From Robotic Toil to Symbolic Theft: Grounding Transfer from Entry-Level to Higher-Level Categories

Angelo Cangelosi
Centre for Neural and Adaptive Systems
University of Plymouth (United Kingdom)

Alberto Greco
Department of Anthropological and Psychological Sciences
University of Genoa (Italy)

Stevan Harnad
Cognitive Science Centre
University of Southampton (United Kingdom)

Corresponding author:

Angelo Cangelosi
School of Computing
University of Plymouth
Drake Circus
Plymouth PL4 8AA (UK)

Email: angelo@soc.plym.ac.uk

Tel. +44 1752 232559
Fax +44 1752 232540

Running title: Symbol grounding transfer in neural nets

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

Note: Submitted to *Connection Scienc*. A preliminary version of this work was presented at ICANN98, the 1998 International Conference on Artificial Neural Networks, Skovde, September 1998.

From Robotic Toil to Symbolic Theft: Grounding Transfer from Entry-Level to Higher-Level Categories

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 analog sensorimotor projections to arbitrary symbolic representations via learned category-invariance detectors in a hybrid symbolic/nonsymbolic system. Our nets are trained to categorize and name 50x50 pixel images (e.g., circles, ellipses, squares and rectangles) projected onto the receptive field of a 7x7 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" vs. "asymmetric") can 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 <u>describing</u> the higher-order category ("symbolic theft"). We analyze 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.

From Robotic Toil to Symbolic Theft: Grounding Transfer from Entry-Level to Higher-Level Categories

1. Introduction

The nonlinguistic or prelinguistic part of us is purely robotic, which is to say purely sensorimotor (Harnad 1995a,b). 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 nonlinguistic interaction with those objects, events and states -- is the traditional domain of robotics: vision, locomotion, object recognition and manipulation. But even in modeling 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, otherwise known as computational (Pylyshyn 1984).

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.

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 with truth-values (true or false). 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. It is merely part of a formal notational convention that its users have agreed 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 <u>those</u> symbols get their meaning? Consider someone who speaks no Chinese trying to find the meaning of a Chinese symbol in a Chinese-Chinese dictionary: All they 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 <u>connecting</u> system now be part of a hybrid symbolic/nonsymbolic "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 onto our sensory surfaces are the same or different. This does not require sophisticated symbolic operations, only a comparison between <u>iconic representations</u>, the internal analogs of the sensory patterns, perhaps by superimposing one onto the other.

But to discriminate inputs is not yet 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 <u>categorical representations</u> (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 nonmembers. 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 & Harnad 1998) and neural nets (Harnad et al. 1991, 1995; Tijsseling & 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 <u>propositions</u> which can be true or false. Only at that stage do they become symbols, and the propositions of which they are components become <u>symbolic representations</u> (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 why. 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 & Harnad 1998.)

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 below, 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 will 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 categorisation (Tijsseling & 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 by 50 pixel images of circles, ellipses, squares and rectangles projected onto the receptive field of a 7 by 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 were used (256 training, 32 test, and 4 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 by 8) stimuli that varied in size (8 sizes generated by reducing the diameter by two pixels) and retinal position (8 positions generated by shifting the center 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).

< Figure 1 about here >

2.2 Neural networks

Ten 3-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 by 7) and the 6 linguistic 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 6 symbolic-name units.

< Figure 2 about here >

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 & Kosslyn (1994) with retinal units receiving activation from their receptive fields in the 50 by 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 7 columns by 7 rows. The activation formula for the retinal units used the Gaussian distribution centered on the receptive field. Hence pixels in the center 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, m is the mean of the Gaussian unit, and s refers to the size of the receptive field. In our case s = .45

2.3 Training procedure

The training procedure consisted of three stages for category learning and naming: (1) prototype-based sorting, (2) E-Level naming, and (3) H-Level naming. In all the learning stages in which names were used, there was a further imitation task (Figure 3).

< Figure 3 about here >

2.3.1 *Prototype-Based Sorting*. The net was first trained, via backpropagation, 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 the same way as the rest of the stimuli).

2.3.2 *Entry-Level Naming*. The net next learned to respond to each stimulus by producing both its prototype shape and its category name.

2.3.3 *Higher-Level Naming*. H-Level categories such as "symmetric/ asymmetric" can be learned in one of two ways, either (1) 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.

