In Cangelosi A & Parisi D (Eds) (2002). *Simulating the Evolution of Language*. London: Springer

Chapter 9

Symbol Grounding and the Symbolic Theft Hypothesis

Angelo Cangelosi, Alberto Greco and Stevan Harnad

The Origin and Grounding of Symbols

Scholars studying the origins and evolution of language are also interested in the general issue of the evolution of cognition. Language is not an isolated capability of the individual, but has intrinsic relationships with many other behavioral, cognitive, and social abilities. By understanding the mechanisms underlying the evolution of linguistic abilities, it is possible to understand the evolution of cognitive abilities. Cognitivism, one of the current approaches in psychology and cognitive science, proposes that symbol systems capture mental phenomena, and attributes cognitive validity to them. Therefore, in the same way that language is considered the prototype of cognitive abilities, a symbol system has become the prototype for studying language and cognitive systems. Symbol systems are advantageous as they are easily studied through computer simulation (a computer program is a symbol system itself), and this is why language is often studied using computational models.

A symbol system is made up by a set of arbitrary "physical tokens" (i.e., symbols) that can be manipulated on the basis of explicit rules (i.e., syntax). Some of the main properties of such a symbol system are: (a) *compositeness*, that is symbols and rules can be recursively composed; and (b) *semantic interpretability*, specifying that the entire system and its parts can be systematically assigned a meaning (Pylyshyn, 1984; Harnad, 1990). Some significant issues arise when studying such symbol systems as a direct metaphor and model of language. These will also have direct implications for the study of the origins and evolution of language. The first issue is to establish exactly what a symbol is, by giving a clear

and unambiguous definition of it. Subsequently, the process of how symbols take their meanings needs to be understood, for example by studying the symbol grounding problem. Finally, questions regarding the evolution of symbols and symbol manipulation abilities need to be addressed.

Definition of a symbol

The definition of a symbol is a yet open and highly debatable issue. Although it is possible to give a precise definition of a symbol in a computational symbol system, it is more difficult when we use this term in the context of language and communication systems. Historically, a semiotic distinction was made between the different constituents of a communication system: icons, indices, and symbols. This distinction, originally introduced by Peirce (1978), is based on the type of reference existing between objects and components of a communication system. Peirce's distinction between icons, indices, and symbols is based on the fact that (1) an "icon" is associated with an object because of its physical resemblance to it, (2) an "index" is associated with an object because of time/space contiguity, and finally (3) a "symbol" is associated with an object due to social convention or implicit agreement and it has an arbitrary shape, with no resemblance to its referent.

Recently, similar distinctions have been proposed. For instance, Deacon (1997) uses a hierarchy of referencing systems based on icons, indices, and symbols. He distinguishes three types of relationships between the means of communication and their referents in the external world and/or in the same communication system. Icons have associations with entities in the world because of stimulus generalization and conventional similarity. Indices are associated with world entities by spatio-temporal correlation or part-whole contiguity. These indexical references are commonly used in animal communication systems. Symbols are characterized by the fact that they have double referential relationships. One type of relationship is based upon the indexical link of a symbol with a referent in the world. The second type of association connects logical and combinatorial relationships with other symbols. For example, in English the verb "to give" is a symbol because it refers to an action, and as a verb it is also associated with nouns that can be used as subject, nouns that can be used as patients, etc. Deacon's definition of symbols is not restricted to language, although symbols express their best potentials in language. There are non-linguistic symbolic tasks, such as the ability to combine elements together using logical combination rules, and general mathematical tasks.

Harnad (1990) distinguishes between three types of mental representations: iconic, categorical, and symbolic. The first two are internal to the individual and non-symbolic. Iconic representations are analogical representations of the proximal sensory projections of distal objects and events. Categorical representations are learned (or innate) feature-detectors that pick out the invariant features of object and event categories from their sensory projections. Elementary symbols are the names of these objects and event categories, assigned on the basis of their non-symbolic categorical representations. Higher-order symbolic representations,

grounded in these elementary symbols, consist of symbol strings (i.e., propositions) mainly describing category membership relationships.

These more recent definitions of symbols share the fact that the real symbolic feature of a communication system relies on the fact that each symbol is part of a wider and more complex system. This system is mainly regulated by compositional rules, such as syntax. In this chapter we will use this characterization of symbols, and in particular we will focus on grounded symbols.

The symbol grounding problem

The symbol systems' property of systematic semantic interpretability implies that any part of the system, and the whole system itself, can be assigned a meaning. Therefore a fundamental question must be asked: How is a symbol given a meaning? This is the problem of symbol grounding. The type of link that exists between symbols and objects is of central importance when using symbol systems as models of language and cognition. Cognitivists avoid this problem by ignoring it or trivializing it. They claim that the autonomous functional module of the symbol system will lately be connected to peripheral devices, in order to see the world of objects to which the symbols refer (Fodor, 1976). In practice, cognitivists often resolve this problem by creating their computational models with another level of yet-to-be-grounded "semantic symbols" that supposedly stand for objects, events, and state of affairs in the world. For example, in a cognitivist model of language it is sufficient to define the set of basic symbols/words (e.g., "John", "Mary", "loves") and some syntactic rules to connect them. Subsequently, each basic symbol will be assigned a meaning (e.g., the meaning of "John" is "the-boy-with-blue-eyes"). This approach is subject to the problem of infinite regression: where does the meaning of the meaningless yet-to-be-grounded "semantic symbols" ("the", "boy", "with", "blue", "eye") come from? It is not enough to have simply a parasitic link of symbols with the meanings in our heads.

