Proposal for a Ph.D. thesis in cognitive science

The origins of symbol use and communication in preverbal categorization: A situated perspective

Ulaş Türkmen

In the doctorate study outlined in this preliminary proposal, I would like to address the following general questions:

1. What are the fundamental processes underlying categorization and symbol acquisition in animals and humans?
2. How can studies on categories and symbols profit from using mobile robots as modeling tools?

More specifically, I would like to combine research on categorization in embodied agents and results from behavioral psychological studies on symbol acquisition to arrive at a model of categorization, symbol acquisition and basic communication phenomena in animals and humans, implemented on robots. To achieve this objective, I will build on results and methods from situated cognitive science, psychology and philosophy.

Scientific background and state of the art

Problems of symbolic artificial intelligence

Symbol use and language is a key feature of human cognition. These two intertwined subjects have been the primary subject matters of artificial intelligence and cognitive science since their beginning. Most investigations on cognition build on the idea of *physical symbol systems*. Newell & Simon (1976, p.113) defined physical symbol systems as follows:

> "A physical symbol system consists of a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure). Thus, a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another). At any instant of time the system will contain a collection of these symbol structures. Besides these structures, the system also contains a collection of processes that operate on expressions to produce other expressions: processes of creation, modification, reproduction and destruction." [1]

[1] A more constraining definition was later proposed by Harnad (1990, p.1):
A symbol system consists of
1. a set of arbitrary "physical tokens" (scratches on paper, holes on a tape, events in a digital computer, etc.) that are
2. manipulated on the basis of "explicit rules" that are
3. likewise physical tokens and strings of tokens. The rule-governed symbol-token manipulation is based
4. purely on the shape of the symbol tokens (not their "meaning"), i.e., it is purely syntactic, and consists of
5. "rulefully combining" and recombining symbol tokens. There are
6. primitive atomic symbol tokens and
7. composite symbol-token strings. The entire system and all its parts – the atomic tokens, the composite tokens, the syntactic manipulations (both actual and possible) and the rules – are all
8. "semantically interpretable", i.e. a meaning can systematically be assigned to the strings of symbols (e.g., as standing for objects, as describing states of affairs, etc.).

This definition was accompanied by the *physical symbol system hypothesis* which set the rules for a research program (Newell & Simon, 1976, p.116):

> "A physical symbol system has the necessary and sufficient means for general intelligent action. By necessary we mean that any system that exhibits intelligence will prove upon analysis to be a physical symbol system. By sufficient we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence. By general intelligent action we wish to indicate the same scope of intelligence as we see in human action."

This idea is a cornerstone of the framework that is now commonly referred to as "classical" AI. The physical symbol systems paradigm, in addition to informing software modeling and technically-oriented work in the field, also laid the grounds for the philosophical and empirical psychological work that followed, and was one of the core assumptions of cognitive science research right from the start (Haugeland, 1981; Fodor, 1975).

However, certain problems started surfacing after intensive research of more than a decade (Dreyfus, 1972). One particularly obvious problem was that symbolic AI took the reference of the used symbols to other entities (e.g. "in the world") for granted. The inherently problematic nature of this premise was spelled out most clearly in the well-known *Chinese room argument* presented by Searle (1980). His argument pointed to the obvious difference between an English-speaking human who generates "answers" to questions in Chinese just by executing a set of rules for manipulating Chinese symbols he does not understand, and a native speaker of Chinese answering with understanding. Harnad (1990) coined a name for the underlying problem: *symbol grounding*.[2] His phrasing of the problem is as follows (Harnad 1990, p. 335):

> "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols?"

Embodied artificial intelligence

Dissatisfactions with traditional robotics and AI, especially with their inability to construct systems that can autonomously interact with complex real-world environments, led to a different approach towards achieving artificial intelligence. Many names have been given to this new approach, e.g. "behavior-based AI", "new AI", "nouvelle AI" or "embodied AI". Embodied AI considers agent-environment interaction (rather than disembodied and detached problem solving) to be the core of cognition and intelligent behavior (cf. e.g. Agre & Chapman, 1990; Brooks, 1991a,b; Pfeifer & Scheier, 1999; Maes, 1994; Agre & Rosenschein, 1995; Arkin, 1998). The aim of this approach is to build artificial agents that successfully interact with and adapt to environments that are previously largely unknown to them (and to their human designers). While classical AI mainly focused on disembodied, detached computer programs, the emphasis is now put on embodied agents that interact with real-world environments, i.e. mobile robots. Through their "embodiment", such agents are continuously coupled to the current real-world situation. Researchers in embodied AI believe that *embodiment* and *situatedness* are also main features of natural intelligent agents. One reason for their importance is that they could form the basis of a solution to the problem of

