

Approaches to Grounding Symbols in Perceptual and Sensorimotor Categories.

Angelo Cangelosi

Adaptive Behaviour & Cognition Research Group

School of Computing, Communication and Electronics

University of Plymouth (UK)

acangelosi@plymouth.ac.uk

Abstract

This chapter presents the Cognitive Symbol Grounding framework for the grounding of language into perception, cognition and action. This approach is characterized by the hypothesis that symbols are directly grounded into internal categorical representations, whilst at the same time having logical (e.g. syntactic) relationships with other symbols. The internal categorical representations, constituting the meanings upon which symbols are grounded, include perceptual, sensorimotor, and social categories, as well as internal state representations. Two main modeling approaches to the symbol grounding are presented: (i) the connectionist approach, based on artificial neural networks for category learning and naming tasks, and (ii) the embodied modeling approach, based on adaptive agent simulations and cognitive robots. These models provide an integrative view of the cognitive systems and help our understanding of the relationships between vision, action and language.

1. Cognitive Symbol Grounding

1.1 *The Symbol Grounding Problem*

Computational cognitive models that focus on linguistic and symbol-manipulation abilities can use symbols that are either grounded or ungrounded. Models using ungrounded symbols require the interpretation of an external user, such as the researcher, to identify and understand the meaning associated to symbols. For example, in a typical symbolic model of language acquisition, IF-THEN rewriting rules are used to represent the knowledge of the language. In the classical example of past tense learning, rules are used to represent the ability to form the past tense of verbs (e.g. rule 1: "IF regular stem, THEN add suffix –ed", rule 2: "IF irregular verb *to go*, THEN use *went*"). In such model, symbols such as the word "go" are ungrounded because the interpretative role of the human modeler is required to understand that "go" means "to move or travel". Even if the same model had an additional set of symbols that provides definitions of the verbs (e.g. a dictionary of verb meanings), these would still be ungrounded and self-referential symbols. Such situation is similar to the well-know Chinese room argument (Searle, 1980) where the symbolic task of responding to questions in Chinese, without knowing the language, can apparently be well performed whilst at the same time not understanding the meaning of the questions and answers.

On the other hand, the cognitive models based on grounded symbols use words that are inherently significant to the cognitive system. They do not

necessitate the interpretation of an external user. In such models, the same cognitive agent is able to link the symbols to their own meanings through perceptual, cognitive and sensorimotor experience. For example, in the grounded model described in section 2.2, adaptive agents are able to name edible mushrooms, necessary to their survival, with the word “eat” (Cangelosi & Harnad, 2000). The same agent is also able to identify the meaning of such a word by referring to the external stimuli, and the corresponding internal states, that produce the word (in this example, the mushrooms that simultaneously possess two perceptual features such as red color and round shape). In addition, the agent’s sensorimotor behavior of approaching and eating the mushrooms also constitute the grounding of the symbol “eat”.

The issue of intrinsically linking the symbols used by a cognitive agent to their corresponding meanings has been called “the symbol grounding problem” (Harnad, 1990). This claims that for a cognitive model to be considered psychologically plausible, symbols should be intrinsically linked to the agent’s ability of acquiring categories through interaction with its environment. In particular, it is essential that at least some basic symbols are directly grounded on sensorimotor categories. Subsequently, new and indirectly grounded symbols (and their corresponding categories) can be formed through the combination of previously grounded basic symbols and categories.

Hybrid symbolic-connectionist models were originally proposed as ideal candidates for solving the symbol grounding problem (Harnad 1990; Sun 2002). These would typically include some connectionist module, i.e. a neural network, that deals mainly with the task of grounding the basic symbols into

perceptual and categorical representations. In addition, the hybrid system would use some symbolic modules and information structures (e.g. scripts, lexicon, and episodic memory). Feedforward neural networks are ideal candidates for performing classification (e.g. categorical) tasks, and the activation of hidden units can be used to analyze their internal, categorical representations. Additional symbolic modules, such as rules sets, can be added on top of the connectionist module to deal with symbol manipulation tasks. These will be mostly used to control high-level cognitive tasks (Miikkulainen 1994).

More recently, alternative modeling approaches have been proposed to deal with the symbol grounding problem. Some are based on fully connectionist systems. This only consists of a neural network, or a connected ensemble of networks, that perform both the basic grounding and the symbol manipulation tasks. Other approaches are based on embodied agent models and cognitive robotics. They focus on the sensorimotor grounding of symbols. In addition, such embodied models also rely on social learning and interaction between agents (including robots, internet agents and humans) for the development of shared grounded symbol communication systems. Some of these robotic approaches also include the use of connectionist networks, while others use different control architectures.

1.2 Grounding Symbols in Cognition

The various modeling approaches to the symbol grounding problem all have in common some core features. First, each symbol is directly grounded into an internal categorical representation. Internal representations include perceptual categories (e.g. the concept of red color, square shape, and

female face), sensorimotor categories (e.g. the concept/action of grasping, pushing, pulling), social representations (e.g. individuals, social groups and relationships) and other categorizations of the organism's own internal states (e.g. emotional states, motivations). Secondly, these categories are connected to the external world through our perceptual, motor and cognitive interaction with the environment. Categorical representation of the organism's internal states can also be mediated by our sensorimotor and cognitive system.