This situation is similar to the paradox of the Chinese Room argument (Searle, 1982; Harnad, 1990). Suppose you don't speak Chinese and you are given the task of replying to some questions asked to you in Chinese. If you used a Chinese–Chinese dictionary alone, you could try to solve this task by looking at the symbols defining the Chinese query words and using them to select (i.e., to chain) a new set of Chinese symbols for your answer. In reality, this trip through the dictionary would amount to a merry-go-round, passing endlessly from one meaningless symbol or symbol-string (the *definientes*) to another (the *definienda*), never coming to a halt on what anything meant (Harnad, 1990). Even if we you were able to do this, you would still not have understood Chinese the same way you understand the meaning of English words. In order to use and fully understand Chinese, it is essential that you link (i.e., ground) at least some essential words to your native language[1]. Therefore, we cannot use this task as an experiment for studying

[1] There is a hard version of the Symbol Grounding problem in which the user of a Chinese–Chinese dictionary does not previously know any other language. This person would have to

Chinese linguistic abilities, nor as an experiment of general linguistic abilities. For the same reason, we cannot use a non-grounded symbol system as a model of linguistic and cognitive abilities.

In order to address the problem of symbol grounding, and to propose workable and plausible solutions, a model needs to include an intrinsic link between at least some basic symbols and some objects in the world. A system must use symbols that are directly grounded through cognitive representations, such as categories. This way symbol manipulation can be constrained and governed not by the arbitrary shapes of the symbol tokens, but by the non-arbitrary shapes of the underlying cognitive representations.

The evolutionary origin of symbols

In language origin research it is important to loot at the issues of the evolutionary acquisition of symbol manipulation abilities and their role in the evolution of language. Deacon (1997) has proposed an integrated neural and cognitive theory of the evolution of symbolic and linguistic abilities. His explanation of the origin of language is based on the evolution of his hierarchical referencing system. This theory relies on the symbol acquisition problem. Under normal circumstances[2], only humans have an ability to acquire symbols and language. Animal communication systems are only based on indexical references, i.e., simple object-signal associations. These associations are mostly innate (e.g., monkeys' calls) and can be explained by mere mechanisms of rote learning and conditional learning. Instead, the symbolic associations of human languages have double references, one between the symbol and the object, and the second between the symbol itself and other symbols[3]. These associations between symbols are reflected by the syntactic rules of human languages. When a complex set of logical and syntactical relationships exist between symbols, we can call these "words" and distinguish grammatical classes of words. A language-speaking individual knows that a word refers to an object and implicitly knows that the same word has grammatical relationships with other words. This combinatorial interrelationship between words can lead to an exponential growth of references. When a new word is learned, it can be combined with other pre-existing words to exponentially increase the overall number of meanings that can be expressed.

Deacon (1997) also gives a neural explanation for this distinctive difference between non-symbolic communication in animals and symbolic languages in humans. He uses neurodevelopmental and neuropsychological data to show that

---

solve the task of connecting Chinese symbols between themselves, and also the task of learning to associate meaning to symbols (Harnad, 1990)

[2] Deacon admits that under specific experimental circumstances, some species of animals, mainly apes, can acquire some type of symbol manipulation abilities. For example, in ape language experiments the acquisition of symbolic communication systems has been shown (Savage-Rumbaugh and Rumbaugh, 1978).

[3] Deacon's use of the term "reference" for the association between symbols has been highly criticized (Hurford, 1998). In semiotics, reference is mainly used to indicate the association between a symbol and the entity it *refers* to.

the enlarged prefrontal cortex in humans allows them extra processing abilities, such as symbol acquisition and symbol manipulation abilities.

In the next sections we will present a theoretical and computational framework for explaining the cognitive mechanisms for symbol grounding and symbol acquisition. In the second part of the chapter we will present a model for the evolution of language based on grounded symbols.

Cognitive Theories and Models for Symbol Grounding and Symbol Acquisition

This section describes a cognitive theory that explains the mechanisms for symbol grounding. It is based on a general psychological theory that sees our basic ability to build categories of the world as the groundwork for language and cognition (Harnad, 1987). This theory focuses on hierarchical mental representations and their role in grounding language. It starts from the principle that symbolic representations must be grounded bottom-up in non-symbolic iconic and categorical representations. This system of hierarchical representations has significant advantages. It restricts the problem of direct symbol grounding to a smaller set of elementary symbols. Any combination of these symbols, through syntactic rules, will inherit the semantic grounding from its low-level elementary symbols. Consider the case of learning a new concept from a pure linguistic definition of it. Let's suppose that you do not know what a zebra is, but are familiar with what horse and stripe patterns look like, because you have seen many real horses and striped patterns. You also know two symbols (names): "horse" for the category of horses and "stripe" for the category of striped patterns. Suppose that the following linguistic definition of zebra: `"zebra"="horse"+"stripe"` is introduced. You can immediately understand that the symbol "zebra" must correspond to a combination of (the categorical representation of) horses with (the categorical representation of) stripes. Moreover, when you see an individual zebra, you will be able to identify it as a member of the linguistically learned category zebra. This example shows how easy it is to learn new categories and new grounded names through the combination of the directly grounded names of basic categories. You would be able to fully understand any English sentence simply by knowing a relatively low number of English words[4] and by using an English–English dictionary to look up unknown words (i.e., grounding the meaning of new words in the known basic words used as definitions).

The cognitive mechanism at the core of this hierarchy of representations and bottom-up groundings is categorical perception. Our ability to build categories results in categorical representations that are a "warped" transformation of iconic representations. This feature filtering ability compresses within-category differences and expands between-category distances in similarity space so as to allow a reliable category boundary to separate members from non-members. Categorical perception consists of this compression/expansion effect (Harnad

[4] About 2000 words is the size of the vocabulary of an average English speaking person.