[2] Several authors have generalized the problem and given it other names: "representation grounding" (Chalmers, 1992), "concept grounding" (Dorffner & Prem, 1993), the "internalist trap" (Sharkey & Jackson, 1994), etc., but the essence of the problem remains the same.

symbol grounding. Ideally, embodied and situated agents would avoid using pre-programmed representations, and form their own behavioral dynamics based on their history of interaction with the environment. However, there still are not any completely autonomous agents, taking autonomy to mean the dependence of the behavior of an agent on its own experience (cf. Russel & Norvig, 1995). If we could successfully avoid coding our own human representations into an agent, and equip it with learning mechanisms to acquire its own representations instead, the meaning of these representations would not be "parasitic on the meanings in our heads" anymore, but rather be grounded in the experiences of the agent itself. In this way, research on autonomous agent learning can provide a basis for solving the symbol grounding problem in AI and for studying the emergence of symbols and communication in a real-world agent.

Embodied models and symbol grounding

The new paradigm in AI also led to new types of models. There have been two major currents in making use of robots as modeling tools. The first is biorobotics, which uses robots to model specific behavioral phenomena observed in animals (for a detailed account, see Webb, 2001). Most notably, the models developed in the field of biorobotics are empirical. Artificial neural networks modeling parts of the neuronal circuitry found in an animal are embodied in robot models that resemble certain aspects of this animal's morphology. These models are then tested under similar conditions with the animals, and similar evaluation methods are used. A well-known example is the model of the coordination of the six legs of the locust *Carausius morosus* and the emergence of different types of gait patterns, by Prof. Dr. Holk Cruse (University of Bielefeld) and his collaborators. This model takes the form of a six-legged mobile robot and the robot and the animals show similar gait patterns under similar conditions (surface structure walked on, walking speed). The experiences with the robot model led to a significant improvement of the previously simulated model: It was realized that the morphology of the agent and its interaction with the environment play a key role in leg coordination and greatly simplify the computations needed within the agent (see Dean et al., 1999). The second current of relating robots to modeling is to construct robots that illustrate how a behavior that alludes to important "cognitive" abilities of natural intelligent agents (e.g. to "learn" or to "categorize") can be implemented. In these models, the aim is not to reproduce data that has been collected in a controlled environment, but rather to get a better understanding of a cognitive ability in a situated and embodied context. Pfeifer & Scheier (1997) have programmed robots to categorize objects by coordinating their sensory input and motor output. By computing correlations between the sensor readings and motor outputs in one time step and the next, the robots were able to solve categorization tasks which would be too difficult to solve by using the sensory input alone. Sensory-motor coordination is thus not only an additional feature that could be implemented "on top" of an otherwise finished cognitive system. And it is not only the source of additional complications we could postpone until we have answered the "more cognitive" questions. Rather, categorization as one key task for every cognitive system can itself be seen as a process of sensory-motor coordination and taking an embodied standpoint does not complicate but rather simplifies the solution.

One eminent scientist that does a similar kind of "illustrating" research is Luc Steels, the director of the AI Lab of the Free University of Brussels. According to Steels & Vogt (1997), robots need to be equipped with at least basic communication abilities in order to move from agents that can solve "low level tasks" such as object avoidance and navigation (which are a usual topic of embodied AI) towards agents that could be said to exhibit "cognition". This communication must however again be autonomously developed by the agents themselves, in the spirit of the embodied AI bottom-up approach, and not designed or programmed in by a

human engineer. The communicated concepts and the means of communication must be grounded in the sensory-motor experiences of the robot (Steels, 1997). This way, robots could be used to study the origins of language and meaning in the self-organization and co-evolution of autonomous agents (Steels, 1996a). Steels and his collaborators carried out a number of experiments with robotic and software agents to study the emergence of reference and meaning, a lexicon, syntax and phonology. Here a brief overview of the first two aspects will be given, because they provide an embodied perspective on categorization and symbol grounding.

