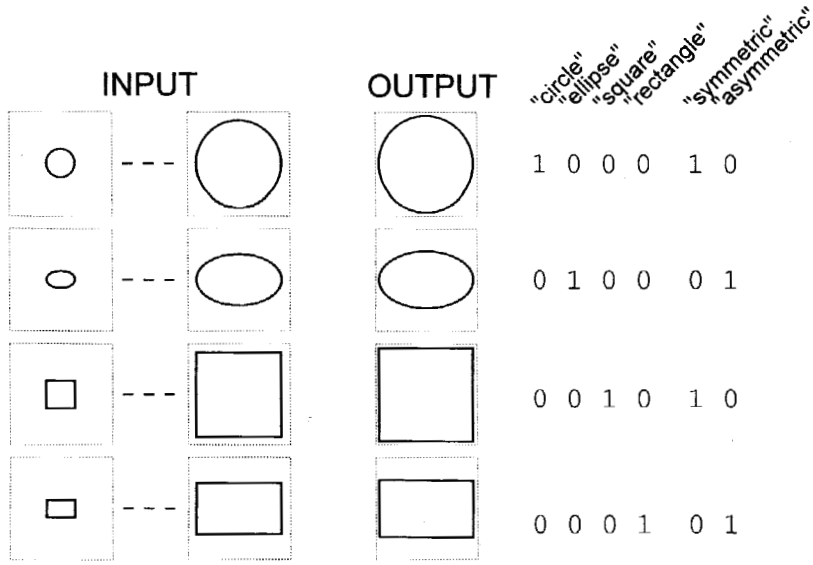

Figure 1. Stimulus set and localist coding of naming units.

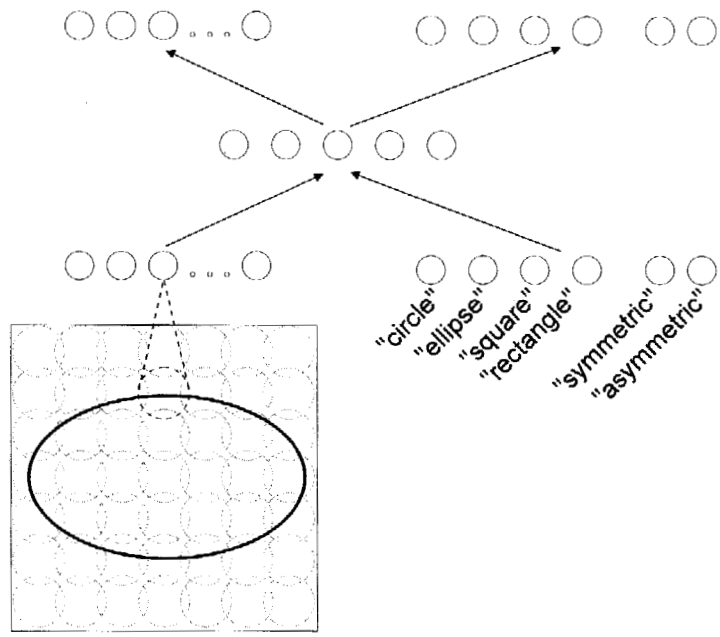

Figure 2. Neural network architecture and stimulus coding.

categorical invariants of the sensory projections of linguistic stimuli (e.g. phonemes, graphemes, etc.). The word 'circle' is to be interpreted as a speech-sound-string category (pronounced in various ways in different contexts), or as written letter-string category (and its contextual variants). Its sensory origins are no longer relevant when 'circle' is in turn used as an arbitrary symbol to stand, not for a class of speech or letter stimuli (things that sound like 'SURKUL' or look like 'CIRCLE'), but for an object (shaped like 'O'). In the present simulations, we accordingly use pre-categorized, localist coding to allow our nets to create links between arbitrary internal symbols and sensory inputs from the retina. The sensory origin of the internal symbols are simply assumed.

Whereas the coding of the symbolic units was localist (i.e. each unit was on when its corresponding label was active), the coding of the retinal units was more complex. We used the coding system of Jacobs an Kosslyn (1994) with retinal units receiving activation from their receptive fields in the 50×50 pixel matrix depicting each of the 256 geometric figures. The receptive field of one retinal unit was a circular area 11 (partially overlapping) pixels in diameter. Because of the receptive field overlap (3 pixels), there were 49 receptive fields arranged in seven columns by seven rows. The activation formula for the retinal units used the Gaussian distribution centred on the receptive field. Hence, pixels in the centre of the field contributed more to the activation of the retinal unit than those in the periphery.

The formula for the activation x of each Gaussian retinal unit is:

$$x = \sum_i \left(\frac{1}{\sigma^2} e - \frac{1}{2\sigma^2} \| p_i - \mu \|^2 \right)$$

where p is the location of the pixel, μ is the mean of the Gaussian unit and σ refers to the size of the receptive field. In our case $\sigma = 0.45$.

2.3. *Training procedure*

Our stimuli and our network architecture partially resemble those used in Plunkett *et al.*'s (1992) work on vocabulary growth. They used the task of learning to name random dot patterns to study the emergence of symbols, but their symbols are only for E-Level categories; they are not combined to denote H-Level categories, which is the crucial feature of the present study.

Our training procedure consisted of three stages: (1) prototype-based sorting; (2) E-Level naming and imitation learning; and (3) H-Level learning (figure 3). This resembles the sequence used in studies on object naming (Braine *et al.* 1990). Our nets use the error backpropagation algorithm because of its efficiency in learning categorization and naming. This is not a biologically plausible learning rule as implemented in artificial nets (the real nervous system does not have antidromic activity of this kind, or an internal supervisor to orchestrate it), but it is easy to conceive of more plausible ways of implementing essentially the same kind of algorithm biologically (with reciprocal feed-forward connections, orchestrated by the reinforcing effects of the immediate or belated Skinnerian 'consequences' of outputs), and implementation is not the issue here. In further developing the model, more 'plausible' learning algorithms for neural networks, such as Bayesian learning (Goodman *et al.* 1992), will also be analysed.