1987). It has been shown to occur in both human subjects (Goldstone 1994; Andrews, Livingstone and Harnad, 1998; Pevtzow and Harnad 1998) and neural networks (Harnad, Hanson and Lubin, 1991; 1995; Tijsseling and Harnad 1997) during the course of category learning.

Connectionism, a recent theoretical and methodological development in psychology and cognitive sciences, proposes the use of artificial neural networks as cognitive models. Neural network models are based on some general structural and functional properties of the brain and permit the modeling of behavioral and cognitive tasks, such as categorization and language (Rumelhart and McClelland, 1986). Various neural networks have proved particularly good at tasks that require the classification of input patterns into separate categories. More neural network models of language have also been developed (Christiansen and Chater, 1999; Elman, 1990). Therefore, connectionism is the natural candidate for learning the invariant features underlying categorical representations and connecting names to the proximal projections of the distal objects they stand for (Harnad, 1990). In this way connectionism can be seen as a complementary component in a hybrid non-symbolic/symbolic model of the mind. Such a hybrid model would not necessarily need an autonomous symbolic module. The symbolic functions could emerge as a consequence of the bottom-up grounding of categories' names in their sensory representations. In this way symbol manipulation would be governed not only by the arbitrary shapes of the symbol tokens, but also by the non-arbitrary shapes of the icons and category invariants in which they are grounded.

In the next two sections we will describe some connectionist models for the phenomena of categorical perception and symbol grounding. Subsequently, we will focus on models of the acquisition of language in which lexicons are directly grounded into sensory and categorical representations.

Models of categorical perception and symbol grounding

We have argued that in a plausible cognitive model of symbol origin, symbolic activity should be conceived as some higher-level process, which is not stand-alone but takes its raw material from non-symbolic representations, i.e., analogue sensori-motor (iconic) in the first instance and then categorical representations. This shift from non-symbolic to symbolic processes is one of the most fascinating aspects to be explained when considering language origins. In this section, we provide a detailed description of the mechanisms for the transformation of categorical perception (CP) into grounded low-level labels, and subsequently into higher-level symbols. Finally, we will describe how new symbols can be acquired from just the combination of already-grounded symbols, a phenomenon called grounding transfer. We shall also show how all such processes can be implemented into a single neural network model.

Neural networks can readily discriminate between sets of stimuli, extract similarities, and categorize. More importantly, networks exhibit the basic CP effect, whereby members of the same category "look" more similar (there is a compression of within-category distances) and members of different categories look more different (expansion of between-categories distances). One of the early

models of CP is Harnad, Hanson and Lubin (1991, 1995). They trained three-layer feed-forward networks to sort lines into categories according to their length. Such lines were represented by 8 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 (coarse representations such as .1 for 0, or .9 for 1 were used, and in some cases boundaries were enhanced by using more distant values for opposite adjacent units). Given that CP is defined as a decrease in within-category inter-stimulus distances and an increase in between-category distances, a baseline for assessing such decreasing or increasing movements is required. The first step in this simulation is simply to allow networks to "discriminate" between different stimuli (to tell pairs of stimuli apart) using a pre-categorization task with auto-associative learning (i.e., networks were trained to produce exactly the same input pattern in the output units). The hidden unit activation vectors were examined and the baseline distances were calculated for each pair of input patterns. After this task the networks were finally trained to sort lines into three categories (short, middle, long) using the back-propagation algorithm.

Such networks not only exhibit successful categorization, which – as we said – is a relatively easy task for neural networks, but they also exhibit the same natural side-effect revealed by human categorization, i.e., CP. In other words, within-category compression and between-categories expansion can be observed both in humans and networks. Another point of interest from CP simulation is that a close scrutiny of hidden representations allows us to propose hypotheses about the factors upon which CP is based. Harnad, Hanson and Lubin (1995) found that the distances between hidden unit pattern representations are already maximized during auto-association (by effect of the baseline discrimination): this could be one source of the maximal interstimulus separation in CP. This separation, however, is not always so clear-cut as to allow linear separability[5], which is a clear-cut categorization, so in some cases there are "bad" or unclear representations, which happen to be close to the plane separating the categories. The back-propagation algorithm, which simulates category learning through supervised feedback, has the effect of "pushing" such unclear representations away from this plane. The result is an improved separation between categories and, at the same time, a smaller distance between representations for the same category; in other words, the CP effect.

Cases where linear separability between categories is attained more easily are not random, but this effect is mostly observed with iconic stimuli. Tijsseling and Harnad (1997), who replicated these results, suggested that CP is strongly related to factors, like the similarity between stimuli. This can lead to different possibilities for the linear separability of representations resulting from simple discrimination (the auto-association phase in the described simulations). When there is either extreme nonseparability or extreme separability, the CP effect is not

[5] Given a space where points represent stimulus dimension values, linear separability is the possibility of drawing a line (in two-dimensional space), a plane (in three-dimensional space), or a hyperplane (in n-dimensional space) to separate points belonging to different categories. In the simulation described, three hidden units were used to represent activation values in a three-dimensional space.

observed. In the former case, due to the fact that the task is too difficult, already at the discrimination level, and in the latter case because it is too easy and there is no need for category learning because categories exist.

CP is a very strong and ubiquitous effect. For example, it was observed in the human categorization of speech sound and of colors (Berlin & Kay, 1969). Nakisa and Plunkett (1998) recently observed it in a simulation of phonological learning. They showed that neural networks could categorize spectral sounds into the phonemes of English. The inputs were sounds, sampled from a single language (the network "native" language) chosen from among 14 natural languages. Nakisa and Plunkett found that networks form similar representations regardless of the particular native language, and such representations exhibit a CP partitioning in the spectral continuum. These networks were trained using genetic algorithms thus showing that CP is not an artifact of particular forms of categorical learning.