In the studies of perceptually grounded meaning creation, meaning is defined as "a conceptualisation or categorisation of reality which is relevant from the viewpoint of the agent" (Steels, 1996c, p.2). The hypothesis tested is that the origins of meaning can be found in construction and selection processes embedded in discrimination tasks. The agent attempts to discriminate one object or situation from others using just low-level sensory processing. Each individual agent is able to construct its own visual features by segmenting the input space of its different sensory channels. The attempts to perform a discrimination based on the current feature repertoire and the adaptation of the repertoire is called a *discrimination game*. In one such game, if the discrimination based on one or more distinctive feature sets fails, the agent will construct new feature detectors. Feature detectors are refined in the process and form discrimination trees. As a result, "the system arrives quite rapidly at a set of possible features for discriminating objects. Most interestingly, the system remains adaptive when new objects are added or when new sensory channels become available" (Steels, 1996c, p.14).

Lexicon formation is based on *language games*, which involve a dialogue between two agents that interact in a common situation. A word is a sequence of letters drawn from a finite shared alphabet, and an utterance is a set of words. Agents have the capability of creating new words (as random combinations of letters from an alphabet) and associating these new words with sets of features they are meant to denote (Steels, 1996b). One agent communicates one word to the other and the first agent tries to guess the set of features of the commonly perceived situation that the other agent might refer to with a word. It might find out that its initial guess was wrong when the same word is later used in another situation that does not contain the same features. Then another feature set would have to be assumed as the meaning of this word. On the other hand, a word can successfully be used by both agents if the assumed meaning (i.e. feature set) fits to all situations they both encounter. By using the common situation for feedback throughout a history of interactions, a set of common words and meanings emerge in both agents. As a result of the experiments, "it was shown that self-organization is an effective mechanism for achieving coherence and many properties of natural languages, in particular synonymy, ambiguity and multiple-word sentences, occur as a side effect of the proposed lexicon formation process" (Steels, 1996b).

The two agents start out with no repertoire of perceptual distinctions and no lexicon. After a number of discrimination and language games, they have acquired (1) a perceptual system for categorizing sensory experiences and identifying distinctive feature sets and (2) a lexicon that associates features or feature sets with words and vice-versa (Steels & Vogt, 1997). In a sense, the agents thus can be said to autonomously acquire grounded means of communication.

There is one important problem with Steels' approach to symbol grounding, however: the agents interact solely for the purpose of playing these pre-programmed games. Thus, the categories they acquire serve no other purpose for the agents than to utter and compare them with other such categories. In this sense, "[t]he system has no concept of [...] what to use the produced labels for, i.e. it is not embedded in any context that would allow/require it to make any meaningful use of these labels" (Ziemke, 1999, p.90). The case is very different with categories and symbols used by natural agents like us. Our categories develop primarily due to and to serve our needs, they are related to our purposes in successfully and autonomously

interacting with the world. A category is "ours" when it has a function in our cognitive machinery that goes far beyond just using it in communication. Therefore, the first step to symbol grounding should be learning categories that serve the agent in its interaction with the environment. And communication itself is used to transfer knowledge between agents, e.g. expressed as rules that tell an agent how things should be done., a task beyond just finding a common sets of words. The inventor of the idea of "language games" himself, Ludwig Wittgenstein, pointed out that we do not just learn words like "chair" and "table", but rather get involved in behaviors like sitting on chairs, putting things on a table etc., and learn to use words in such contexts in which they play a role. Luc Steels' approach resembles rather the simplistic model of Augustinus explicitly criticized by Wittgenstein, where words are labels attached to features of a commonly perceived scene that one agent points to to instruct the other (Wittgenstein, 1953, p.4).