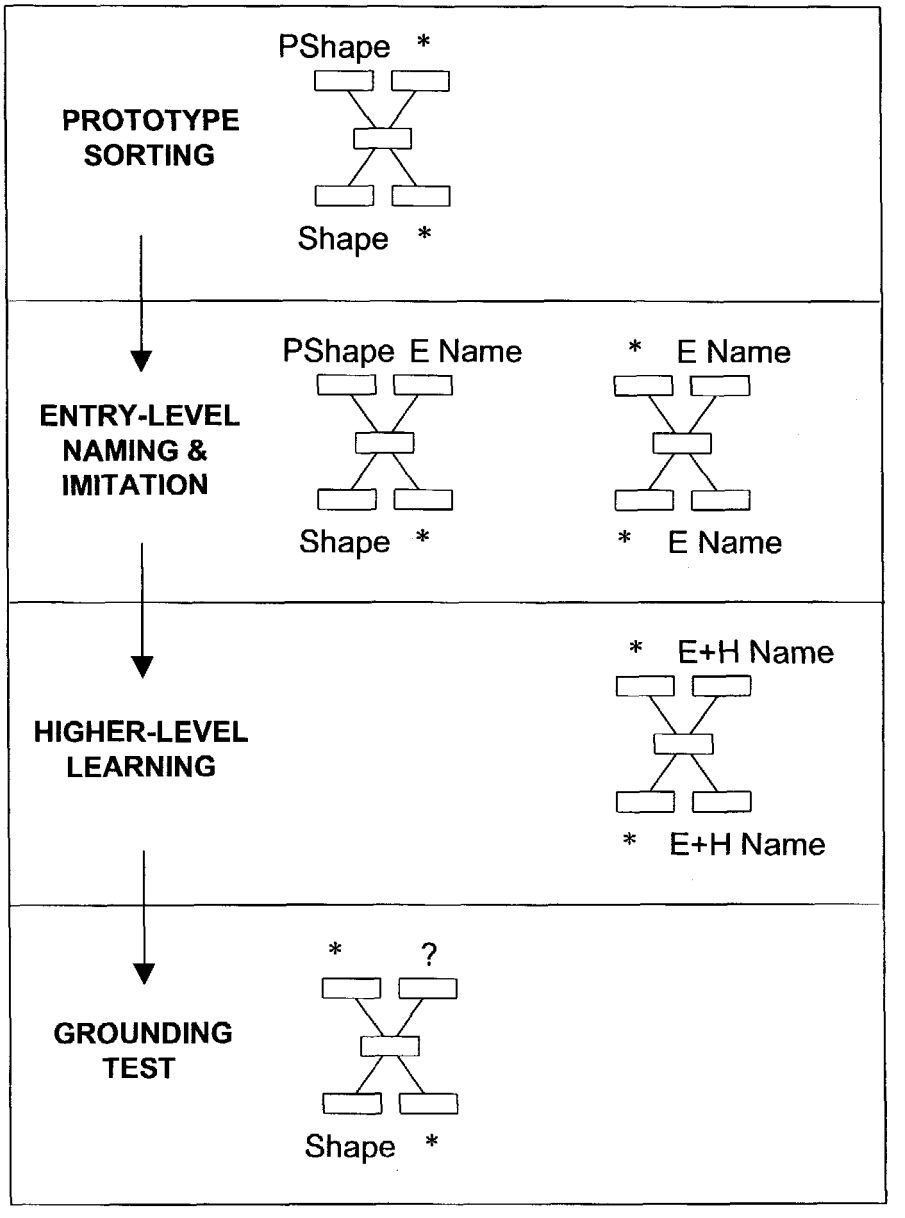

Figure 3. Input and output in learning and test stages. Nets on the left perform naming, those on the right, imitation learning. Absence of input or output is indicated by an asterisk. Null input to the input units is all zeros. Null output corresponds to setting the units' error to zero, so that no changes occur in the connection between them and the hidden layer.

2.3.1. *Prototype-based sorting.* The net was first trained, via back propagation, to sort the 256 training stimuli into the four categories (64 stimuli each) by producing as output the 'prototype' of each category in the form of the largest circle, ellipse, square or rectangle (coded in the same way as the rest of the stimuli).

2.3.2. *Entry-level naming and imitation.* The net next learned to respond to each stimulus by producing both its prototype shape and its category name. An imitation task was interposed between each trial of the naming task, consisting of an extra activation cycle to allow the net to 'practise' on the category name learned in the preceding naming trial. These paired learning cycles strengthen the mapping between retina and linguistic input units and the linguistic output nodes.

2.3.3. *Higher-level learning.* H-Level categories such as 'symmetric/asymmetric' can be learned in one of two ways; either (1) through naming directly from the retinal input, as with the E-Level categories ('sensorimotor toil'); or (2) from Boolean combinations of the grounded category names ('symbolic theft'). We investigated (2): the net received as input the conjunction of the grounded name plus a new name (either 'asymmetric' or 'symmetric') and was required, through error-correcting feedback, to generate both names as output. (Simultaneous presentation of E-Level and H-Level names makes it unnecessary to use a recurrent network to learn the association.) A net that learns that two different grounded names, 'circle' and 'square', are always combined with the same new name, 'symmetric', should be able to name a circle both 'circle' on the basis of the prior sensorimotor grounding, and 'symmetric' on the basis of the new symbolic grounding. This learning task is based on imitation rather than naming, because networks learn to map the combination of linguistic units into linguistic output units only.

2.4. *Backpropagation*

One learning epoch consists of the presentation of all 256 training stimuli. The first learning stage (prototype-based categorization) consists of 10 000 epochs. This is necessary because of the large number of retinal units (49) that need to be trained. The two E-Level and H-Level naming tasks last 2000 and 1000 epochs, respectively. Each learning condition is replicated with 10 nets. In the prototype-sorting task 10 nets having different initial random weights are used (in the range ± 1). In the following learning stages, the connection weights of the previous trained nets are used. The backpropagation learning rate for all learning tasks is 0.01. The node activation follows the standard sigmoid function, with the activation range of 0–1. The neural network software package TLEARN (http://crl.ucsd.edu) was used.