The functional role of CP in symbol grounding is clearer as an interaction between discrimination and identification. To discriminate is to distinguish some (undefined) pattern in sensors ("there is something"), and it is a relative judgment because something and something else are logically implied in any distinction. To identify is to assign a stable identity to what has been discriminated; this is revealed by a consistent system reaction when the "same" pattern is presented again. Identification is an absolute judgment, and – since it necessarily comes from a "sameness" judgment – it has a categorical nature. The CP process attains the identification result precisely by acting upon discriminability or separation between different categories. Subsequently, CP is a basic mechanism for providing more compact representations, compared with the raw sensory projections where feature-filtering has already done some of the work in the service of categorization. Identification does not presuppose naming. To react consistently to some category of stimuli does not require being able to say what such stimuli are, e.g., by using labels that act as names for them. However, compact CP representations are more suitable than the sensory ones in the subsequent process of learning labels for categories. These labels, or names for categories, can be further combined into propositions and become symbols.

So far, the process we have described is based on a direct sensorimotor interaction with the environment. Symbols derived from this can be called "grounded symbols". There is, however, a different way of acquiring new categories, namely by combining grounded symbols. The previous example of learning that a zebra is a horse with stripes should be recalled to illustrate this point. Cangelosi and Harnad (in press) called the first method of acquiring categories "sensorimotor toil" and the second method "symbolic theft", to stress the benefit of not being forced to learn from a direct sensory experience for every new category.

A recent model by Cangelosi, Greco and Harnad (2000) simulated this overall process of CP, subsequent acquisition of grounded names, and learning of new high-order symbols from grounded ones (grounding transfer). Three-layer feed-forward neural networks were used (see Figure 9.1), having two groups of input units: 49 units simulating a retina and 6 units simulating a linguistic input. The networks had five hidden units, and two groups of output units replicating the organization of input (retina and verbal output). The retinal input depicted

geometric images (circles, ellipses, squares, rectangles) with different sizes and positions[6]. The activation of each of the verbal input units corresponded to the presentation of a particular category name. The training procedure had four learning stages. In the prototype sorting task, the networks were trained to categorize figures: from input shapes they had to produce the categorical prototype as output. The same networks were subsequently given the task of associating each shape with its name. This task is called "entry-level naming". An imitation learning cycle was also used for the linguistic input and output units. Names acquired this way, however simple, can be considered grounded because they were explicitly connected with sensory retinal inputs.

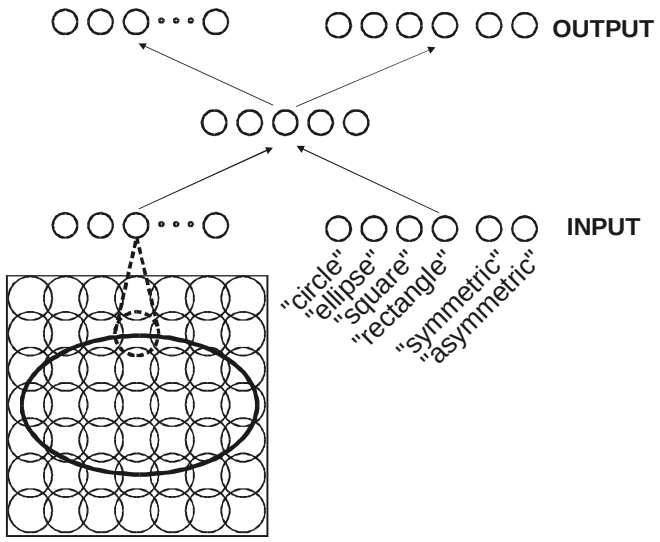

Figure 9.1 Network architecture in the Cangelosi, Greco and Harnad (2000) model. The encoding of output units exactly reflects the structure of input units.

The most interesting part of the simulation is the final stage, where the same networks learn the conjunction of such grounded names (for example, "square" or "rectangle") with new arbitrary names (e.g., "symmetric" or "asymmetric"). This higher-level learning was accomplished by simple imitation learning of the combination of names. It is like teaching "square [is] symmetric" or "rectangle [is] asymmetric". The retinal part of the network was not used during this task. After this procedure, networks were used for the grounding test. The shapes were presented as retinal inputs and output names were checked. In 80% of cases, the networks were able to produce the basic name of each shape and its high-order name (symmetric/asymmetric). As this learning comes from the association of grounded names with new names, the grounding is "transferred" to names that did not have such a property. This is why the process is called "grounding transfer".

[6] Pixels were pre-processed in order to compress information using the receptive fields technique (Jacobs and Kossylyn, 1994).

This model has been extended to use the combination of the grounded names of basic features in order to learn higher-order concepts. The same architecture and training procedure were kept. In this case, networks first learned to sort prototypes of shapes depicting turtles, spots, horses and stripes, and then associate such shapes with their names, thereby grounding them. They were then taught to associate the conjunction of grounded names ("horse"+"stripes" or "turtle"+"spots") with a new name ("zebra" or "sportoise"). Networks were able to name the pictures of zebras they had never seen before with the name "zebra". This was achieved solely on the basis of a linguistic combination of grounded names, which in fact can be considered the effect of a true prepositional definition.

Models of the acquisition of grounded symbols

The path we have followed, starting from stimulus discrimination and leading to categories and grounded names for them, actually describes the first stages of language acquisition. Language is not a common sensorial input. It is not like commonly perceived objects, but has something special, because it acts like a "comment" on the world. We shall now focus on some models of the acquisition of grounded lexicons. The most natural source of inspiration for this kind of simulation is language acquisition in children.