Proposed approach for the thesis

Similarity-based categorization

To get to a more satisfactory approach towards studying the acquisition of symbols and communication in autonomous agents, I propose to start with a more realistic account of categorization. The categorization scheme I aim to implement will be a pattern-based one. Pattern-based approaches try to avoid generating abstractions and rules as far as possible and rather work by matching previously and currently encountered patterns according to their similarity (Hahn & Chater, 1998). Exemplar- (instance-) based models and case-based reasoning in machine learning are examples of such approaches (see e.g. Smith & Medin, 1981). To avoid the trap of encoding human categories in the design of the agent, the categorization process should work on raw sensor data and the interaction between agent and environment should be structured primarily according to the experience of the agent itself. The potential of exemplar-based systems to implement situated concepts has been pointed out by Barsalou (2000). The defining attribute of an exemplar-based categorization system is the similarity measure. As a most parsimonious approach, I will start with a pixel-based similarity measure. If the robot is unable to acquire the discriminations it needs to successfully interact with its environment using this measure, it will carefully be extended in the most parsimonious manner. The key issue in developing an appropriate similarity measure will be to avoid coding in features that are symbolized in language before the agent has actually engaged in a process of symbol acquisition. As seen above, Steels assumes that first representations of categories defined by abstract features have to be formed. Only after this representation has been formed, a name will be attached to the ready-made category. In this approach, as criticized above, the sole purpose of forming a category is to have a representation that a name can be attached to. In my approach, on the other hand, I will assume that the first and foremost purpose of categorization processes is to enable the agent to differentiate between situations in which it has to act differently. Whether or not there is any name for this difference in any language is secondary. The idea is to explain how mechanisms of categorization can be employed in the acquisition of symbols and symbolic communication even though they have not been designed for this purpose in the first place. But then it would be a mistake to assume symbol-like structures in the categorization mechanisms right away. This is why I will try to use a similarity measure that does not depend on symbolic feature codes, but works on a less abstract, e.g. pixel-based level.
The applicability of such a similarity measure in modeling categorization has been demonstrated in the EROSAL Master student project. The aim of the project was to build an empirical robot model of the performance of animals in categorization experiments carried

out by Christian W. Werner et al. (see our project report for details (Deiwiks et al., 2003)). The learning mechanism was an exemplar-based one which operated on raw camera input data encoded in the L*a*b color space. Our model managed to learn without the help of categories or input features built in by the designer, solely by pixel-based comparison.

Steels & Kaplan (2001) have opted for a very similar approach in their model of social learning of language, too. They also used a pixel-based similarity measure for category learning. While their model at first was based on the standard RGB color space used by computer cameras, they later switched to the L*a*b* space and report an improvement of the results. This color space was designed to match the perceived color dissimilarity found in psychophysical experiments with humans onto Euclidean distances of the positions of the respective color codes in an abstract 3D space. I also intend to use this color space in my similarity measure. Using a similar pixel-based similarity measure as the one used in the EROSAL project, Steels and Kaplan try to develop a robot model of learning words for single objects (names). Like Steels' earlier studies, the experiments are again constructed as "games". The game the robot is engaged in this time is called "categorization game". The robot (in this case a Sony AIBO™ robot dog) is shown an object by a human experimenter and is expected to classify it by uttering the same name that was uttered by a human instructor before. If the robot successfully classifies the object, it gets positive reinforcement. In case of error, the experimenter gives negative feedback. An associative memory stores relations between object views and words. Note that the object views are not matched on a single category which is then associated with a word, as Harnad's theory and Steels' earlier models would assume. Rather, an input pattern is directly associated with an output pattern (here: a name), without an intervening abstraction step. In the EROSAL project, we used the same idea and called it "categorization without categories". In our model, every visual pattern is individually matched to an action and the resulting feedback. The emergence of categorical behavior is due to the increasing number of exemplars and the effect of the reinforcement received as a result of the actions performed.

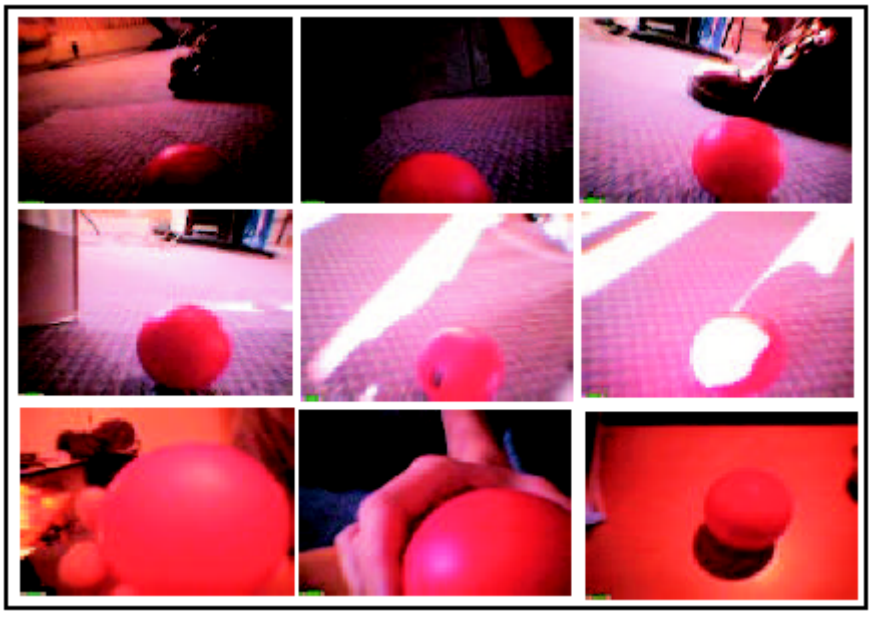

Figure 1: Different views of a red ball as captured by the robot's camera
(from Steels & Kaplan, 2001, p.12).