2.3.4 *Imitation learning*. The imitation task is alternated with each trial of the naming task. It consists of an extra activation cycle that allows the net to "practice" on the category name learned in the previous naming cycle.

2.4 Backpropagation

One learning epoch consists in 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 require 2,000 and 1000 epochs, respectively. Each learning condition is replicated with 10 nets having different random weight initializations. The 10 replications of the second and third stages, however, are constrained to the weights that were trained in their previous stage.

3. Results

3.1 Learning error and generalization

All ten nets learned the three tasks successfully. The final average error for the first stage, Prototype-Based Categorization, was .09 after 10,000 epochs. (Figure 4a). This error is not very low, but in most of the nets it was less than .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 .08 (Figure

4b).

The error in the H-Level learning was very low, about .01, because 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 4c).

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.

< Figure 4 about here >

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 feature-filter that reduces the category confusability by decreasing the within-category differences among the icons and increasing the between-category differences as needed to reliably master 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 5-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 (.315) to after (.2). That is, during the course of the prototype learning the 64 points of each category move closer to one another [MANOVA: $F(9,1)=6.12$, $p<.035$]

A further within-category compression from prototype matching (.2) to naming (.172) shows the effects of arbitrary naming on categorical representations (prototypes are analog, names are arbitrary) [$F(9,1)=14.9$; $p<.004$].

< Figure 5 about here >

The same effects are observed with the between-category differences (the distances between the centers of the four categories). From before learning (.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.

< Figure 6 about here >

After prototype-based categorization, the within-category-to-be distances between the two symmetric shapes (Circle [C] vs. Square [S], .82) and the two asymmetric ones (Ellipse [E] vs. Rectangle [R], .91) were smaller than the distances between the four between-category-to-be pairs (C vs. E and C vs. R both, 1.12; S vs. R, 1.32; E vs. 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 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 & 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 (criterion for all conditions: correct bit > 0.5, others < 0.5)

Table I reports percent 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 ten nets met this criterion for Entry-Level names and eight did for H-Level names (see shaded columns in Table I). Assuming chance to be .5, the binomial probability of 9/10 nets successful by chance is .0098 and (and for 8/10, .044). Hence the E-Level grounding successfully transferred to the H-Level categorization.

We also did a control to see whether this outcome depended on some uncontrolled variable rather than grounding transfer. For a set of nets, the E-Level learning stage was skipped; H-Level learning followed immediately after prototype learning. The results are shown in Table II. Based on the same criterion as in Table I, none of the ten nets was successful in the E-level naming, and only three were successful for the H-level naming.

< Figure 7 about here >

< Table I about here >

< Table II about here >

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

$$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 * Q_1}{N} + \frac{P_2 * Q_2}{N}}}$$

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<.0001), confirming that prior direct grounding is essential for grounding transfer.

3.4 Extending the simulation from extensional to intentional categories.

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 Figure 8). 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 then 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 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 III and IV report percent 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 percent 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.

< Figure 8 about here >

< Table III about here >

< Table IV about here >

4. Discussion

These results confirm and extend findings with other connectionist models of categorical perception (Harnad, Hanson & Lubin 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 & Harnad 1998).

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 percent 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 categorisation 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 may explain how some of 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 gets 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 & 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 & Harnad 1998). But a paradigmatic example of the 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.

References

Andrews, J., Livingston, K. & Harnad, S. (1998) Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**(3), 732-753.

Cangelosi, A. & Harnad, S. (1998) The adaptive advantage of symbolic Theft over sensorimotor Toil: Grounding language in perceptual categories. Presented at the 2nd International Conference on the Evolution of Language, London, April 1998. Submitted to the Journal *Evolution of Communication*. http://cogsci.soton.ac.uk/harnad/Papers/Harnad/harnad98.theft.toil.html

Cangelosi A. & Parisi, D. (1998). The evolution of a 'language' in an evolving population of neural nets. *Connection Science*, **10**(2), 83-97.

