# Categories and Word Meanings in Adaptive Organisms

Domenico Parisi
Institute of Psychology
National Research Council, Rome (Italy)

Daniele Denaro
Department of Computer Science
University of Ancona, Ancona (Italy)

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

**Abstract**

Categories and word meanings tend to be interpreted as single entities that are activated in the mind when an individual categorizes the world or uses language. We argue that this view is mistaken by using models based on neural networks and adaptive behavior simulations. Connectionist simulations of populations of organisms that live and reproduce in an environment show that categories and word meanings are open and changing collections of activation patterns in the network's internal units rather than single activation patterns.

**Corresponding author**:
Angelo Cangelosi
Centre for Neural and Adaptive Systems
University of Plymouth
7 Kirkby Place
Plymouth PL4 8AA

tel. 44 1752 232619
fax 44 1752 232540
email: angelo@soc.plym.ac.uk

**Keywords**: adaptive behavior, artificial life, collective/social behavior, communication, agent-environment interaction, cognition, neural networks

**Module**: ARAB - Autonomous Robotics and Adaptive Behaviour
**Presentation**: oral

# 1 Categories and word meanings as mental entities

When we describe and explain the behavior of organisms, in particular humans, we are likely to use at some point or another such expressions as categories (or concepts) and (word) meanings. Folk psychology and the use of language may induce us to believe that categories and word meanings are entities, either in the mind (if we are psychologists) or in a Platonic world of ideas (if we are philosophers) or in a dictionary (if we are linguists or lexicographers). But if we wish to understand and explain the behavior and cognition of organisms we should abandon this well-entrenched belief and begin to think that there is no single entity in the mind which corresponds to a category or a word meaning.

Categories and word meanings should be viewed not as single entities but as open and constantly changing sets of entities in the mind (or, rather, brain) of an organism. This view of categories and word meanings can better be explicated using a distributed connectionist model in an adaptive behavior simulation framework [2]. Neural networks allow us to study distributed representations of concepts and word meanings. Moreover, the simulation of adaptive organisms interacting with world entities and communicating about these entities, give us a potential modeling approach for studying these problems. In this paper we describe simulations of evolving populations of artificial organisms. We will use these simulations to support our alternative view of concepts and word meanings.

# 2 Evolution of categories in populations of organisms

Imagine a population of organisms living in an environment that contains 10 edible and 10 poisonous mushrooms. Edible mushrooms are functionally defined as those mushrooms by eating which an individual increases its survival and reproductive chances. Poisonous mushrooms are those mushrooms that when eaten decrease the fitness of the individual. Since each mushroom is perceptually different from all other mushrooms, a well adapted organism must possess two abilities: it must be able to categorize or recognize mushrooms as either edible or poisonous and it must be able to respond to edible mushrooms by approaching them and to poisonous mushrooms by going away from them.

The nervous system of an organism is modeled by the feedforward neural network shown in Figure 1a. For organisms that can also communicate about mushrooms, we will use the architecture in Figure 1b.

In the input layer one unit encodes the location of the (single) mushroom which at any given time is nearest to the organism. The remaining 10 input units encode the perceptual properties of the mushroom. The binary patterns of edible mushrooms are derived by flipping some bits of the string 1111100000. The poisonous mushrooms are derived in the same way from the opposite pattern 0000011111. The output layer contains 2 units encoding the organism's movements.

We apply a genetic algorithm to a population of such organisms starting from a first generation with random connection weights and selecting for reproduction the individuals with highest fitness. The fitness is the number of edible mushrooms

eaten minus the number of poisonous mushrooms eaten. Small random changes are applied to the connection weights inherited by the offspring. The simulation continues for 1000 generations. Average fitness increases evolutionarily and at the end of evolution organisms are able to discriminate and approach/avoid any edible/poisonous mushroom. Model details and results are fully described in [1].