This view of the symbol grounding process will be referred to as "Cognitive Categorical Perception". It is consistent with growing theoretical and experimental evidence on the strict relationship between symbol manipulation abilities and our perceptual, cognitive and sensorimotor abilities (e.g. Pecher & Zwaan, in press). For example, some authors have explicitly supported the fact that symbols are grounded in our ability to form categories. Harnad (1990; 1987) identifies our innate ability to build discrete and hierarchically-ordered representations of the environment (i.e. categories) as the basis of all higher-order cognitive abilities, including language. The categorization of the external and internal world is adaptive to the organisms since it helps them sorting things out and knowing how to interact with them. This ability is called Categorical Perception (Harnad, 1987). In particular, it refers to the process of re-representation of the external environment into internal categories and to the process of "warping" of the similarity space of internal categorical representations. This re-representational process results in the compression of within-category differences between members of the same category, and the expansion of between-category distances amongst members of different

categories. Categorical perception is a widespread ability in natural and artificial cognitive systems. It has been shown to occur in animals (e.g. Zentall et al. 1986) and human subjects (e.g. Goldstone 1994). The warping effects have also been analyzed in real neural systems (Kosslyn et al. 1989) and in artificial neural networks (Tijsseling & Harnad 1997; Cangelosi, Greco & Harnad 2000; Nakisa & Plunkett 1998).

The phenomena of within-category compression and between-category expansion can be graphically represented through the process of the formation of clusters of points in the similarity space of categories (Figure 1). Before category learning (Figure 1 Left), category members produce an undifferentiated similarity space. For example, points representing square objects overlap with those representing circles. The other diagram (Figure 1 Right) represents the formation of two distinct clusters (cluster of squares vs. cluster of circles) after category learning has occurred. The diagrams represent an abstract two-dimensional similarity space, where each dimension may correspond to some classification component (e.g. geometrical feature) or to the hidden unit activation of a neural network. Relative distances in the similarity space can be calculated using Euclidean measures between points. The two dotted circles in each diagram represent the within-category distances, corresponding to the standard deviation of the Euclidean distances between each point and the center of its cluster. The continuous straight line represents the between-category distance, e.g. the Euclidean distance between the centers of the two clusters.

Other researchers have highlighted the relationship between perception, language and action. Barsalou (1999; see also Joyce et al. 2003 for a related

connectionist model) supports a view of our cognitive system based on perceptual symbol systems. Perceptual experience, through association areas in the brain, captures bottom-up patterns of activation in sensorimotor areas. In addition, in a top-down manner, association areas partially reactivate sensorimotor areas to implement perceptual symbols. This implements a mental simulator that produces limitless simulations of schematic representations of perceptual components. Simulators implement a basic conceptual system that supports categorization, produces categorical inferences and supports productivity, propositions, and abstract concepts.

Coventry and Garrod (2004) propose a cognitive system grounded in both perceptual and action abilities. They hypothesize the on-line activation of situation-specific models for tasks involving spatial cognition and spatial language judgments (e.g. when subjects are asked to evaluate the use of specific spatial terms). For example, they have extensively studied the appropriateness of the locative prepositions *over* and *above* for describing a visual scene depicting a man holding an umbrella and some pouring rain. Experimental and modeling evidence (e.g. Coventry et al. 2001; Cangelosi et al., in press) shows that subjects take into consideration a series of factors activated by their previous experience and by the input stimuli involved in the spatial cognition task. These factors include geometric information (relative orientation of an umbrella respect to the direction of the rain and the position of the human being protected), object-specific knowledge (e.g. typical rain protection function performed by an umbrella), sensorimotor experience with the objects involved (e.g. force dynamics factors on the direction of the rain).

The grounding of language into action has been extensively studied by Glenberg and collaborators. They have developed an embodied theory of cognition (see also Clark 1997), where meaning consists of the set of actions that are a function of the physical situation, how our bodies work, and of our experiences (Glenberg & Kaschak, 2002; Borghi et al. in press). For example, Glenberg demonstrated how language comprehension takes advantage of our knowledge of how actions can be combined and how linguistic structures coordinate with action-based knowledge to result in language comprehension.

In addition to experimental evidence, the computational approaches to the symbol grounding problem have also provided further evidence in support of the Cognitive Symbol Grounding framework. Various connectionist, robotic, and hybrid symbolic-connectionist models provide a working framework for the implementation of symbol grounding in artificial cognitive systems. The modeling approaches based on classical connectionist networks primarily focus on the grounding in perception and the linking of vision and language. The embodied approaches, based on robots and hybrid robotic/connectionist models, tend to take into consideration both perceptual and sensorimotor components and focus on the linking between vision, action and language. In the next sections we will review some of these models and will highlight the main findings in support of the cognitive symbol grounding view. The review will mainly focus on models developed by the author and his collaborators at the Adaptive Behaviour & Cognition Research Group¹ of the University of Plymouth (UK). However, other relevant models and simulations will also be briefly referred to and discussed.

¹ <http://www.tech.plym.ac.uk/soc/research/ABC>

2. Linking Vision and Language: Connectionist Approaches to Category Learning and Symbol Grounding

2.1 Connectionist Modeling of Category Learning and Naming

The great majority of neural network models of symbol grounding usually concern (a) the use of a dual-route architecture and (b) the simulation of language production (naming) and comprehension tasks. The dual-route architecture (Figure 2) typically involves both visual input (e.g. retina projection or visual feature list) and linguistic input (e.g. localist or graphemic/phonetic encoding of symbols). The output layer will have symbolic units for representing names and words (e.g. with a phonetic encoding of the lexical items), and a categorical representation of input stimuli (e.g. a localist node for each category, or a visual representation of category prototypes). Some models can use an action-based representation of categories, although this is more typical of connectionist modules within embodied agent systems (see Section 3.1). All input and output layers are connected via some hidden units. The route from visual input to symbolic output is used for language production tasks, such as naming of the object represented in the visual scene and of their categories. This is the essential route of a symbol grounding network, since the vision→language link constitutes the core mechanism in perceptual grounding. The route from linguistic input to visual/categorical output is used for language understanding tasks.