Csato, L., Kovacs, G, Harnad, S. Pevtzow, R. & 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

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

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 http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad90.sgproblem.html

Harnad, S. (1993). Grounding symbols in the analog world with neural nets. *Think*, **2**, 12-78. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad93.symb.anal.net.html

Harnad, S. (1995a) Does the mind piggy-back on robotic and symbolic capacity? In: H. Morowitz (Ed.) *The Mind, the Brain, and Complex Adaptive Systems*. Santa Fe Institute Studies in the Sciences of Complexity. Volume XXII. pp. 204-220. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad95.mind.robot.html

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

http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad95.robot.html

Harnad, S. (1996) The origin of words: A psychophysical hypothesis. In Velichkovsky B & Rumbaugh, D. (Eds.) *Communicating Meaning: Evolution and Development of Language*. NJ: Erlbaum: pp. 27-44.

http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad96.word.origin.html

Harnad, S., Hanson S.J., & Lubin J. (1991). Categorical perception and the evolution of supervised learning in neural nets. In D.W. Powers & L. Reeker (Eds.) *Proceedings* *of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology.* http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad91.cpnets.html

Harnad, S., Hanson, S.J. & Lubin, J. (1995) Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honavar & L. Uhr (Eds,) *Symbol Processors and Connectionist Network Models in Artificial Intelligence and Cognitive Modelling: Steps Toward Principled Integration*. Academic Press. pp. 191-206. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad95.cpnets.html

Pevtzow, R. & Harnad, S. (1997) Warping Similarity Space in Category Learning by Human Subjects: The Role of Task Difficulty. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.) *Proceedings of SimCat 1997:Interdisciplinary Workshop on Similarity and Categorization*. Department of Artificial Intelligence, Edinburgh University, pp. 189-195. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad97.textures.html

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

Tijsseling A. & Harnad S. (1997). Warping Similarity Space in Category Learning by Backprop Nets. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.) *Proceedings of SimCat 1997:Interdisciplinary Workshop on Similarity and Categorization*. Department of Artificial Intelligence, Edinburgh University, pp. 263-269. http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad97.cpnets.html