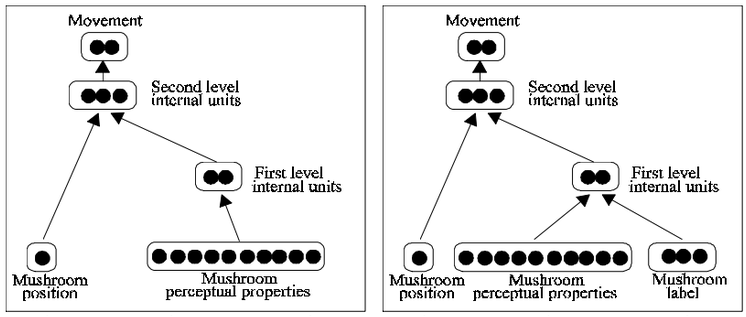



Figure 1. Network architecture of categorizing organisms (a) and of communicating organisms (b)

Notice that even if the ability to categorize the mushrooms appropriately must necessarily be attributed to our evolved organisms, they categorize implicitly rather than explicitly the mushrooms. Categorization is explicit when the organism generates a single output which is the same in response to all relevant inputs. Categorization is implicit when the organism's response can be called "the same" at some high level of description but the actual behavior exhibited by the organism varies with the circumstances. Our organisms respond to edible mushrooms by approaching and reaching them and to poisonous mushrooms by avoiding them. "Approaching" and "avoiding" are high level descriptions of behavior. But at a lower level what our organisms do is moving toward edible mushrooms in ways that vary as a function of the location of the mushroom relative to the organism, the organism's orientation, and perhaps even the perceptual properties of the mushroom. In the same way "avoiding" a poisonous mushroom can mean different things in different circumstances. In fact, our organisms (almost) never generate an identical succession of outputs in response to mushrooms that are categorized as belonging to the same category.

Since the organisms of the later generations are apparently able to implicitly categorize the mushrooms, we can ask how they succeed in doing so. Returning to the question raised in the preceding Section, we can ask: Do they evolve categories as single entities that can be observed inside their neural networks and that are evoked by all encountered mushrooms of the same class; in our case, one category for edible mushrooms and one category for poisonous mushrooms? Or they succeed in categorizing the mushrooms appropriately without developing categories as single entities?

We can try to answer this question by looking at the internal representations of the mushrooms. We call “internal representation” of a mushroom the activation pattern observed in the first layer of internal units when the input units encode the perceptual properties of an encountered mushroom.

Since there are only two units in the first layer of internal units, the internal representation of a mushroom is a vector of two numbers in the interval between 0 and 1. We can therefore represent all possible internal representations of mushrooms as points in a bidimensional space with the activation value of one unit in the abscissa and the activation value of the other unit in the ordinate.

If we examine the internal representations of the 20 mushrooms in the neural networks of the individuals of the last generation, i.e., individuals that are clearly able to categorize the mushrooms, what we find is that in each individual network not only the mushrooms belonging to the two different categories but also the mushrooms belonging to the same category tend to evoke different internal representations. The edible (or the poisonous) mushrooms tend to evoke not a single activation pattern in the first layer of internal units, i.e., a single point in the space of internal representations, but a collection of points.

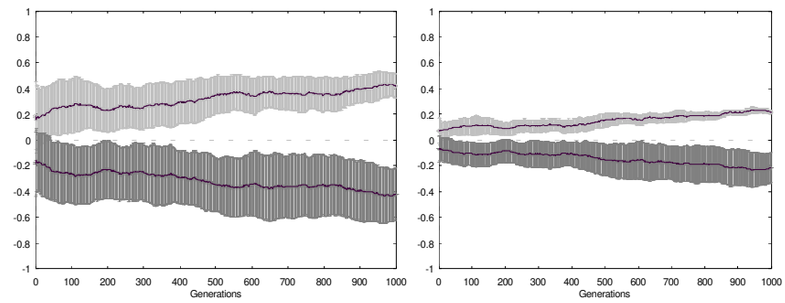
We can conclude that the ability to categorize appears not to be based on the existence of some single entity (a category) in the mind (neural network) of our organisms. What is it based on then? What are the properties of the collections of points that correspond to a category?

We can answer this question by comparing the internal representations evoked by the mushrooms in the individuals of successive generations, that is, as the ability to implicitly categorize the mushrooms emerges in the population. In particular, we can measure the distance between the centers of the two collections of points of edible and poisonous mushrooms, respectively, and we can determine the boundaries of the two collections of points in the space of internal representations across the 1000 generations. What we find is that the distance between the centers of the two collections of points increases evolutionarily and, furthermore, there is no overlap but an increasing distance between the boundaries of the two collections (Figure 2a).