3. Results

3.1. *Learning error and generalization*

All 10 nets learned the three tasks successfully. The final sum square error for the first stage, prototype-based categorization, was 0.09 after 10 000 epochs (figure 4(a)). This error is not very low, but in most of the nets it was less than 0.05; it was only in a few that it was about 0.1. Nevertheless, the categorization of all the stimuli was unambiguous, that is, each shape was always categorized correctly; the errors pertain only to some imperfections in generating the right prototype (the largest figure for each shape) in this hybrid iconic/categorical task. The same level of error was attained in the E-Level naming stage, with a final error of 0.08 (figure 4(b)).

The error in the H-Level learning was very low, about 0.01. In fact only the error in the name units is computed. The pattern in all three conditions is a rapid initial decrease in the early training epochs. After that, the error decreases very little (figure 4(c)).

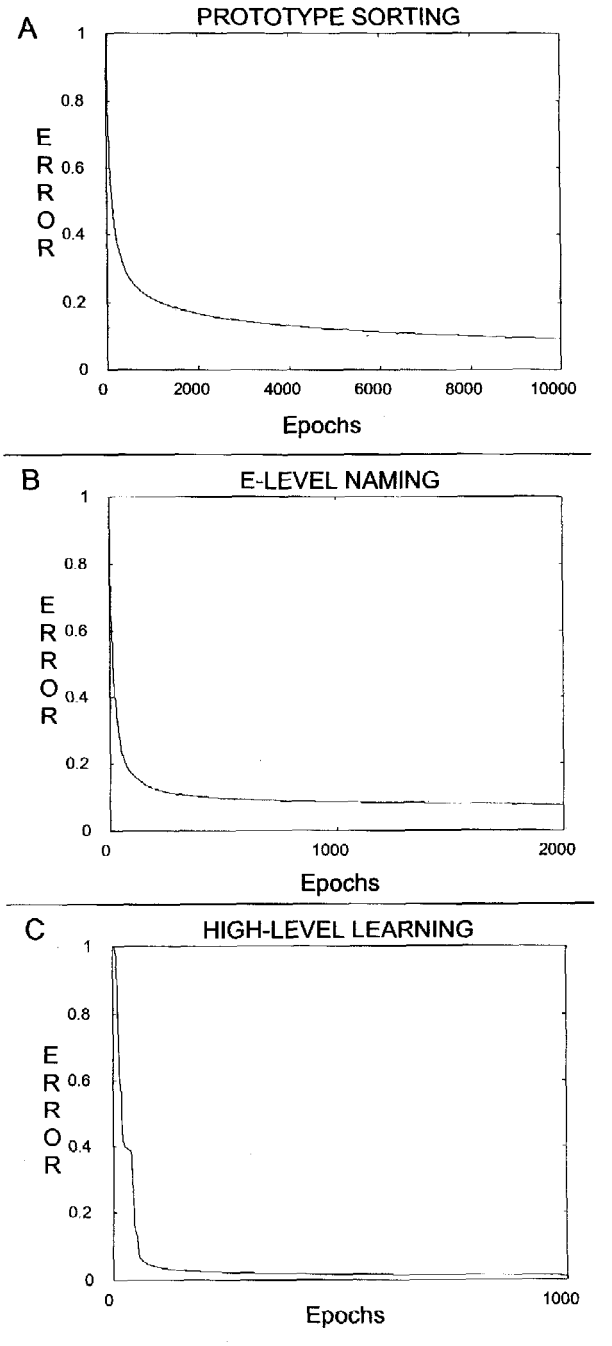

Figure 4. Learning error for (a) the prototype sorting, (b) Entry-Level naming and (c) H-Level learning.

The results of the generalization test showed that after the prototype learning the 32 test stimuli were properly categorized in the four E-Level categories. The same good generalization performance was obtained in the other two learning stages.

3.2. *Categorical perception effects*

At the level of the hidden units, the net builds categorical representations which must sort each icon reliably and correctly into its own category. This can be thought of as a featurefilter that reduces the category confusability by decreasing the within-category differences among the icons and increasing the between-category difference as needed to master reliably the sorting task (Harnad 1987).

For the three learning stages of each of the 10 nets, we computed means and variances in the Euclidean distances for all 256 representations in the five-dimensional hidden unit activation space. We first computed the central (mean) points for the four categories. These were then used to compute both within- and between-category distances. The within-category variance is a measure of the distance between each of the 64 points and its respective category mean. There is a clear decrease in within-category variance from before prototype learning (0.315) to after (0.2). That is, during the course of the prototype learning the 64 points of each category move closer to one another [MANOVA: $F(9,1)=6.12, p<0.035$].

A further within-category compression from prototype matching (0.2) to naming (0.172) shows the effects of arbitrary naming on categorical representations (prototypes are analogue, names are arbitrary) [$F(1,1)=14.9, p<0.004$] (figure 5).

The same effects are observed with the between-category differences (the distances between the centres of the four categories). From before learning (0.15) to prototype matching (1.14), the average between-category distance increases for all six pairwise comparisons between the four category means [$F(9,1)=1034, p< 0.0001$)]. A further but smaller increase occurs with naming [1.16; $F(9,1)=28, p< 0.0001$]. Figure 6 shows the between-category distances for a sample of pairwise comparisons.