One of the first and most influential models of naming and symbol grounding was that developed by Plunkett and collaborators (Plunkett et al. 1993; Plunkett & Sinha 1992). The neural architecture is based on the

standard dual-route network. The model had two distinct sensory modalities (retinal and verbal) in the input and output layers, and two hidden layers. An auto-associative learning task was used to map both input modalities into the corresponding output nodes. During testing, either the verbal or retinal input was presented, and the network was requested to produce the output corresponding to the opposite modality. The most interesting result was that training performance was not linearly related to the extent of training. The networks went through stages of sudden improvements, exhibiting something like a “vocabulary spurt” without any apparent reason. This happened both for comprehension and for production, but at different times. The simulation exactly reflects what is observed in children, or in adults when learning a new language: comprehension precedes production. At some stage the network was able to “understand” what image a name referred to, but not yet to produce this name when given the corresponding image. But at a later stage, a new and sudden improvement would also be observed in production.

The connectionist models developed by Harnad, Hanson and Lubin (1991; 1994) specifically focused on the symbol grounding and categorical perception effects. They trained three-layer feed-forward networks to sort lines into categories according to their length. Such lines were represented by input units using two basic coding schemes, iconic (e.g., a length-4 line could be coded as “11110000”) vs. positional (e.g., the same line coded as “00010000”). Single bit values could also be more or less discrete (e.g. coarse representations such as .1 for 0, or .9 for 1 were used). Training consisted of two sequential back-propagation learning tasks: autoassociation and category learning. The first task allowed networks to “discriminate” between different

stimuli using a pre-categorization task with autoassociative learning. The hidden unit activation vectors were examined to record the baseline categorical perception distances for each pair of input patterns. After the autoassociation task, the networks were trained to categorize stimuli by sorting lines into three categories: short, middle, long. The comparison of baseline and post-categorization distances in the networks' hidden activations showed the natural side-effect of within-category compression and between-category expansion revealed by human categorical perception. Another point of interest from this simulation is that a close scrutiny of hidden representations permits a better understanding of the factors influencing categorical perception. Harnad et al. (1991) found that the distances between hidden unit activations are already maximized during auto-association, by effect of the baseline discrimination. This separation, however, is not always so clear-cut as to allow linear separability in the hyperspace of hidden activation, as happens with perfectly categorized stimuli. The back-propagation algorithm, which simulates category learning through supervised feedback, has the effect of "pushing" such unclear representations and form a hyperplane that separates members of different categories. This results in an improved organization of categorical representations. Tijsseling and Harnad (1997) later replicated these results and additionally manipulated other factors such as the similarity between stimuli. When there was either extreme nonseparability or extreme separability, the categorical perception effects were not present. In the extreme nonseparability case, this is due to the fact that the task is too difficult to learn, already at the discrimination level. In the latter case, compression effects are not present because the task is too easy.

In fact, there is no need for category learning because categories already implicitly exist in input stimuli that are extremely separable.

The models above described mostly refer to “naming” tasks, that is the acquisition of a series of labels for naming (visual) objects and their categories. They can mostly investigate how representations of name-objects associations are learned and how they contribute to the grounding of symbols. However, in human language not all words are always directly associated with their referents. In fact, words are mostly associated with other words through syntactic rules and relationships. As shown in (ungrounded) connectionist networks (Elman, 1990) and in studies based on lexical analyses (Landauer & Dumais, 1997), words contain “latent” semantic information that is implicitly expressed through symbol-symbol grammatical relationships. As a matter of fact, the correct definition of symbol implies the presence of symbol-symbol relationships (Cangelosi et al. 2002; Harnad, 1990; Deacon, 1997). These can be logical Boolean relationships, or typical grammatical rules as in human languages.

Some models of language acquisition have employed a grounded approach that also considers the use of symbol-symbol relationships (e.g. Dyer, 1994). In particular, various connectionist models have focused on the perceptual grounding of spatial quantifiers. This is because spatial terms have been shown to depend on a variety of semantic and contextual factors such as geometric constraints and extra-geometric information (Coventry & Garrod, 2004). Regier (1996) developed a computational model of spatial prepositions using a method called constrained connectionism. The model learns various spatial prepositions for static (e.g. over and above) and moving (e.g. through)

objects, and makes explicit use of the processing of geometrical information. An image of two objects (ground and figure) is given in input to the lower layer of the network. Then the image goes through several levels of geometrical processing. The output units, corresponding to spatial prepositions, are activated according to the geometrical position of the figure object with respect to the central ground.

Cangelosi, Coventry and collaborators (Cangelosi et al. in press) have recently developed a connectionist model that produces linguistic description of dynamic scenes involving spatial relations between objects. In addition to the geometric constraints considered in Regier's model, this new study focuses on the role of extra-geometrical factors, such as object knowledge and interaction. The model processes 60-second movies showing a located object (e.g. teapot) pouring a liquid (e.g. water) into a reference object (e.g. cup). The task of the network is to name the objects, and more importantly, to select the most appropriate spatial preposition (e.g. over, above, under and below) describing the spatial relationships between objects. The model consists of three modules: (i) a neurally-inspired vision module based on Ullman-type routines (Ullman 1996), (ii) an Elman recurrent neural network to learn compressed neural representations of the dynamics of a scene, and (iii) a dual-route network for expressing the names of objects and the spatial terms. The dual-route network is the core component of the model because it integrates visual and linguistic knowledge to produce a description of the visual scene. The activation values of the linguistic output nodes correspond to rating values given by subjects for the rating of the four prepositions. The multi-layer perceptron is trained via error backpropagation, using rating data

collected during experiments. Some of the ratings are also used for the generalization test. Simulation results consistently show that the networks produce rating values similar to that of experimental subjects. It also accurately predicts new experimental data on the ratings of scenes where only the initial frames are shown and the subjects must “mentally replay” the scene and predict its end frame (i.e. where the liquid ends). This model is also consistent with Barsalou’s perceptual symbolic systems theoretical framework (1999). Currently, such a model is being extended to deal with further linguistic terms, such as the vague quantifiers some, few, many. The working hypothesis is that this grounded connectionist approach will permit the identification of the main mechanisms responsible for quantification judgment and their linguistic expression.