We conclude that the organism's ability to categorize is based on the possession of categories as collections of entities (activation patterns of a set of network units) with the following properties: (a) the center of each collection of points corresponding to a category in the abstract space of activation patterns tends to be separated and distant from the center of other collections (other categories); (b) the collection of points representing a category tends not to overlap with other collections; (c) the different collections of activation patterns support the appropriate behavioral responses (in our case, approaching edible mushrooms and avoiding poisonous mushrooms) to all instances of each category.

Notice that the connection weights linking the first to the second layer of internal units do not change this state of affairs. One might think that although categories are collections of entities at the level of the first layer of our network architecture, the connection weights linking the first to the second layer transform them into single entities at the level of the second layer. This is not so. We have analyzed the activation patterns evoked by the 20 mushrooms in the 3 internal units of the

second layer in controlled conditions, i.e., by blocking the information from the input unit encoding the position of the mushroom and allowing only the information from the 10 inputs units encoding the mushrooms's perceptual properties to determine the activation pattern. The results are shown in Figure 2b.




(a) (b)

Figure 2. Internal representations of edible (light grey) and poisonous (dark grey) mushrooms. The distance between the two midlines corresponds to the average inter-category distance. The width of each line corresponds to the standard deviation of the intra-category distances. Representations observed in the first (a) and second (b) hidden layer.

As this Figure makes clear, we still observe collections of activation patterns for the mushrooms of the same category at the second level of internal units, although the distance between the centers of the two collections of points and the sizes of the two collections are less great than those of Figure 2a (first level of internal units).

Categories are not only collections of entities in the mind but they are open and changing collections of entities. An organism can be said to possess a given category if the organism is able to respond appropriately to new instances of the category, that is, instances not encountered before. New instances evoke new activation patterns in the internal units. The new instances are recognized as members of a category not because they evoke the same activation pattern of past instances but because they evoke activation patterns with the properties (a)-(c).

# 3 Word meanings

One might accept the notion that the categories implicitly used by organisms in their interactions with the environment are not single entities in their mind (neural network) but rather open collections of entities with particular properties. However, the same analysis appears to be less easily applicable to word meanings. Word meanings, it would be seen, cannot but be single entities in the mental lexicon of a language-using organism. In this section we will try to show that this is not so. Word meanings, like categories, are collections of varying internal representations, not single representations.

Imagine that our organisms have a form of primitive language. More specifically,

when an individual encounters a mushroom the individual is also exposed to a "signal", defined as an additional sensory input from the environment (presumably from a conspecific) with the property that this additional sensory input is always the same for all edible mushrooms and for all poisonous mushrooms but very different in the two cases. Therefore, the "language" of this population includes only two "words", one for edible and the other for poisonous mushrooms. The signal is encoded in an additional set of input units of the neural network of the organism (Figure 1b). These additional units project their connections to the first layer of hidden units, where the internal representations of mushrooms are formed. We have utilized this scenario to study the evolutionary emergence of a shared language among organisms [1, 2]. This approach follow the recent use of adaptive behavior models to simulate the evolution of language and communication [5]. Here we will use some of these language evolution simulations to analyze the features of the evolved word meanings.

The organisms with language receive signals as sensory input and insofar as they react appropriately to this sensory input they can be said to be able to understand the signals. Since the input units encoding the signals send their connections to the first layer of internal units we can interpret the activation pattern observed in these internal units as the meaning of the signal - what we have called word meaning.

A signal can evoke its meaning in the internal units in two different conditions: (a) the activation pattern in the internal units (the meaning) is caused by both the signal and the perceptual properties of the encountered mushroom (when the mushroom is near to the organism); (b) the activation pattern is caused by the signal alone (when the organisms cannote see the mushroom but receives only the word input). We ask: What are the characteristics of the meaning of the signal in the two cases?

When the signal is perceived by the organism together with the perceptual properties of the mushroom, the activation pattern observed in the internal units varies with the different perceived mushrooms belonging to the same category. Therefore, in this case we have the same situation we found with categories. One and the same signal evokes a meaning which is not always the same single entity (internal representation) but a collection of internal representations. The meaning of each of the two signals is a collection of activation patterns with basically the same properties (a)-(c) we have seen before for categories.