After prototype-based categorization, the within-category-to-be distances between the two symmetric shapes (circle [C] versus square [S], 0.82) and the two asymmetric ones (ellipse [E] versus rectangle [R], 0.91) were smaller than the distances between the four between-category-to-be pairs (C versus E and C versus R both, 1.12; S versus R, 1.32; E versus S, 1.42; figure 6). This means that when the four prototype-based categories are formed, the two symmetric pairs and the two asymmetric ones are

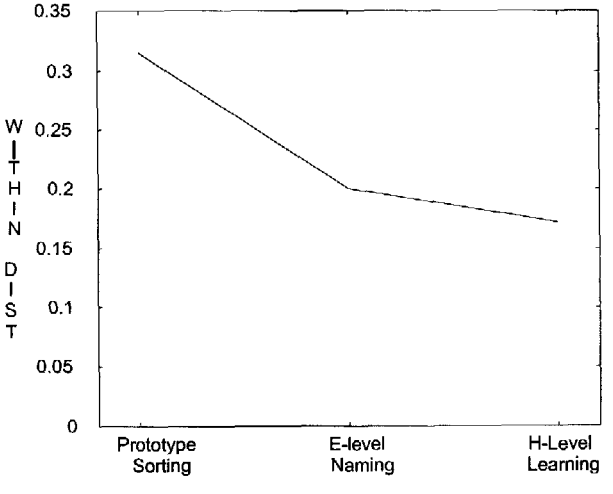

Figure 5. Average within-category distances.

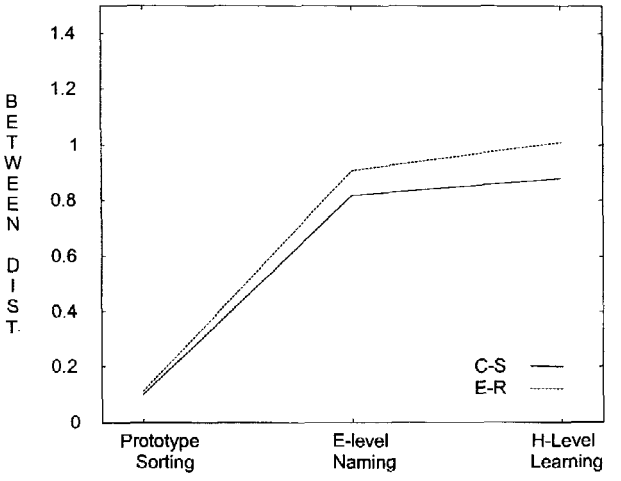

Figure 6. Between-category distances for the pairs circles–squares and ellipses–rectangles.

already closer to one another than the between-category pairs are. The higher-order categorization task starts with this initial similarity structure.

In this sense, the symmetric/asymmetric distinction can be thought of as a somewhat 'prepared' category, as there is already an intrinsic bias in their similarity structure. A harder task would be one in which the within and between distances for the (future) categories are initially equal, but if the distances are also small, this can run the risk of making the categorization task unlearnable (Pevtzow and Harnad 1997).

3.3. *Grounding transfer*

We next tested whether grounding could be 'transferred' from directly grounded names to H-Level ones. Can a net that has learned the category 'symmetric' indirectly through symbolic theft generalize it to the direct retinal input? To test this, after the H-Level training we presented the retinal stimuli alone (see figure 3, last column) and computed the frequency of correct responses for the E-Level names and H-Level names (criterion for all conditions: correct bit > 0.5, others < 0.5).

Table 1 reports per cent correct for the E-Level names (left column for each net) and the H-Level names (right column). A net's success criterion was at least 50% correct. Nine of the 10 nets met this criterion for Entry-Level names and eight did for H-Level names (see italic columns in table 1). Assuming chance to be 0.5, the binomial probability of 9/10 nets successful by chance 0.0098 (and for 8/10, 0.044). Hence, the E-Level grounding successfully transferred to the H-Level categorization.

We also did a control to test whether this outcome depended on some uncontrolled variable rather than grounding transfer. This could be tested by eliminating the grounding stage for the E-Level categories (i.e. no E-Level naming and imitation) or by randomizing the grounding of the E-Level categories. Both methods are valid, but the first is preferable because it is a more thorough way to eliminate the grounding of low-level categories, on which we infer that the grounding transfer to H-Level categories is based. For the control we repeated the training with 10 new nets. Now the E-Level learning stage was skipped and H-Level learning followed immediately after prototype learning followed immediately after prototype learning (figure 7).