A plausible task to be modeled is the presentation of words along with their referents, something like what happens when a parent shows her child a ball while uttering "ball". One example of such a kind of simulation is the work of Plunkett, Sinha, Møller and Strandsby (1992). They designed a network that had to associate simple pictures with labels. The network architecture was similar to the one used by Cangelosi, Greco and Harnad (2000) and described earlier (Figure 9.1). There were two distinct sensory modalities (retinal and verbal) in the input and output layers, and two hidden layers. An auto-associative learning task was used. During testing, only either the verbal or retinal input was given and the net was requested to give the corresponding other output. As often happens with neural networks, they were not able to correctly perform these tasks at all training stages; performance was obviously poor at early stages and better with intensive training. The interesting result was that performance was not linearly related to the extent of training. It suddenly improved at some point, exhibiting something like a "vocabulary spurt" without any apparent reason. This happened both for comprehension and for production, but at different times. This exactly reflects what is observed in children, or in adults when learning a new language: comprehension precedes production. In other words, at some stage the net 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, sudden improvement would be also observed in production.

A similar network architecture and learning scheme using auto-association were used in a model by Schafer and Mareschal (2001). Networks learned to associate names (coded as phonemes) with objects (arbitrary binary vectors). In this study, the network's capability to distinguishing different names when associated with the same objects was tested. This task models an observation

coming from studies with infants, reporting that younger (8-month-old) infants are able to produce finer phonetic discriminations than older (14-month-old) infants. Schafer and Mareschal use this model to claim that there is a possibility of obtaining similar developmental discontinuities without necessarily hypothesizing a difference in processing strategies. The disruption in low-level processing (discrimination) does not necessarily arise from a higher cognitive load in higher-level (semantic) processing. Note that this model considers the role of discrimination in language acquisition. In this case, a hindrance to name (phonetic) discrimination is subsequent to a well-established association with some object pattern. The perspective can also be reversed, with names seen as emerging from a need for discrimination, as long as they are able to capture differences in perceived objects. This fundamental property reveals the critical role of language as a symbolic tool.

A model by Greco and Cangelosi (1999) investigates the role of linguistic labels in categorization. It focuses on the feature-extraction process which is affected when names already exist for perceived patterns. They trained four-layered networks to associate names with pictures. Names referred to different features of the input (name, color, function) and there were three input conditions (visual features, name, features+name). Analyses of hidden activation show that representations were different in the feature+name condition and these strongly depended on the name. For example, the visual input of a blue pen presented together with the label "blue" activated the *blue* units in the network, while the name "pen" activated the *pen* units. This shows the mediating role of language in categorization. The same model was successfully extended to display this knowledge explicitly by using a further module that re-described the hidden-layer representations using a competitive learning algorithm.

All these toy models show that simulations can fit observed behavioral results or generate reliable predictions about them, but they are obviously simplifications of the real lexicon acquisition task in children. Inasmuch as these models investigate how representations are constructed of name-objects associations, they are also symbol grounding models. However, words acquired by children are not always associated with their referents, but mostly associated with other words. Such models should also be extended to the expressive functions of language, as we know them from developmental psychology studies.

Evolving Grounded Languages

We have established the importance of direct grounding in models of categorization and in language acquisition. Models of the evolution of symbols and language should also include the grounding of symbols into the world. The computational study of the evolution of language has the additional objective of understanding how and when symbol acquisition abilities originated and how the ability to ground symbols in real world meaning emerged. By using models with emerging symbol grounding properties the researcher is released from the task of deciding which meanings to input to the system in the different evolutionary stages. For example, a non-grounded approach would have significant limitations

in investigating the possible existence (or not) of sequential stages of syntactic complexity in the evolution of language. The researcher would have to define an a priori series of stages of semantic complexity upon which syntax would be biased to gradually develop. Instead, in a symbol grounding approach, other autonomous factors, such as the emergence of different stages of behavioral complexity during an organism's adaptation, would be free to affect (or not) the evolution of different stages of syntax complexity.

Current computational models of the evolution of language deal with the symbol grounding problem in different ways. Some models simply avoid the problem by ignoring it, or by assuming that this is not a real problem because it is easily solved in later stages. They think, as cognitivists do, that the researcher will connect the symbols in the simulated communication system with the meaning in the real world. This is the case of models that use a self-referential system where some symbols are used for communication and other symbols are used to represent semantics (e.g., when a list of words is used to denote the list of semantic categories). For example, in models that study the auto-organization of signal-meaning tables (e.g., Steels, 1996; Oliphant and Batali, 1997), the researchers provide the system with a fixed list of *N* symbols denoting "meanings" and *M* symbols denoting communication signals. In other models (e.g., Kirby 2000), the symbols used for communication vary whilst an invariant semantic layer is provided by means of a list of names of semantic categories (e.g., "John", "Mary", "love"). This represents an intermediate layer between the real referents (Mr. John, Ms Mary, the feeling of love) and the communicating symbols associated to them ("blap", "blop", "blup"). However, the missing link between the real feeling of love and the semantic category "love" is what makes symbol grounding interesting. We cannot ignore the implication of this (cognitive) process in the investigation of the evolution of language.