FIGURES

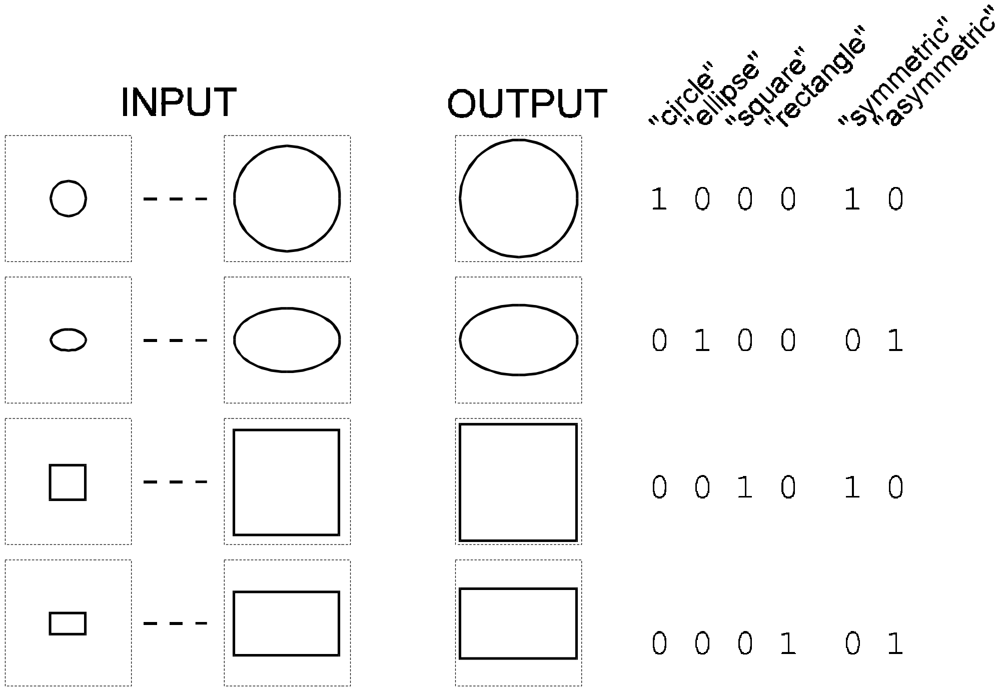

Figure 1 - Stimulus set and localist coding of naming units

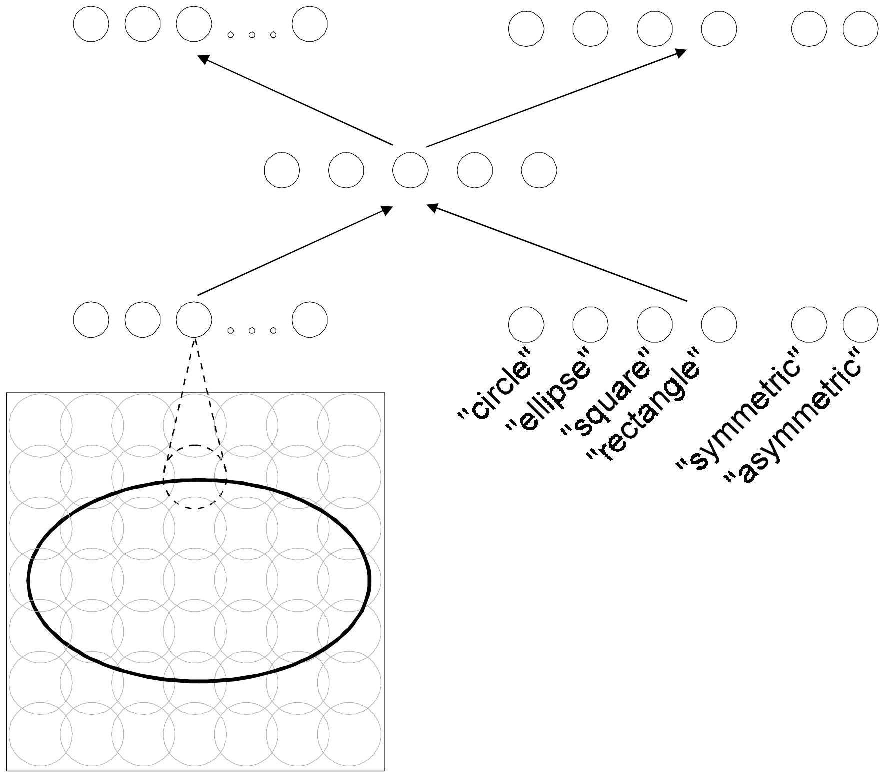

Figure 2 - Neural network architecture and stimulus coding

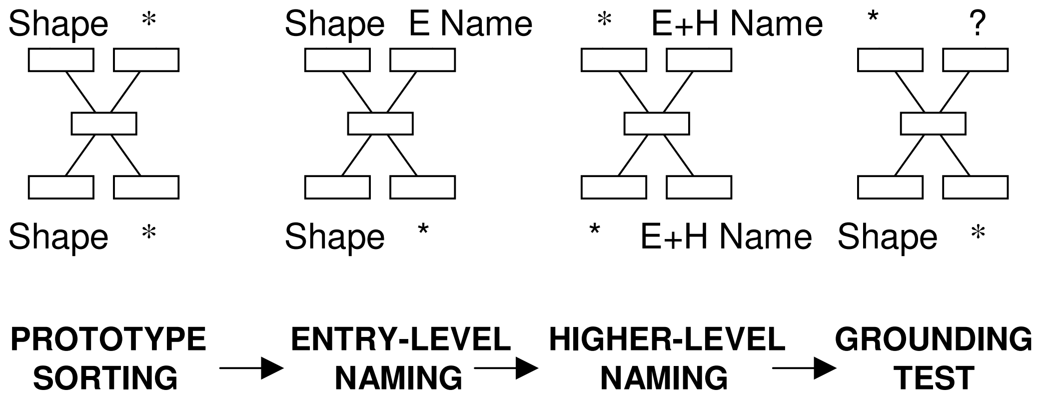

Figure 3: Neural network input and output in the learning and test stages. Absence of input or output in the specified set of units is indicated by *.