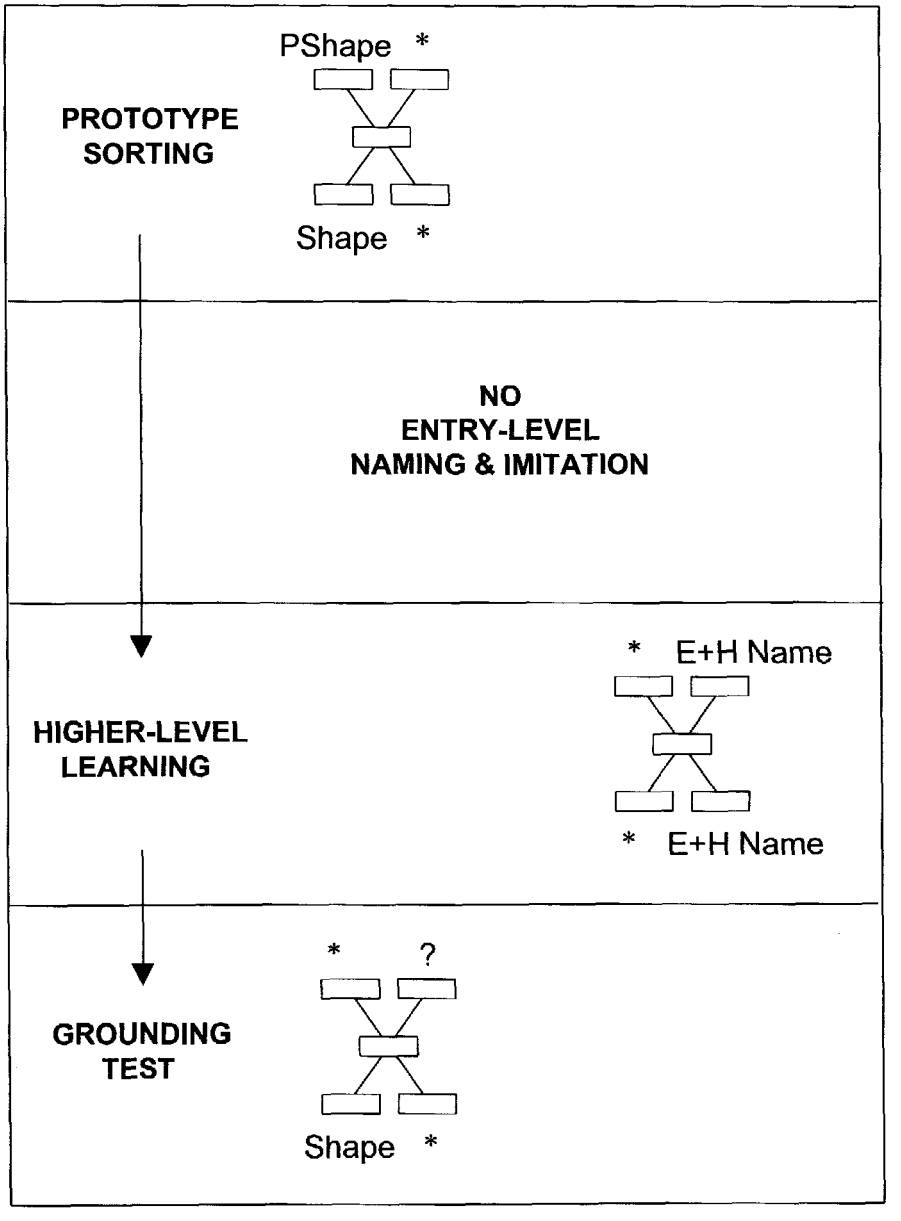

Figure 7. Neural network input and output in the control simulations.

The results are shown in table 2. Based on the same criterion as in table 1, none of the 10 nets was successful in E-Level naming, and only three were successful in H-Level naming.

We can also count the total number of correct responses instead of the number of correct nets. Since the total number of naming trials is high (2560 for E-Level plus H-Level), we can use the Gaussian distribution and compute the z value for the difference between the two probabilities. For E-Level naming, the per cent correct is

Table 1. Per cent correct in grounding transfer test.[a]

	Net 1		**Net 2**		**Net 3**		**Net 4**		**Net 5**		**Net 6**		**Net 7**		**Net 8**		**Net 9**		**Net 10**	
	E	*H*	*E*	*H*	*E*	*H*	*E*	*H*	E	*H*	*E*	*H*	*E*	*H*	*E*	H	*E*	*H*	*E*	H
C	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	100	100	100	100
E	100	100	75	100	100	100	100	100	12	100	100	100	100	100	100	100	100	100	100	37
S	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	0	100	100	100	100
R	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	37

[a] For each net, the number on the left is correct responses for E-Level names and on the right for H-Level names. Rows are for the 64 circles (C), ellipses (E), squares (S) and rectangles (R). Italic cells indicate success in E-Level (E) or H-Level (H) categorization in the grounding transfer (criterion: at least 50%).

Table 2. Per cent correct in grounding transfer controls.[a]

	Net 1		**Net 2**		**Net 3**		**Net 4**		**Net 5**		**Net 6**		**Net 7**		**Net 8**		**Net 9**		**Net 10**	
	E	H	E	*H*	E	H	E	H	E	H	E	*H*	E	H	E	H	E	H	E	*H*
C	100	100	0	100	0	8	0	0	0	100	0	100	0	100	100	100	100	100	100	100
E	0	0	0	100	0	100	100	100	0	0	0	100	0	0	0	0	0	0	0	100
S	0	100	100	100	0	0	0	0	0	100	100	100	0	100	0	100	0	100	0	100
R	0	0	0	87	0	58	0	100	0	0	0	100	0	0	0	0	0	0	0	100

[a] For each net, the number on the left is correct responses for E-Level names and on the right for H-Level names. Rows are for thhe 64 circles (C), ellipses (E), squares (S) and rectangles (R). Italic cells indicate the nets that succeeded in or H-Level (H) grounding transfer (criterion: at least 50% correct).

97% for the grounding transfer test and 15% for the controls (prototype learning only). For H-Level naming, the per cent correct is 92%, compared with 63% for the controls. Here we will compare only the probabilities for H-Level naming. z is computed using the formula:

$$z = \frac{P - P_2}{\sqrt{\left(\frac{P_1 * Q_1}{N} + \frac{P_2 * Q_2}{N}\right)}}$$

where P_1 and P_2 are, respectively, the two positive probabilities in the test and counter-test, and Q_1 and Q_2 are the reciprocal percentages ($Q = 100 - P$). N is 2560. For the difference between the two H-Level probabilities, z is 30.3 (N=2560; p<0.0001), confirming that prior direct grounding is essential for grounding transfer.

The results of the grounding test show that the proposed sequence of learning tasks allows nets to learn H-Level categories via either imitation learning or name composition. Because the names are grounded directly in retinal projections, the new symbols inherit this grounding. After H-Level learning, the retinal inputs activate the correct, symmetric category. But what is the mechanism that allows such grounding transfer to occur? How are categorical representations involved in this process? These questions can be answered by analysing the nets' hidden representations.

We examined the hidden representations produced by nets after each naming and imitation learning stage. Figure 8 shows the hidden unit activations for one net (black square size proportional to activation). Activation values for the four categories (square, circle, rectangle, ellipse) are reported. For each category, the activation used is the average for the 64 stimuli of each shape. We have already noted that, owing to categorical perception effects, the hidden representations of the stimuli belonging to a category are very similar and have low within-category distances (section 3.2).

In the Entry-Level tasks—naming (left group, top window) and imitation (right group, top window)—three hidden units (h3–h5) have very similar activation patterns in both tasks, whereas two (h1–h2) have different patterns. What the two patterns have in common is their contribution to the four linguistic output units (the two high-order linguistic units are not yet used). Wherein they differ is the activation of the retinal output units. The three units with similar activations (h3–h5) will effectively influence the linguistic output units. The two that differ (h1–h2) will control the activation of the retinal output units.