A different group of computational models of language evolution deals directly with the symbol grounding problem using simulated languages with grounded semantics. An example is the embodied approach to the evolution of communication between robots (Steels and Vogt, 1997; this volume). Robots interact in a real environment with physical entities (walls, obstacles, other robots) through sensorimotor devices (video cameras, radio receivers, wheels, arms). This experience constitutes the basis for extracting meanings to communicate. Recently, robotic models have been extended to the Internet and to communication with humans. In Steels and Kaplan's (1999) "Talking Heads" experiment, two robotic agents have the task of describing the location of colored geometrical shapes in a whiteboard. Through various Internet sites, human subjects can be remotely "embodied" in one of the robotic talking heads and can participate in language games. A similar methodology has been applied to the evolution of direct communication between robots and humans. The SONY entertainment robot AIBO is being trained to evolve a lexicon for communicating with humans (Kaplan, 2000).

Additionally, direct grounding of symbols can be obtained through simulation, such as artificial life simulations (Parisi, 1997). This type of model achieves symbol grounding by explicitly simulating the environment in which the communicating agents live and interact. Simulated agents can perform foraging

tasks by learning to classify different sources of energy (e.g., mushroom types) and to communicate their attributes. The agents' behavior is controlled by neural networks, which we have shown to be ideal candidates for dealing with categorization and symbol grounding. This categorization of food provides the basic meaning upon which agents will ground their communication symbols. A detailed example of this approach is presented in the following section. A specific theory of the origin of language based on hearsay and symbolic theft will be tested using the symbol grounded metaphor of a "mushroom world" (Cangelosi and Harnad, in press).

The symbolic theft hypothesis of the origins of words and language

We have already discussed categorical perception and the ability to build categories of objects, events and states of affair in the world. These constitute the groundwork of cognition and language. There are two opposite ways of acquiring categories. First, we can use "sensorimotor toil", in which new categories are acquired through real-time, feedback-corrected, trial and error experience. Secondly, we can use "symbolic theft", in which new categories are acquired through language, based on hearsay from propositions (e.g., through boolean combinations of symbols describing them). In competition, symbolic theft always outperforms sensorimotor toil. It is more efficient than toil because only one propositional description of a new category is enough to learn it. In contrast, repeated experience is required to learn a category by sensorimotor toil. Due to this significant advantage, it has been hypothesized that symbolic theft is the basis of the adaptive advantage of language (Harnad, 1996). However, some basic categories must still be learned by toil to avoid an infinite regress in the symbol grounding problem. The picture of language origins and evolution that emerges from this hypothesis is that of a powerful hybrid symbolic/sensorimotor capacity. Initially, organisms evolved an ability to build some categories of the world through direct sensorimotor toil. They also learned to name such categories. Subsequently, some organisms must have experimented with the propositional combination of the names of these categories and discovered the advantage of this new way of learning categories, stealing their knowledge by hearsay. The benefits of the symbolic theft strategy must have given these organisms the adaptive advantage in natural language abilities. This is infinitely superior to its purely sensorimotor precursors, but still grounded in and dependent on them.

To test this hypothesis of language origin Cangelosi and Harnad (in press) developed a computational model which simulates a community of foraging organisms. They rely on learning categories of foods to survive. Category formation is achieved through toil or theft strategies. The model tests the prediction that acquiring categories through symbolic theft is more adaptive than acquiring them through sensorimotor toil. Moreover, the model should help us to understand the mechanisms central to symbol grounding. For example, it should show that new categories learnt by theft inherit their grounding from the low-level categories.

Computer simulation

The computational model uses the mushroom world scenario (Harnad 1987) to simulate the behavior of virtual organisms that forage among the mushrooms, learning what to do with them. 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 that process the sensory information about the closest mushroom and activates the output units corresponding to the movement, action and call patterns. For each action, the forager first produces a movement and an action/call output using the information about the physical features of the mushroom. The network's action and call outputs are compared with their expected output and this difference is then backpropagated to adjust connection weights. In this way the forager learns to categorize the mushrooms by performing the correct action and call. In the second spread of activation the forager also learns to imitate the call. It only receives the correct call for that kind of mushroom as input, which it must imitate on its call output units. This learning is likewise supervised by back-propagation.

The population of foragers is also subject to selection and reproduction through a genetic algorithm (Goldberg, 1989). The initial population consists of 100 neural networks with a random weight matrix. 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 (eat/mark/return). At the end of their life-cycles, the 20 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.

Adaptive advantages of Theft versus Toil learning

Our hypothesis is that the theft strategy is more adaptive (i.e., results in greater fitness and more mushroom collection) than the toil strategy. To test this, we compare foragers' behavior for the two learning conditions. In the first simulation, two experimental groups were directly compared: Toil and Theft. 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: EAT and MARK. The return behavior, and its name are not yet taught. From generation 200 to 210, 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. 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, we compare foragers' behavior for the two conditions by counting the number of AB mushrooms that are correctly returned to. Thieves successfully return to more AB mushrooms (55) than Toilers (44). 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 is to see how they fare in competition against one another. We performed some competitive simulations. Again, foragers live for two life stages. In the first, all learn to eat and mark through Toil. In the second life stage, the foragers are 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. In the present ecology, the assumption is that mushrooms are abundant and that fitness efficiency only affects the selection of the top 20 foragers. Figure 9.2 shows the proportion of Thieves in the overall population of Theft vs. Toil. Thieves gradually come to outnumber Toilers, so that in less than 10 generations the whole population is made up of Thieves.