2.2 Connectionist Modeling of Symbol Grounding Transfer

In the connectionist models of category learning and naming discussed above the focus is on the direct grounding of symbols into perception. However, not all symbols need to be directly grounded in perception. In fact, directly grounded symbols can be combined together, through grammatical rules, to produce definition of new symbols. For example, we can describe a zebra through the definition: “A zebra is an animal with horse shape and stripes”. If a person has never seen a zebra, but has direct grounding experience of the two words “horse” and “stripes”, she can easily infer the perceptual meaning of “zebra”. This process is called “grounding transfer”, i.e. when the grounding of basic words is transferred to that of new symbols and categories acquired via linguistic descriptions. This phenomenon has been

studied by Cangelosi and colleagues (Cangelosi et al. 2000; Greco et al. 2003) using neural networks. The model consists of a network that has to categorize abstract images consisting of combinations of three different shapes (square, cross, dots) and three different colors (red, green, blue). A modular dual-route neural network was used, in which the hidden layer was organized into two separated groups. The output units indicating shapes were only connected to the first group of hidden units, whilst those indicating colors had connections solely to the other units. This modular connectivity forces the functional division of the hidden layer into a group dedicated to categorizing shapes and a group to classifying colors.

This type of models is trained through a series of sequential stages: Prototype-sorting, Entry-Level naming, Entry-Level imitation and naming, and Higher-Level learning and Grounding transfer test (Figure 3). In the Prototype-sorting and Entry-Level naming stage the neural nets are initially trained to categorize and name the color and shape of objects perceived on the retina. In the Entry-level imitation and naming stage an extra learning cycle is executed. This consists on the use of only the symbolic units in both the input and the output layers. During the first two training stages, the networks learn through direct trial and error experience supervised by corrective feedback. In the third training phase (Higher-Level learning), networks acquire new higher-order categories solely through symbolic descriptions. New categories are built by combining grounded names. Each description contains the name of a shape, of a color and the name of an object that is new to the network. The grounding test is performed at the end of training by presenting in input the visual representation of the new object.

The networks were able to categorize the colors and shapes of the training stimuli correctly, with a success rate of 100%. During the grounding transfer test, novel retinal stimuli depicting new objects were presented to the networks for the first time, in order to check if grounding had been “transferred” from directly grounded names to higher-order categories. The rate of correct test responses for all networks was 85%. This indicates that this model is able to do both basic grounding, and above all, to transfer this grounding to new concepts and symbols. This second ability highly depends on the modular organization of the connections between hidden and linguistic output nodes. (Greco et al. 2003). Moreover, such a study further supports a fully-connectionist approach to symbol grounding, since the same network also demonstrates symbol manipulation abilities.

3. Linking Vision, Action and Language: Embodied Approaches to Language Learning and Evolution

Embodied agent models and cognitive robotic research has recently contributed to the issue of symbol grounding. Many robotics models focus on the role of social learning and interaction between agents as the basis of language (e.g. Steels 1999, 2002; Vogt 2002). Others robotic (Marocco et al. 2003; Cangelosi et al. 2004) and adaptive agent models (Cangelosi & Harnad 2000) also center their attention on the cognitive grounding of symbols. However, what all these approaches have in common is the focus on the role of action and sensorimotor knowledge in the grounding of language. In this section we will describe some of these models to show their contribution to

symbol grounding research, and the linking between vision, action and language.

3.1 Grounding symbols in simulated agents: The symbolic theft hypothesis

Cangelosi & Harnad (2000) have employed an adaptive agent model to study the role of symbol grounding and categorical perception in the origins of language. They consider two ways of acquiring categories. In the first method, new categories are acquired through feedback-corrected, trial and error experience with the environment. This is the “sensorimotor toil” approach. Alternatively, new categories are acquired through language, i.e. through hearsay from linguistic propositions provided by language-speaking adults. This is called the “symbolic theft” strategy. In competition, symbolic theft always outperforms sensorimotor toil because it is more efficient than toil (e.g. only one propositional description of a new category is enough to learn it – as in the case of the zebra example). In contrast, repeated experience is required to learn a category by sensorimotor toil. The significant advantage of symbolic theft has been hypothesized to produce an adaptive benefit for language and can help explaining the origins of language (Harnad, 1996, 2002).

Cangelosi and Harnad (2000) developed a computational model based on simulated adaptive agents using a mushroom world scenario (Cangelosi & Parisi 1998). This approach is characterized by the simulation of the on-line interaction between the agent and its environment for a foraging task. The behavioral and cognitive experience of the agent is determined by its embodiment features (e.g. set of sensors and actuators) and the physical aspects of the environment. In this simulation, agents rely on learning

categories of foods to survive. For example, mushrooms with feature A (i.e., those with black spots on their tops) are to be eaten; mushrooms with feature B (i.e., a dark stalk) are to have their location marked, and mushrooms with both features A and B are to be eaten, marked and returned to. All mushrooms have three irrelevant features (C, D and E) that the foragers must learn to ignore. When organisms approach a mushroom, they emit a call associated with their functionality (“EAT”, “MARK”). Both the correct action pattern (eat, mark) and the correct call (“EAT”, “MARK”) are learned during the foragers’ lifetime through supervised learning (sensorimotor toil). Under some conditions, the foragers also receive the call of another forager as input. This will be used to simulate theft learning of the return behavior.

The behavior of organisms is controlled by neural networks, similar to those described in the previous sections. Category learning and naming also resemble those described in Figure 3 (although in a less sequential way). The main difference consists in the ability to process the sensory information about the closest mushroom to activate the output units corresponding to the action to perform on the environment. The agents learn to categorize the mushrooms by performing the correct action and naming.

The population of foragers is also subject to selection and reproduction through a genetic algorithm. During the forager’s lifetime, the fitness is computed by assigning points for each time a forager reaches a mushroom and performs the correct action on it. At the end of their life-cycles, the best foragers with the highest fitness in each generation are selected and allowed to reproduce by engendering five offspring each. The population of newborns is subject to random mutation of their initial connection weights.