During the H-Level learning, the net is trained to activate the two linguistic output units for the symmetry/asymmetry categories. The middle window of figure 8 shows that after H-Learning the net keeps the same hidden unit activation pattern as in the previous E-Level imitation, but uses it for adjusting the weights of the connections from the hidden units to the two new output units. At the beginning of the H-Learning these weights are random and near 1, whereas at the end they differentiate. Figure 9 shows that at the end of H-Learning these 10 hidden-output connection weights change and that the two weights from the third hidden unit h3 are very high and have opposite sign. This unit is contributing in a significant way to the activation of the linguistic unit for 'symmetric' (weight +9). At the same time, h3 is inhibiting (weight

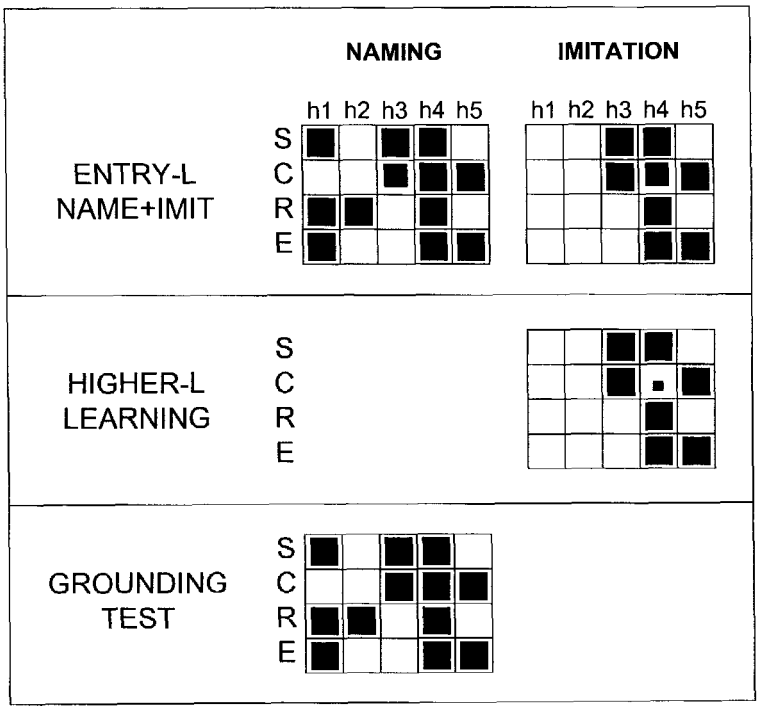

Figure 8. Hidden unit activations for the four categories (S=square, C=circle, R=rectangle, E=ellipse) in the two learning tasks and the symbol-grounding test. The position of the four activation groups is as in figure 3. For each category, the activation is the average for the 64 stimuli of each shape. The size of the black square is proportional to the average activation (biggest square for activation = 1, empty white square for activation = 0). Note that only the third hidden unit can discriminate between symmetric (S, C) and asymmetric (R, E) shapes. See text for full explanation.

–9), the output unit for the category 'asymmetric'. The activation of h3 is maximal for the two symmetric shapes, square and circle, and zero for the asymmetric shapes.

Analysis of the hidden unit activations during the symbol grounding test (figure 8, bottom window) reveals that the activation produced by the retinal input has not changed much from what it was in E-Naming. The pattern of units h3–h5 is very similar to E-Learning. In particular, h3 is what makes the discrimination between the symmetric and asymmetric shapes possible. Its activation, in conjunction with the newly learned weights connecting it to the two high-order linguistic units, allows the net to turn on the right output unit.

Analysis of the three nets that did not pass the grounding transfer test reveals that their hidden representations are more distributed than in the other nets. There are more units whose activations differ for naming and imitation. It is accordingly more difficult for the net to find a good set of hidden-output connection weights that can discriminate between symmetric and asymmetric shapes with either the retinal or the linguistic input.

What this analysis tells us is that the transfer of grounding from the low-level categories to the higher-level ones is mediate by the hidden representations. Because of categorical representation effects, these units partition the net's representational space into distinct regions, one per category. These regions tend to have high

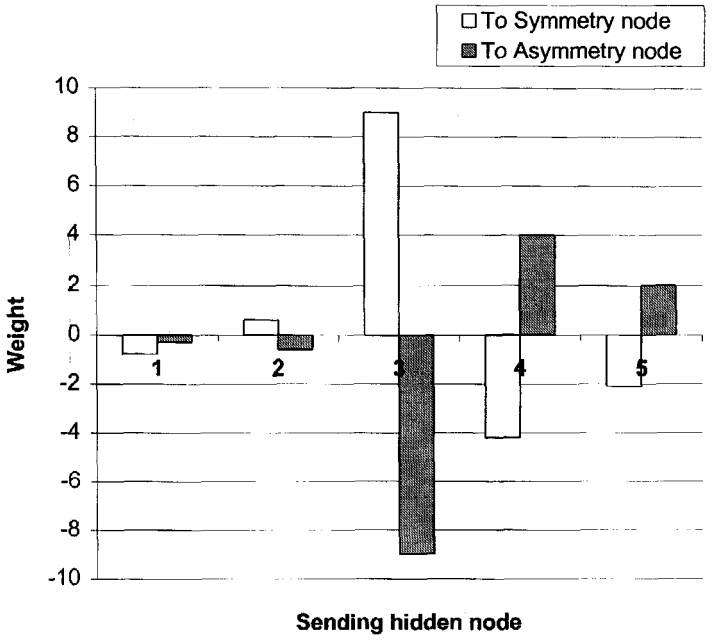

Figure 9. Weights of connections between the five hidden units and the two output units for the higher-order (symmetric/asymmetric) categorization. Note that the highly contrastive weights from the third hidden unit are the ones mainly responsible for the differentiation of the two output units categories (see figure 8). See text for full explanation.

between-category distances. Imitation learning creates links between well-differentiated categorical representations and discrete symbols. When these symbols are combined, they inherit their links to low-level categorical representations.