However, when the signal is perceived alone, i.e., as a substitute for its referent (a distant mushroom), the meaning of the signal necessarily is a single entity. The signal by itself always evokes the same internal representation since neural networks cannot but respond exactly in the same way to the same input. Hence, at least when signals are perceived by organisms in the absence of their referent, it seems to be correct to say that (word) meanings are single mental entities; in our case, always the same activation pattern in a set of internal units.

We think this conclusion would not be justified for at least three reasons. First, as models of a biological system (the nervous system) neural networks should be functioning on a constant baseline of statistical noise. (This can be realized by adding in each cycle some ramdomly chosen quantity to connection weights or activation levels, etc.) Unlike symbol manipulation systems, neural networks can

benefit from limited noise rather than be harmed by it (cf., for example, [4]). The addition of noise would make the activation patterns evoked by one and the same linguistic signal, even when perceived alone, different in different occasions and, therefore, the meaning of the signal would be a collection of different activation patterns.

Second, one and the same signal (for example, some particular English word) is not encoded as the same sensory, say acoustic, input in all occasions. Some of this variability may be eliminated by a purely acoustical preprocessing of the sensory signal by some additional layer of network units interposed between the input units and our first layer of internal units. However, some physical differences among different pronounciations of the same word may be important and preserved in the internal representation of the word.

Third, and most important, words are always experienced in particular contexts, i.e., together with additional input, and the different contexts are reflected in the activation pattern evoked by any particular word as the word's meaning. A particularly important type of context that accompanies the perception of linguistic signals in the real life of humans is represented by the words that precede any particular word in the succession of words that constitute a phrase or sentence or piece of text. The words that precede any perceived word in a sentence can be called context if we assume that the neural networks of humans include what Elman calls "context units" [3]. Context units are special units that, in any input/output cycle, copy the activation pattern evoked in the network's internal units by any particular word in a succession of words. These context units send their connections to the internal units of the network. Therefore, the activation pattern evoked by the currently perceived word in the network's internal units is determined by both the perceived word and the activation patterns evoked by the previous words - and, as a consequence, is likely to be different as a function of previous context.

# 4 Conclusions

Perhaps because we refer to categories and word meanings using well defined linguistic signals (words) we tend to believe that categories and word meanings are well defined single entities in the mind of organisms that categorize the environment and use language. Distributed connectionist models of the nervous system of organisms living in an environment and using categories and language as adaptations to the environment show that this view of categories and word meanings is mistaken. More specifically, our simulations show that categories and word meanings are open and always changing collections of activation patterns rather than single entities in the nervous system (neural network). Collections of activation patterns are virtual, not actual entities, in the brain. What is observed at any particular time inside a neural network is the particular activation pattern evoked by the particular mushroom perceived by the organism. Categories are virtual collections of activation patterns with properties that allow the organism to respond adaptively to environmental input.

# References

[1] Cangelosi A, Denaro D, Parisi D. Language evolution as an aid to categorization. *Technical Report*, Institute of Psychology, CNR, Rome. 1997

[2] Cangelosi A, Parisi D. The emergence of a "language" in an evolving population of neural networks. *Connection Science* (in press)

[3] Elman JL Finding structure in time. *Cognitive Science*, 1990; 14:179-211.

[4] Hanson SJ. A stochastic version of the delta rule. In: Forrest S (ed) *Emergent Computation*. MIT Press, Cambridge-MA. 1991

[5] Steels L. Synthesizing the origins of language and meaning using co-evolution, self-organisation and level formation. In: Hurford J. & Studdert-Kennedy (eds) *Evolution of Human Language*. Edinburgh University Press. 1997

| | |
|---|---|
| Filename: | PARISI.RTF |
| Directory: | L:\USERS\ACANGELO |
| Template: | C:\MSOFFICE\WINWORD\TEMPLATE\NORMAL.DOT |
| Title: | This is an Example of a Title with only |
| Subject: | |
| Author: | Roger eriksson |
| Keywords: | |
| Comments: | |
| Creation Date: | 03/23/98 2:52 PM |
| Revision Number: | 1 |
| Last Saved On: | 03/23/98 2:52 PM |
| Last Saved By: | Angelo Cangelosi |
| Total Editing Time: | 0 Minutes |
| Last Printed On: | 03/23/98 2:55 PM |

As of Last Complete Printing

| | |
|---|---|
| Number of Pages: | 8 |
| Number of Words: | 2,913 (approx.) |
| Number of Characters: | 16,608 (approx.) |