To test the adaptive advantage of symbolic theft versus sensorimotor toil, we compared foragers' behavior for the two learning conditions. In one simulation, the two strategies were directly compared. In the first 200 generations, all organisms learn through sensorimotor Toil to eat mushrooms with feature A and to mark mushrooms with feature B. They also learn the names of the basic categories. The return behavior and its name are not yet taught. In the following 20 generations, organisms live for a longer life stage. In the second part of their lifetime, they are divided into the two groups of Toilers and Thieves. Toil foragers go on to learn to return to AB mushrooms in the same way they had learned to eat and mark them through honest toil. In contrast, Theft foragers learn to return on the basis of hearing the vocalization of the mushrooms' names (e.g. "EAT" + "MARK" = "RETURN"). They rely completely on other foragers' calls to learn to return as they do not receive the feature input. To test the adaptive advantage of Theft versus Toil learning, the foragers' behavior for the two conditions was compared by counting the number of AB mushrooms that are correctly returned to. Thieves successfully return to more AB mushrooms than Toilers. This means that learning to return from the grounded names "EAT" and "MARK" is more adaptive than learning it through direct toil based on sampling the physical features of the mushrooms.

A more direct way to study the adaptive advantage of Theft over Toil was to see how they fared in direct competition against one another. At the second stage after generation 200 stage, the 100 foragers were randomly divided into 50 Thieves and 50 Toilers who must all learn to return. Direct competition only occurs at the end of the life cycle, in the selection of the fittest 20 foragers to reproduce. Results consistently showed that thieves gradually came to

outnumber Toilers, so that in less than 10 generations the whole population was made up of Thieves.

All these results support the original hypothesis that a Theft learning strategy, based on language hearsay, is much more adaptive than a Toil strategy. This adaptive advantage could have constituted the basis for the origin of language and its adaptive advantage. In addition, categorical perception analyses on the neural network hidden activation showed that symbolic theft produced enhanced warping effects. The categorical representations for the category "return" in the Toil vs. Theft organisms were contrasted. Data on the Euclidean distance comparisons showed that categories acquired via theft (i.e. via language) have higher between-category distances and lower within-category distances (Cangelosi & Harnad, 2000). This suggests that Theft learning not only is more advantageous for survival (Thieves collect more return mushrooms than Toilers), but it also optimizes the internal categorical representation by the categorical perception effects. Language learning is based on categorization, but in return it improves categorical learning.

This effect has consistently been reported in connectionist and experimental models of category learning and naming (e.g. Cangelosi et al. 2000; Lupyan, in press). In addition, different word classes produce different levels of enhancement of categorical perception effects. For example, in a related model of the evolution of syntax, agents evolve two different classes of words, namely verbs (names of actions) and nouns (names of objects). Comparison of language-induced categorical effects showed that verbs produce further enhanced warping effects compared to nouns (Cangelosi &

Parisi, 2000). These enhancement effects have been explained by the sensorimotor component of the grounding of verbs (Cangelosi & Parisi 2004). Overall, all these results support the importance of modeling the grounding of language into perceptual and sensorimotor abilities.

3.2 The Emergence of Language in Robots

Cognitive robotics has been recently used to model the emergence of language in groups of robotic agents. This has been demonstrated in groups of autonomous robots (e.g. Vogt, 2002; Marocco et al. 2003) and hybrid groups of robots and humans (Steels 1999; Roy et al. 2003). These models further permit a deeper understanding of the embodied basis of cognition, and in particular of the grounding of language into action (Glenberg & Kaschak 2002). For example, in an evolutionary robotic model of the emergence of language, Marocco et al. (2003) showed that the ability to form categories from direct interaction with the environment constitutes the ground for subsequent evolution of lexicons based on names of actions and names of objects. In this model, agents use proprioceptive and tactile information to actively explore objects in the environment, build categories and a shared lexicon. The controller of each robotic agent consists of an artificial neural network in which, in addition to proprioceptive sensors, two symbolic neurons receive their input from the other agents. The output layer has motor neurons, which control the actuators of the corresponding joints, and two additional symbolic output neurons, which encode the signal to be communicated to the other agents. A genetic algorithm is used to evolve the behavior of agents.

The evolutionary robotics model was used to run a series of experiments on the role of various social and evolutionary variables in the emergence of

shared communication. Experimental design variables included the selection of speakers (the parent or all peers in the population) and the evolutionary time in which communication is allowed (in parallel with the evolution of manipulation abilities, or subsequently to the pre-evolution of good behavior). The simulation results showed that populations evolve stable shared communication mostly when the parents act as speakers and when signaling is introduced in the second stage. Additional analyses of the evolutionary data and the neural network behavior also supported the findings that: (i) the emergence of signaling brings direct benefits to the agents and the population, in terms of increased behavioral skills and comprehension ability; (ii) there is a benefit in direct communication between parents and children, not only because of kinship mechanisms, but also because parents produce more stable and reliable input signals; (iii) the pre-evolution of good sensorimotor and cognitive abilities (i.e. sensorimotor grounding) permits the establishment of a link between production and comprehension abilities, especially in the early generations when signaling is introduced.

A second robotic model developed (Cangelosi et al. 2004) focused on human-robot communication. This study simulated epigenetic robots that observe and execute actions via imitation learning. An artificial language was used to communicate about the names of actions and objects. Robots first learned a set of basic actions by mimicking them, while simultaneously learning words corresponding to these actions (direct grounding). Subsequently, they learned higher-level composite behaviors by receiving linguistic descriptions containing the words previously acquired. The agents merged basic actions into composite actions by transferring the neural

grounding of the words referring to basic actions to the word indicating the higher-level behavior (cf. grounding transfer in section 2.2).