3.4. *Extending the simulation from extensional to intensional categories*[2]

To control for the possibility that our findings applied only to conjunctions of individuals and conjunctions of symbols, we replicated and extended the grounding transfer test from merely *extensional* H-Level categories (based on Boolean combinations of individuals) to *intensional* ones (based on Boolean combinations of features) using a second set of stimuli: animal shapes (horse and turtle) and texture features (stripes and spots) see (see figure 10). With this combination of individuals and features (e.g. horse and stripes) as E-Level stimuli (rather than only individuals and individuals, as in the prior simulations), it was possible to teach the H-Level names by combining them into Boolean descriptions of new H-Level individuals (e.g. zebras). The H-Level 'zebra' name was trained in one stage using the name conjunction: 'horse + stripes'. The test for the H-Level 'zebra' category was when whether the zebra shape (an image of a striped horse) could be correctly named. In the prior shape experiment, the H-Level names had been derived by conjoining two *individuals* (e.g. circle and square) to learn a new abstract feature category (symmetric). The training had been in two stages, one for learning that 'circle' was 'symmetric' and the other for learning that 'square' was likewise 'symmetric'. The grounding transfer test was also in two stages, one for each symmetric shape. The zebra simulations used the same method

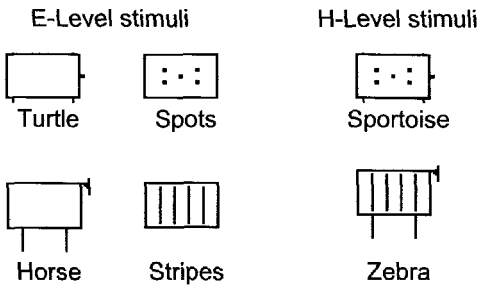

Figure 10. Stimuli used in the zebra simulations.

as in section 2, except that (apart from the new stimuli) the H-Level training and testing involved only one stage for each H-Level category ('horse' + 'stripes' = 'zebra', 'turtle' + 'spots' = 'sportoise').

Tables 3 and 4 report per cent correct for grounding transfer for the H-Level stimuli with the standard and control nets (omitting the E-Level naming), respectively. Eight of the 10 experimental nets but none of the 10 control nets were successful.

The per cent correct for instances of naming (rather than of successful networks) was 83% in the experimental condition and 7% in the control (N=900). The difference was highly significant.

These results are similar to those for the shape simulations. Only the nets that learned the direct grounding of the E-Level names ('horse' and 'stripes' were able to ground the H-Level names, correctly naming the zebra shape they had never encountered during training. The control nets could not name the H-Level categories because they had no grounding for the E-Level names.

Table 3. Per cent correct in grounding transfer test for zebra simulations.[a]

	n1	***n2***	***n3***	***n4***	***n5***	**n6**	***n7***	**n8**	***n9***	***n10***
Zebra	62	100	100	100	100	20	100	33	66	100
'Sportoise'	100	100	100	100	100	0	100	100	100	100

[a] Numbers refer to *H-Level* names. Italic cells refer to the eight successful H-Level nets in the grounding transfer (criterion: at least 50% correct).

Table 4. Per cent correct in grounding transfer controls for zebra series.[a]

	n1	**n2**	**n3**	**n4**	**n5**	**n6**	**n7**	**n8**	**n9**	**n10**
Zebra	100	42	67	100	53	100	20	30	0	100
'Sportoise'	0	100	0	0	0	0	0	0	0	0

[a] Numbers refer to *H-Level* names. No net met the 50% success criterion.

4. Discussion

These results confirm and extend findings with other connectionist models of categorical perception (Harnad *et al.* 1995, Csato *et al.* submitted). When trained to

categorize, neural nets build internal representations that compress differences within categories and expand them between. These data are also consistent with related findings in a connectionist model with localist encoding of perceptual features (Cangelosi and Harnad in press).

Ours is a 'toy' model, but it is hoped that the findings will contribute toward constructing hybrid models that are immune to the symbol-grounding problem. Names (symbols) are grounded via net-based connections to the sensory projections of the objects they stand for. The grounding of E-Level symbols can then be transferred to further symbols through Boolean combinations of symbols expressing propositions.

The control simulation showed that direct grounding of at least some names is necessary. We grounded the names of the four E-Level shapes directly in their retinal projections. The same retinal projections then also activate the new H-Level name, 'symmetric', through their indirect grounding. Circles and squares activate some common categorical representation in the hidden layer that in turn activates 'symmetric'; rectangles and ellipses activate 'asymmetric'.

The conditions that lead to grounding transfer require further simulations and analysis. E-Level naming proved sufficient for grounding transfer in most of the nets (80%). Thirty per cent of the control nets were likewise able to transfer grounding to the H-Level names, probably because compression/separation induced by their training in E-Level categorization and naming reduced the variability in the hidden layer. This can be tested with further randomized and biased control conditions.

During the prototype-based categorization, the nets learn to produce four separable hidden representations for each of the categories (64 shapes in each), with very similar activation patterns within categories and very different ones between. In addition, there is already some compression of the symmetric and asymmetric shapes at the prototype level. These 'head-starts' in similarity space, together with the analysis of hidden representations, explain how the nets managed to master the H-Level naming without being taught the E-Level naming: they already had the categories, just not yet their names. And so it may well be with many categories; random seeding is an unlikely model for the initial conditions of biological categorization.

Some categories will already be 'prepared' by evolution; others will be acquired on the basis of shared iconic or functional responses, rather than arbitrary naming. But when naming does occur, it will benefit from following these pre-existing gradients or boundaries in similarity space—as long as the requisite new category goes with them rather than against them. This too is a form of grounding transfer.

This explanation is confirmed by the analysis of the naming errors for the E-Level names in the control condition. Nets named only a very low proportion of shapes correctly in this condition (15%) because it becomes harder to be right by chance as the number of bits increases. With two possibilities, symmetric/asymmetric, nets can achieve 50% by chance, but with four (circle, square, etc.), chance is 25%. Moreover, the E-Level control errors reveal that circles are often called 'circle + square' or simply 'square', and conversely. This interconfusability of circles and squares is what one would expect from their close categorical representations.