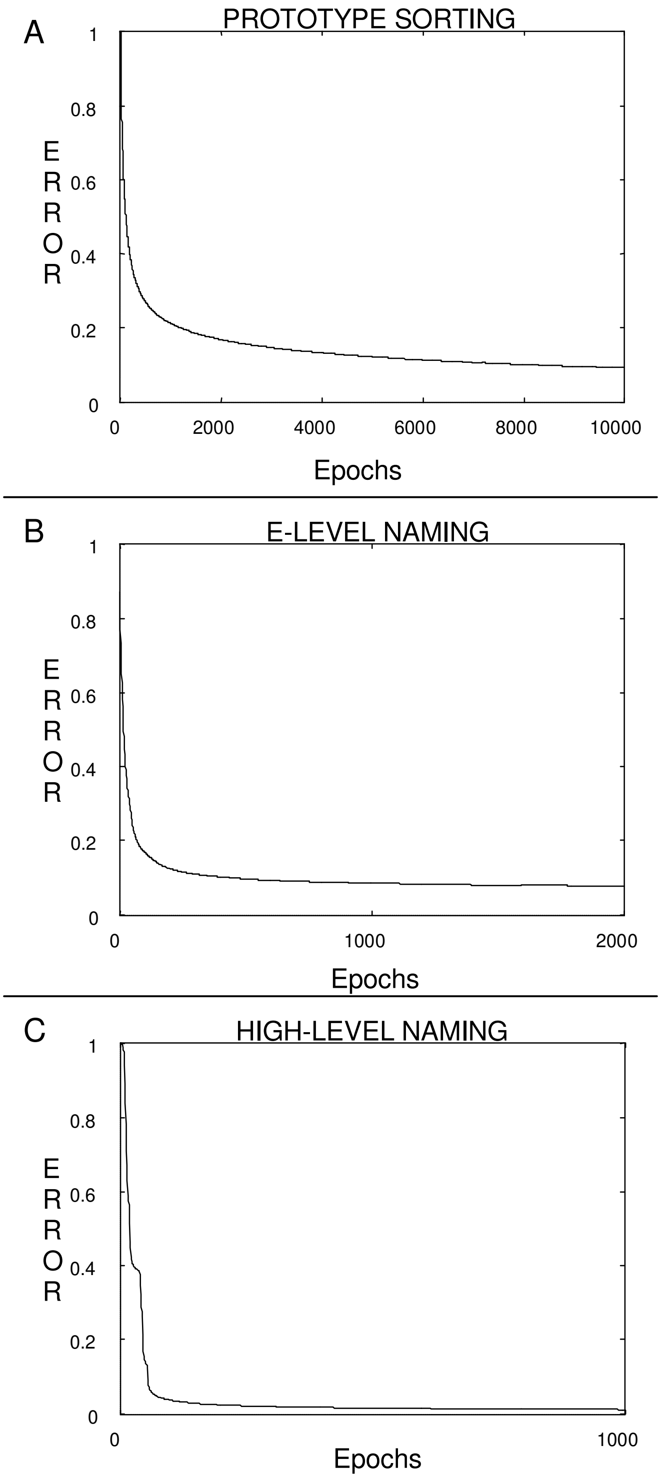

Figure 4 - Learning error for the Prototype Sorting (4a), Entry-level Naming (4b), and H-level naming (4c).

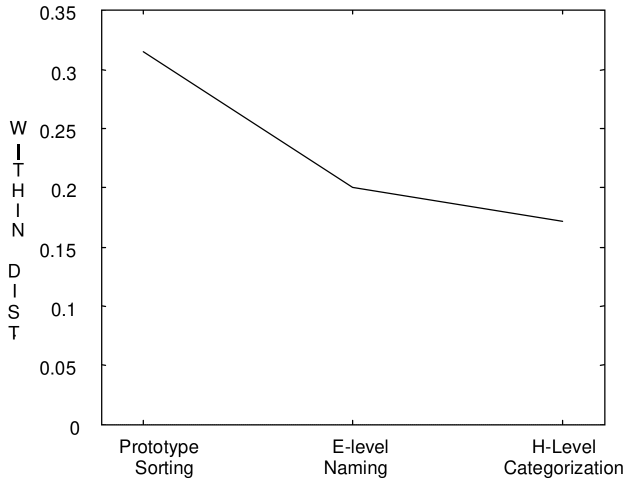

Figure 5 - Average within-category distances

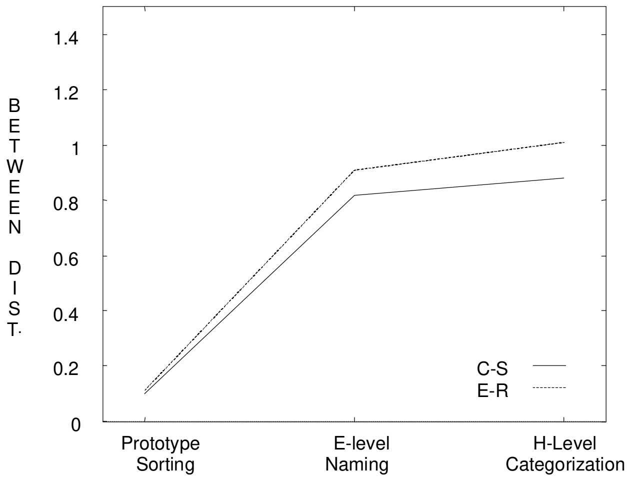

Figure 6: Between-category distances for the pairs Circles-Squares and Ellipses-Rectangles