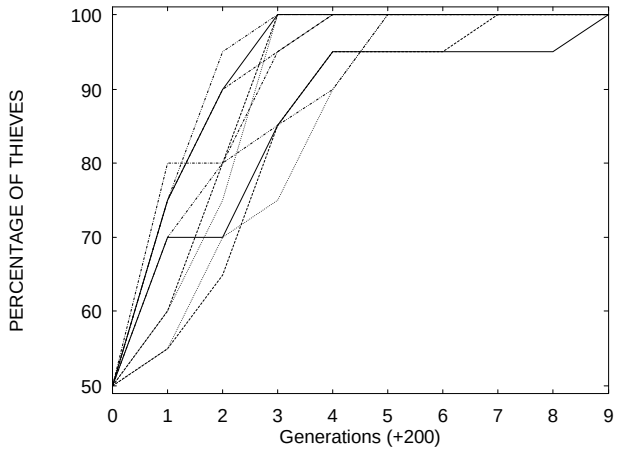

Figure 9.2 Proportion of Thieves in the ten competitive simulations.

The direct competition between Toilers and Thieves has been studied in more detail in other simulations. In one study we varied the availability of mushrooms to see the effects of scarcity/abundance of mushrooms for competition. In another simulation, kinship relationships determined the choice of the listener organism to which the names of mushrooms was vocalized. Results show that when the scarcity of the mushrooms is varied, Theft beats Toil provided there are plenty of mushrooms for everyone. However, when the mushrooms are scarce and vocalizing risks losing the mushroom to the Thief, Toil beats Theft and the

foragers are mute. Further studies analyzing kinship showed that under conditions of scarcity vocalizing only to relatives beats vocalizing to everyone.

All these results support the original hypothesis that a Theft learning strategy, based on language, is much more adaptive that a Toil strategy. This adaptive advantage could be basis for the origin of language and its adaptive advantage.

Categorical perception effects

These computational models are also useful in the investigation of the changes that communication and linguistic abilities cause in the organism. We have already stressed the importance of internal categorical representation in the grounding of symbols. Previously we showed the compression of within-category distances and the expansion of between-category distances in categorization and naming tasks. Now we will show how these phenomena are also present in the model of the evolution of Theft learning and communication. We will study the changes in the foragers' hidden-unit representations for the mushrooms to determine internal changes during Toil and Theft. We compare categorical representations in four different experimental conditions: (1) Pre-learning, for random-weight networks before learning; (2) No-return, for foragers' networks that were only taught to eat and to call EAT, and to mark and to call MARK, (3) Toil-return, for networks that also learned to return and to call RETURN with feature input, and (4) Theft-return for learning to return from calls alone.

We recorded the Euclidean distances between and within categories using the coordinates of the five hidden unit activations. At the end of each simulation, the five fittest foragers in each condition were tested based on the measurement of within- and between-category distances. For each type of distance, there are four means for the distances between the internal representations of the Do-nothing (neither Mark nor Eat nor Return), Eat only, Mark only, Eat+Mark+Return. The average within-category distances in three experimental conditions are shown in Table 9.1. Statistical tests on these data suggest that within-category distances decrease significantly from Pre-learning to No-return to Toil. As expected, the greatest decrease is between the (random) Pre-learning and all the post-learning nets. When we compared the four types of categories, all means differed from each other except the Eat and Mark within-distances. That is, the within-category distance for Eat and Mark are the same, whereas the within distance of Do-nothing is the greatest and that of Return the smallest. These results are consistent with categorical perception effects. There is a compression of the category from the pre-learning condition to all other post-categorization cases.

Table 9.1 Table of means for the within-category distances. Values for the Theft condition are not reported because the distance is always 0 (all ten samples of mushrooms use the same call input)

CATEGORY	Pre-learning	No-return	Toil-return
Do-nothing	0.34	0.16	0.14
Eat	0.32	0.14	0.12
Mark	0.30	0.13	0.12
Eat+Mark(+Return)	0.29	0.11	0.09

The between-category distances amongst the four categories in the different experimental conditions are described more clearly in Figure 9.3. The quadrilateral represents the 2D projections of the between-category distances in the four conditions (= four sides of the quadrilateral shape). All distances, except Eat vs. Mark, are directly comparable and reflect the actual Euclidean distances between categories. The figure shows an expansion of the between-category distances from Pre-learning to Theft Learning. The thin dashed rectangle refers to the between-category distances before learning (random weights). The thick dashed line represents what they look like after Toil learning of Eat and Mark without Return. The thin continuous line refers to the Toil learning of Eat and Mark, with Return, and the thick continuous line is for Theft learning of Return. First, this chart shows the expected categorical perception effect of differentiation between (the centers of) categories from the Pre-learning condition to the other post-learning cases. What is new and more important is the effect of Theft learning relative to the other learning conditions/categories. The distances between Return and all the other categories (Return vs. Eat, Return vs. Mark, Return vs. Do-nothing) are the highest in the Theft condition. This suggests that Theft learning not only is more advantageous for survival (Thieves collect more return mushrooms that Toilers), but it also "optimizes" the internal categorical representation by the categorical effects of within-category compression and between-category expansion. That is, language learning is based on categorization, but in return it improves categorical learning.

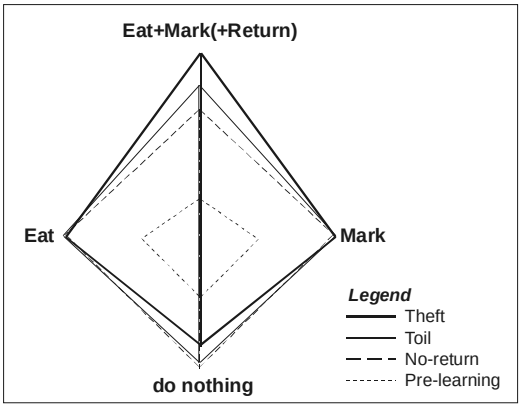

Figure 9.3 Two-dimensional projections of the between-category distances (quadrilateral sides) in the four learning conditions.