Our model for categorization and naming can also test hypotheses about the origin of cognition and of language (Cangelosi and Parisi 1998). The proposition describing the H-Level categories in the present simulation ('circle [is] symmetric' 'ellipse [is] asymmetric', etc.) came as a kind of 'Deus ex Machina': the E-Level categories could have been acquired by ordinary trial and error reinforcement in the world, through

learning supervised by the consequences of categorizing and miscategorizing. This is what we have called learning by 'sensorimotor toil'. But in a realistic world the symbolic propositions on which the H-Level categories were based would have had to come from someone who already knew what was what.

To get categories by 'symbolic theft', then, is to get them on the basis of the grounded knowledge of others, transferred to us via symbolic propositions whose terms—all but one—are already grounded for us too. This new way of acquiring categories spares us a great deal of sensorimotor toil. (Imagine if everything we learned from books and lectures instead had to be learned directly through trial and error experience!) Hence, gaining intellectual goods via hearsay is a kind of theft, but in most cases it is also a victimless crime, as the provider of the knowledge loses nothing by giving it away; perhaps it is more like a form of reciprocal altruism. There are exceptions, such as when the knowledge concerns scarce resources for which there is competition (Cangelosi and Harnad in press). But a paradigmatic example of victimless nature of linguistic theft would be this article itself, which, if its reader has gained anything from it, certainly leaves the authors none the worse off for it.

Notes

1. A preliminary version of this work was presented at ICANN98, the 1998 International Conference on Artificial Neural Networks, Skovde, September 1998.
2. A set or category can be defined by its 'extension' or its 'intension'. Its extension is its membership. One way to define the set would hence be to enumerate exhaustively all of its members, one by one. (The extension of the set of 'even numbers' is 2, 4, 6, 8 [.]). A category that was internally represented extensionally would consist of internal representations of each and every member ('instance-based' representation). The intension of a set consists of the properties of its members that determine that they belong to that set. (The intension of the set of 'even numbers' is those numbers that are divisible without remainder by 2.) A category that was internally represented intensionally would consist of an internal detector of those properties or 'invariants' that make an instance a member of that set. An extensional internal representation of the category 'circle' would consist of all instances of circular shapes. An intensional internal representation of that same category would consist of a detector for equidistance of a continuous set of points around a midpoint.

References

Andrews, J., Livingston, K., and Harnad, S., 1998, Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **24**: 732–753.

Braine, D. S., Brody, R., and Brooks, P. J., 1990, Exploring language acquisition in children with miniature artificial language: effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language*, **29**: 591–610.

Cangelosi, A., and Harnad, S., in press, The adaptive advantage of symbolic theft over sensorimotor toil: grounding language in perceptual categories. *Evolution of Communication. http://cogsci.soton.ac.uk/harnad/Papers/Harnad/harnad98,theft.toil.html*

Cangelosi, A., and Parisi, D., 1998, The evolution of a 'language' in an evolving population of neural nets. *Connection Science*, **10**: 83–97.

Csato, L., Kovacs, G., Harnad, S., Pevtzow, R., and Lorincz, A., submitted, Category learning, categorisation difficulty and categorical perception: computational modules and behavioural evidence. *Connection Science*.

Fodor, J. A., 1975, *The Language of Thought* (New York: Thomas Y. Crowell).

Goldstone, R., 1994, Influences of categorization of perceptual discrimination. *Journal of Experimental Psychology: General*, **123**: 178–200.

Goodman, R. M., Higgins, C. M., Miller, J. W., and Smyth, P., 1992, Rule-based neural networks for classification and probability estimation. *Neural Computation*, **4**: 781–804.

Hadley, R. F., 1994, Systematicity in connectionist language learning. *Mind and Language*, **9**.

Hadley, R. F., and Cardei, V. C., 1999, Language acquisition from sparse input withouth error feedback. *Neural Networks*, **12**: 217–235.

Harnad, S. (ed.), 1987, *Categorical Perception: The Groundwork of Cognition* (New York: Cambridge University Press).

Harnad, S., 1990, The symbol grounding problem. *Physica D*, **42**: 335–346.

Harnad, S., 1993, Grounding symbols in the analog world with neural nets. *Think*, **2**, 12–78.

Harnad, S., 1995, Grounding symbolic capacity in robotic capacity. In: L. Steels and R. Brooks (eds) *The Artificial Life Route to Artificial Intelligence: Building Embodied Situated Agents* (New Haven: Lawrence Erlbaum), pp. 277–286.

Harnad, S., 1996, The origin of words: a psychophysical hypothesis. In B. Velichkovsky and D. Rumbaugh (eds) *Communicating Meaning: Evolution and Development of Language* (New Jersey: Erlbaum), pp. 27–44.

Harnad, S., Hanson, S. J., and Lubin, J., 1995, Learned categorical perception in neural nets: implications for symbol grounding. In V. Honavar and L. Uhr (eds) *Symbol Processors and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling: Steps Toward Principled Integration* (New York: Academic Press), pp. 191–206.

Jacobs, R. A., and Kosslyn, S. M., 1994, Encoding shape and spatial relations: the role of receptive field size in coordinating complementary representations. *Cognitive Science*, **18**: 361–386.

Pevtzow, R. and Harnad, S., 1997, Warping similarity space in category learning by human subjects: the role of task difficulty. In M. Ramscar, U. Hahn, E. Cambouropolos and H. Pain (eds) *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization*, Department of Artificial Intelligence, Edinburgh University, pp. 189–195.

Plunkett, K., Sinha, C., Moller, M. F., and Strandsry, O., 1992, Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science*, **4**: 293–312.

Pylyshyn, Z. W., 1984, *Computation and Cognition* (Cambridge MA: MIT/Bradford).

Tijsseling, A., and Harnad, S., 1997, Warping similarity space in category learning by backprop nets. In M. Ramscar, U. Hahn, E. Cambouropolos and H. Pain (eds) *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization*, Department of Artificial Intelligence, Edinburgh University, pp. 263–269.

van Gelder, T., 1990, Compositionality: a connectionist variation on classical theme. *Cognitive Science*, **14**: 335–364.