The imitator robot, during training, learned the basic actions and names of opening and closing their left and right arms (upper arms & elbows), lifting them (shoulders), and moving forward and backward (wheels), together with the corresponding words. After few training epochs, robots received 1st level linguistic descriptions of combined actions. This consisted of a new word (for the new higher-order action) and two known words referring to basic actions. For example, the action of grabbing the object in front of the agent was described as: “close_left + close_right = grab”. Grounding was transferred from “close_left” and “close_right” to “grab”. Consequently, when the agent was given the command “grab”, it was able to successfully execute the combined action of pushing its arms towards the object and grabbing it. After few more training epochs, the same robot received 2nd order descriptions. These consisted of a new word (for a novel higher-order action) and a combination of a basic word and a 1st order word. For example, “move_forward + grab = carry” combines the grounding of the actions of grabbing (1st order) and moving forward (basic order) and produces the behavior of carrying (2nd order). This higher-level grounding was also successfully transferred to the new 2nd-order word, enabling the agent to correctly perform the action of carrying on hearing the word “carry”. The system learned several of these combined actions simultaneously, and also four-word definitions and grounding transfers of up to three levels have been realized. This study provides a further demonstration of the process of

grounding, and grounding transfer, of symbols in sensorimotor categories in a context of human-robot interaction.

4. Discussion and Conclusion

The main features of the Cognitive Symbol Grounding approach will be summarized in this section, together with the highlight of the most relevant characteristics and contributions of the above connectionist and embodied models. This will permit a consideration of the open research issues and the identification of the future research directions on symbol grounding research.

The Cognitive Symbol Grounding approach is characterized by the following principles:

- A symbol is directly grounded into an internal categorical representation (Harnad, 1990), and at the same time it has logical (e.g. syntactic) relationships with other symbols.
- The internal categorical representations include perceptual, sensorimotor, and social categories, as well as internal state representations
- Categories are connected to the external world through our perceptual, motor and cognitive interaction with the environment
- Such view is consistent with current theoretical and experimental psychology research in the grounding of language and cognition in perceptual abilities and embodiment factors (e.g. Barsalou 1997; Glenberg & Kaschak, 2002).

Among the various cognitive modeling approaches to cognitive modeling, those based on artificial neural networks and on embodied agents provide a

theoretical framework consistent with the Cognitive Symbol Grounding framework. In particular, connectionist models based on grounded neural networks are characterized by:

- main focus on perceptual grounding
- categorization ability that produces categorical perception effects
- the use of dual-route neural architectures that permit the simultaneous simulation of language production (vision→language) and language understanding (language→vision/action) abilities
- the transfer of grounding from directly-grounded symbols to higher-order symbols (e.g. Greco et al. 2003). This permit the simultaneous control of basic grounding and of symbol combination tasks.

The embodied approach to cognitive symbol grounding is mostly based on adaptive agent and robotic models. Such methodologies are characterized by:

- main focus on grounding in action
- effects of sensorimotor knowledge in the differentiation of symbol (word) classes
- consideration of the social and interaction factors in the development of shared lexicons
- analysis of evolutionary factors in the origins of language and cognition (e.g. symbolic theft hypothesis, Cangelosi & Harnad 2000)

The two computational approaches to the grounding of symbols and language only apparently differ in the focus they give to various cognitive and social interaction abilities. In fact, the embodied approach encompasses most of the characteristics of the grounded connectionist modeling. This is

particularly true for adaptive agent and robotic models that use neural networks to control the behavior and cognitive system of the agents. All the embodiment modeling examples reported above use neural controllers to organize the sensorimotor, cognitive and linguistic behaviors of the foraging agents and of the evolutionary and epigenetic robots.

The integration of connectionist networks and embodied agent models has important theoretical and methodological implications. It provides an integrative view of the cognitive system, in contrast to other cognitive modeling approach that only simulated isolated abilities (e.g past tense connectionist model that only focus on morphology). This is because all sensorimotor, cognitive and linguistic abilities are controlled by the same (connectionist) network, or a modular set of networks. This is particularly important for symbol grounding research, because it provides a means for linking vision, action and language and is consistent with embodiment views of cognition (Clark, 1997; Varela et a. 1991).

The choice and use of either of the two modeling approaches also has some methodological implications, in addition to the core characteristics listed above. Connectionist models tend to be used in studies that use cognitive tasks with clear and pre-defined symbol-meaning sets. When there is a pre-structured stimulus set and a predefined number of categories (and names) to which the individual stimuli belong, a typical connectionist simulation will suffice. The training of a connectionist network on category learning and naming is similar to a laboratory study on categorization where the researcher decides the objects and categories to use in the experiment. Adaptive agent and robotic approach, instead, permit the on-line formation of categories

during the organism's interaction with the environment. This is essential for the simulation of tasks and environments where the process of stimulus grouping and meaning formation is flexible and not defined a priori. For example, simulations of the emergence of communication allow agents to construct an autonomous categorical representation of the environment, while also developing the names for such categories (Cangelosi & Parisi, 2002).

Another methodological implication for the selection of the most appropriate modeling approach is dependent on the aims of the research. Studies interested in the neural basis of the symbol grounding will use connectionist models, or embodied models using neural controllers (e.g. Cangelosi & Parisi, 2004). Instead, research on the social origins of shared symbols may only require the use of robots with symbolic architectures (e.g. Vogt, 2002).

The computational models discussed in sections 2 and 3 only provide some examples of models of the cognitive symbol grounding hypothesis. More models and experimental studies are still necessary to develop a deeper understanding of the mechanisms of meaning formation and the grounding of language into cognitive and sensorimotor categories. To conclude this review, we highlight some of the most promising research directions for current and future studies. These regard the study of neural mechanisms and modularity, the scaling up of the meaning and symbol sets, and the development of coordinated computational models and experimental studies.