Summary

In this chapter we have focused on the simulation of the acquisition and evolution of grounded languages. We have given a definition of symbol based mainly on cognitive (Harnad, 1987) and neurally-related semiotic factors (Deacon, 1997).

The real symbolic feature of communication relies on the fact that each symbol is part of a wider and more complex system. This is mainly regulated by compositional rules, such as syntax. Subsequently, the problem of symbol grounding in cognitive models was illustrated. Psychologically plausible models of language and cognition should include an intrinsic link between at least some basic symbols and some objects in the world. These basic symbols must be directly grounded in cognitive representations, such as categories. This way symbol manipulation can be constrained by the non-arbitrary shapes of the underlying cognitive representations.

Various computational models of categorization, symbol grounding transfer, and language acquisition have been described. They are primarily based on the use of neural networks. These models can easily abstract from similarities between stimuli and achieve categories. Moreover, they can associate names with categories. They exhibit the basic categorical perception effect, whereby internal representations of members of the same category look more similar and members of different categories look more different. In the evolutionary model of the symbolic theft acquisition of categories and language, it has been shown that such cognitive factors for category learning and symbol grounding can be integrated to test hypotheses on the evolution of language.

We suggest that the inclusion of direct grounding in simulation models of the evolution of syntax can improve their potential to explain the emergence of linguistic and cognitive abilities. The simulation approach proposed here, and other methodologies such as the robotic modeling of the evolution of communication (see next chapter), are clear examples of how symbols can directly and autonomously ground their meaning.

References

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

Berlin B, Kay P (1969) *Basic color terms: Their universality and evolution*. University of California Press, Berkley

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: 143-162

Cangelosi A, Harnad S (in press) The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*

Christiansen MH, Chater N (1999) Connectionist natural language processing: The state of the art. *Cognitive Science*, 23: 417-437

Deacon TW (1997) *The Symbolic Species: The coevolution of language and human brain* London: Penguin

Elman JL (1990) Finding structure in time. *Cognitive Science*, 14: 179-211

Fodor JA (1976) *The Language of thought*, Thomas Y Crowell, New York

Goldberg DE (1989) *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, Reading MA

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

Greco A, Cangelosi A (1999) Language and the acquisition of implicit and explicit knowledge: A pilot study using neural networks. *Cognitive Systems*, 5: 148-165

Harnad S (ed) (1987) *Categorical perception: The groundwork of cognition*. Cambridge University Press, New York

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

Harnad S (1996) The origin of words: A psychophysical hypothesis. In: Velichkovsky BM, Rumbaugh DM (eds) *Communicating meaning: The evolution and development of language*. Lawrence Erlbaum Associates, Mahwah NJ

Harnad S, Hanson SJ, Lubin J (1991) Categorical perception and the evolution of supervised learning in neural nets. In: Powers DW, Reeker L. (eds) *Proceedings of the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology*

Harnad S, Hanson SJ, Lubin, J (1995) Learned categorical perception in neural nets: Implications for symbol grounding. In: Honavar V, Uhr L (eds) *Symbol processors and connectionist network models in artificial intelligence and cognitive modelling: Steps toward principled integration*. Academic Press, pp 191-206

Hurford J (1998) Review of Terrence Deacon, 1997 The Symbolic Species: The co-evolution of language and the human brain. *The Times Literary Supplement*, October 23rd, 1998, 34

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

Kaplan F (2000) Talking AIBO: First experimentation of verbal interactions with an autonomous four-legged robot. In: Nijholt A, Heylen D, Jokinen K (eds) *Learning to behave: Interacting agents*. CELE-TWENTE Workshop on Language Technology, pp 57-63

Kirby S (2000) Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In: Knight C, Studdert-Kennedy M, Hurford J (eds) *The evolutionary emergence of language: Social function and the origins of linguistic form*. Cambridge University Press, pp 303-323

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

Oliphant M, Batali J (1997) Learning and the emergence of coordinated communication *Centre for Research in Language Newsletter*, 11(1)

Parisi D (1997) An Artificial Life approach to language. *Mind and Language*, 59, 121-146

Peirce CS (1978) *Collected Papers (Vol II: Element of Logic)*. In: Hartshorne C, Weiss P (eds), Belknap Cambridge, MA

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

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

Pylyshyn ZW (1984) *Computation and cognition*. MIT Press, Bradford Books, Cambridge MA

Rumelhart DE, McClelland JL, the PDP Research Group (eds) (1986) *Parallel distributed processing: Explorations in the microstructure of cognition (Vol 1: Foundations)*. MIT Press, Cambridge MA

Savage-Rumbaugh S, Rumbaugh DM (1978) Symbolization, language, and Chimpanzees: A theoretical reevaluation on Initial language acquisition processes in four Young Pan troglodytes. *Brain and Language*, 6: 265-300

Schafer G, Mareschal D (2001) Modeling infant speech sound discrimination using simple associative networks. *Infancy*, 2(1)

Searle JR (1982) The Chinese room revisited. *Behavioral and Brain Sciences*, 5: 345-348

Steels L (1996) Self-organising vocabularies. In: Langton CG, Shimohara K (eds) *Proceedings of the ALIFE V*. MIT Press, Cambridge MA, pp 179-184

Steels L, Kaplan F (1999) Collective learning and semiotic dynamics. In: Floreano D, Nicoud JD, Mondada F (eds) *Proceedings of ECAL99 the Fifth European Conference on Artificial Life (Lecture Notes in Artificial Intelligence)*. Springer-Verlag, Berlin, pp 679-688

Steels L, Vogt P (1997) Grounding adaptive language games in robotic agents. In: Husband P, Harvey I (eds) *Proceedings of the Fourth European Conference on Artificial Life*. MIT Press, Cambridge MA, pp 474-482

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