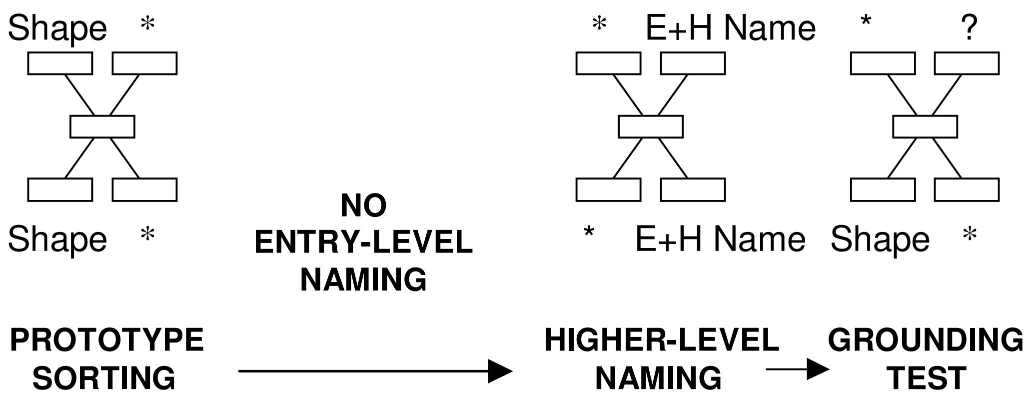

Figure 7: Neural network input and output in the control simulations.

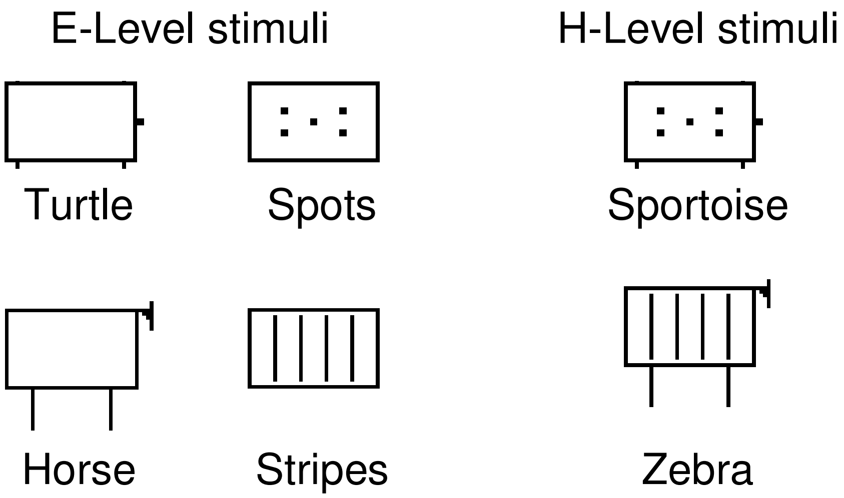

Figure 8 - Stimuli used in the zebra simulations

TABLES

	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

Table I - Percent correct in grounding transfer test. For each net, number on the left is correct responses for E-level names and on right for H-level names. Rows are for the 64 circles (C), ellipses (E), squares (S), and rectangles (R). Shaded cells indicate success in E-level (light grey) or H-level (dark grey) categorization in the grounding transfer (criterion: at least 50%)

	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

Table II - Percent correct in grounding transfer controls. For each net, number on left is correct responses for *E-level* names and on right for *H-level* names. Rows are for the 64 circles (C), ellipses (E), squares (S), and rectangles (R). Shaded cells indicate the nets that succeeded in E-level (light grey) or H-level (dark grey) grounding transfer (criterion: at least 50% correct).

	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

Table III - Percent correct in grounding transfer test for Zebra simulations. Numbers refer to *H-level* names. Shaded cells refer to the nine successful H-level nets in the grounding transfer (criterion: at least 50% correct)

	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

Table IV - Percent correct in grounding transfer controls for Zebra series. Numbers refer to *H-level* names. No net met the 50% success criterion.