Considerable progress has been done in our investigation and understanding of the neuroscience of language (Pulvermuller, 2003). Current knowledge on the neural mechanisms for semantic and syntactic processing can be used to design neurally-inspired neural network models of symbol

grounding. For example, experiment on the modular organization of the organism's neural network was performed using an adaptive agent model for the evolution of language. Through the application of the synthetic brain imaging techniques (Cangelosi & Parisi, 2004), it was possible to design a modular network architecture that closely (though only qualitatively) resembled the organization of the human speaking brain. The artificial neural networks consistently showed the same functional differentiation between motor areas, specialized for verb processing (i.e. action names), and sensory processing areas specialized for the processing of nouns (object names).

This study also supports the research on the modular organization of neural networks for symbol grounding. The work by Greco et al. (2003 – see section 2.2) demonstrated the importance of separating the hidden-output connections that resulted in the modularization of the hidden layer into groups of units specialized for different category classes. Modularity proved essential for the scaling up of previous experiments on the symbol grounding (Cangelosi et al. 2000).

Another important issue for future research regards the scaling up of the cognitive agent's meaning set, lexicons and syntactic rules. Current models mostly deal with few categories/symbols (e.g. about 10 in Greco et al. 2003). Symbol grounding models worthy of addressing language grounding should include a much larger lexicon and more rich syntactic structure.

Finally, future research should look towards a more integrative approach between computational modeling and experimental studies. This will lead to direct comparison of simulation data and experimental data. It will also permit the empirical testing of predictions produced in simulation models. For

example, connectionist models of category learning and naming have indicated that language (and category labels) produces enhanced categorical perception effects (Cangelosi et al. 2000; Lupyan in press). This prediction should be also investigated in cognitive psychology experiments to assess the psychological plausibility of such an effect. In more general term, such coordinated empirical and simulation efforts will test the validity of the modeling results on the symbol grounding in perception, cognition and action.

References

- Barsalou L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22: 577-609.
- Borghi, A.M., Glenberg, A., Kaschak, M. (in press). Putting words in perspective. *Memory and Cognition*.
- Cangelosi A., Coventry K.R., Rajapakse R., Bacon A., Newstead, S.N. (in press). Grounding language into perception: A connectionist model of spatial terms and vague quantifiers. *9th Neural Computation and Psychology Workshop*, Plymouth, September 2004.
- Cangelosi A., Greco A., Harnad S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science*, 12(2): 143-162
- Cangelosi A., Greco A. & Harnad S. (2002). Symbol grounding and the symbolic theft hypothesis. In A. Cangelosi & D. Parisi (Eds), *Simulating the Evolution of Language*, London: Springer, pp. 191-210
- Cangelosi A., Harnad S. (2000). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*. 4(1): 117-142
- Cangelosi A., Parisi D. (2002). *Simulating the Evolution of Language*. London: Springer.
- Cangelosi A., Parisi D. (2004). The processing of verbs and nouns in neural networks: Insights from Synthetic Brain Imaging. *Brain and Language*.
- Cangelosi A., Riga T., Giolito B. & Marocco D. (2004). Language emergence and grounding in sensorimotor agents and robots. *First International Workshop on Emergence and Evolution of Linguistic Communication*, Kanazawa Japan

- Clark A. (1997). *Being There: Putting Brain Body and World Together Again*. Cambridge, MA: MIT Press.
- Coventry, K. R. & Garrod, S. C. (2004). *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*, Essays in Cognitive Psychology Series, Psychology Press, Hove and New York.
- Coventry, K. R., Prat-Sala, M. & Richards, L. V. (2001). The interplay between geometry and function in the comprehension of 'over', 'under', 'above' and 'below'. *Journal of Memory and Language*, 44, 376-398.
- Deacon T.W. (1997). *The Symbolic Species: The coevolution of language and human brain*, London: Penguin.
- Dyer M.G. (1994). Grounding language in perception. In V. Honavar, L. Uhr (Eds.), *Artificial Intelligence and neural networks: Steps toward principled integration*. Boston: Academic Press.
- Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.
- Glenberg A., Kaschak M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9 (3), 558-565.
- Goldstone R. (1994). Influences of categorization of perceptual discrimination. *Journal of Experimental Psychology: General*, 123: 178-200
- Greco A., Riga T., Cangelosi A. (2003). The acquisition of new categories through grounded symbols: An extended connectionist model. In O. Kaynak, E. Alpaydin, E. Oja & L. Xu (Eds.). *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003*. Berlin: Springer, pp. 773-770
- 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. (1996) The origin of Words: A psychophysical hypothesis. In Velichkovsky, B. & Rumbaugh, D. (Eds.), *Communicating meaning: The evolution and development of language*. Lawrence Erlbaum Associates, Mahwah NJ, pp. 27-44.
- Harnad, S., Hanson, S. J. and Lubin, J. (1991) Categorical Perception and the Evolution of Supervised Learning in Neural Nets. In D.W. Powers & L. Reeker (Eds.), *Proceedings of*

Proceedings of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology.

Harnad, S., Hanson, S.J. & Lubin, J. (1994) 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.

Joyce D., Richards L., Cangelosi A., Coventry K.R. (2003), On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. *Proceedings of the 5th Intl. Conference on Cognitive Modeling (ICCM 2003)*. Bamberg

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104, 211-240.

Lupyan G. (in press). How labels augment category representations: A connectionist model. *9th Neural Computation and Psychology Workshop*, Plymouth, September 2004.

Marocco D., Cangelosi A., Nolfi S. (2003), The emergence of communication in evolutionary robots. *Philosophical Transactions of the Royal Society London – A*, 361: 2397-2421

Miikkulainen R. (1994). Integrated Connectionist Models: Building Ai Systems On Subsymbolic Foundations, 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. 483-508.

Nakisa R.C., Plunkett K. (1998) Evolution of a rapidly learned representation for speech. *Language and Cognitive Processes*, 13: 105-127

Pecher D., Zwaan R.A. (Eds.) (in press). *The Grounding of Cognition: The role of perception and action in memory, language, and thinking*. Cambridge University Press

Plunkett, K. & Sinha, C. G. (1992) Connectionism and developmental theory. *British Journal of Developmental Psychology*, 10, 209-254.

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

- Regier T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. Cambridge MA: MIT Press
- Roy D., Hsiao K., Mavridis N., (2003). Conversational robots: Building blocks for grounding word meanings. In: *Proceedings of the HLT-NAACL03 workshop on learning word meaning from non-linguistic data*,
- Searle J.R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3, 417-457
- Steels L., (1999). *The Talking Heads Experiment. Volume 1. Words and Meanings*, Antwerpen.
- Steels L. (2002). Grounding symbols through evolutionary language games. In A. Cangelosi, D. Parisi (Eds.), *Simulating the Evolution of Language*. London: Springer-Verlag.
- Sun R. (2002), Connectionist implementationalism and Hybrid systems, in "Encyclopedia of Cognitive Science", Nature Publishing Group – MacMillan.
- Tijsseling A., Harnad S. (1997). Warping similarity space in category learning by backprop nets. In M. Ramscar, U. Hahn, E. Cambouropoulos, H. Pain (Eds.). *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization*, Edinburgh University, 263-269.
- Ullman, S. (1996). *High-level Vision. Object recognition and visual cognition*. Cambridge, MA; MIT Press.
- Varela, F., Thompson, E., Rosch, E. (1991). *The Embodied Mind*. Cambridge, MA: MIT Press.
- Vogt P. (2002). The physical symbol grounding problem. *Cognitive Systems Research*, 3(3): 429-457
- Zentall T.R., Jackson-Smith P., Jagielo J.A., Nallan G.B. (1986). Categorical shape and color coding by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 12(2): 153-159.

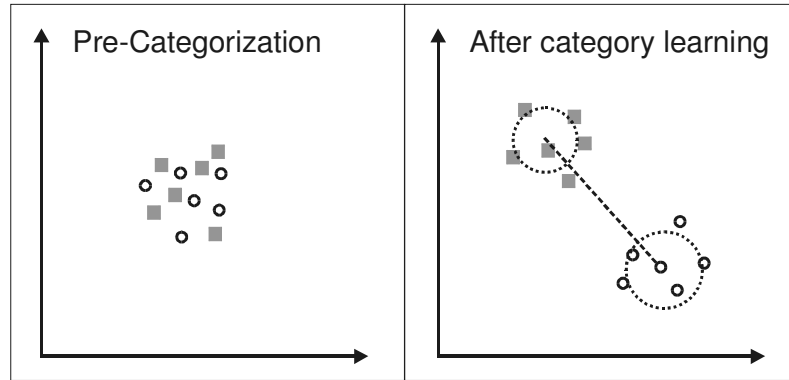


Figure 1: Typical formation of clusters of points (i.e. square and circle categories) during category and language learning. Before category learning (Left), points corresponding to different categories overlap. After categorization, points group in distinct areas (Right).

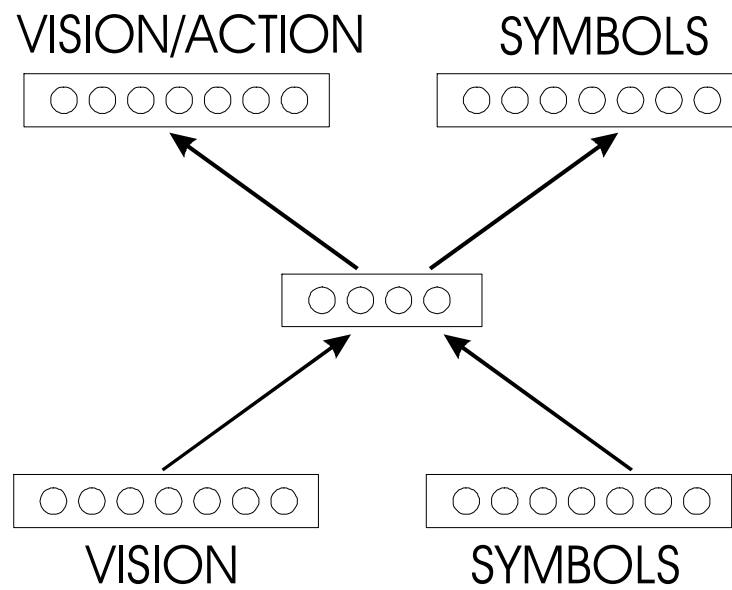


Figure 2. Typical dual-route architecture for connectionist model of symbol grounding

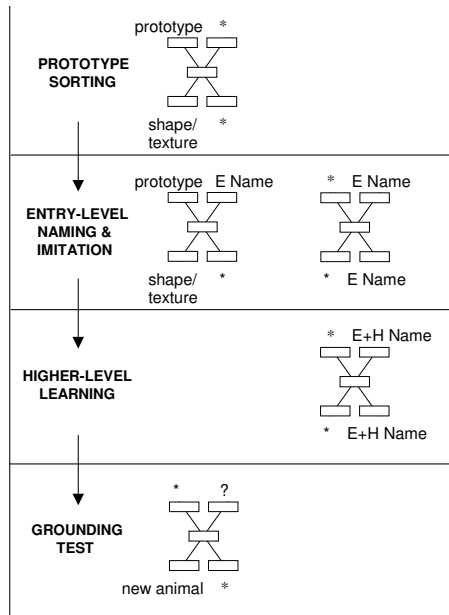


Figure 3. Training stages for the symbol grounding transfer simulation

Words for subject index

Symbol grounding

Language grounding

Cognitive symbol grounding

Categorical perception

Connectionist models

Neural networks

Robotics

Adaptive agents

Sensorimotor abilities

Language acquisition

Language evolution

Embodiment

Computational modeling

Category learning

Category naming