As can be seen in figure 1, the views which are associated with the same name ("ball") are actually quite different. Extracting their common, "invariant features" as Harnad proposed

would be a very difficult, if not impossible task. And without pre-coded detectors for the "relevant" features and other constraints implemented by a human engineer, we could also not be sure that the "correct" defining features are associated with the name "ball" (e.g., features of the floor might be included in the definition). But building in such pre-coded mechanisms would contradict the goal of behavior-based AI to construct robots that are autonomous, and it would also not lead to a realistic model of categorization and naming. I will therefore stick to an exemplar-based model that uses an "categorization without categories" mechanism.
The main drawback of Steels & Kaplan's approach is that they used pre-coded names for objects within their model. While the category is constructed by the robot itself by assembling object views, the robot has ready-made detectors for all object names that might possibly occur in the experiments, and simply attaches the respective pre-coded label to the current object view. In addition, while this model certainly is a more realistic account of symbol acquisition than Steels' earlier models, the authors do not relate to any existing studies on learning in animals or humans.

Research on stimulus equivalence as a possible empirical basis to study symbol acquisition

Once a successful model of categorization is framed, the next step will be to investigate the acquisition of symbols. A very recent approach in behavioral psychology, which still has largely gone unnoticed in cognitive science (embodied or otherwise) sees the roots of symbolization in a phenomenon dubbed *stimulus equivalence*. Pioneering work in psychology on this phenomenon has been done by Murray Sidman, who used symbolic match-to-sample procedures to teach conditional discriminations to youths with mental retardation (Sidman 1971; Sidman & Cresson 1973). In these experiments, the subjects first learned to select a particular comparison stimulus (e.g. the picture of a dog) upon hearing a word (e.g. "dog"), establishing a relation between these two types of stimuli (spoken words (A) and pictures (B), relation "AB"). Subjects were next taught to select printed words (e.g. DOG) upon hearing the corresponding spoken words (e.g. again "dog"), establishing an AC relation. Establishing such conditional relations is of course not new and has been demonstrated with many animal species. However, when the human subjects were tested further, new untrained behavioral relations emerged, which were seemingly not predicted by known laws of learning. These emergent relations were CB and BC relations, from printed words to pictures and from pictures to printed words, respectively. Sidman and colleagues argued that the subjects learned a new kind of relation between spoken words, written words and pictures that did not occur in earlier learning experiments with animal subjects. And they tried to show that the concept of equivalence as used in mathematics (defined by reflexivity, symmetry and transitivity) can be applied to further characterize this learned relation (Sidman et al., 1982; Sidman & Tailby, 1982). In their analysis, *reflexivity* is inferred from match-to-sample performance when subjects show generalized identity matching; *symmetry* is inferred when, having trained subjects to select stimulus B upon

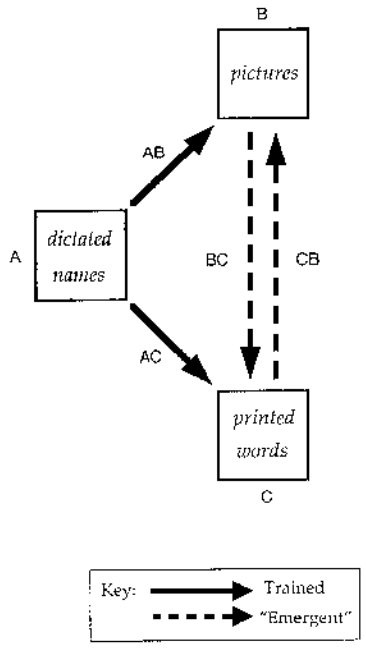

Figure 2: A schematic representation of Sidman's (1971) equivalence paradigm (from Horne & Lowe, 1996, p. 188)

presentation of stimulus A, the subjects proceed without further training to select A when presented with B; *transitivity* is inferred when, having established A-B and a second relation B-C, subjects proceed without further training to select C when presented with A.

The ability to form such stimulus equivalence classes purportedly gives rise to symbolic behavior and some other crucial linguistic capacities, and research has suggested that (non-human) animals lack this ability (Hayes, 1989). According to Sidman, stimulus equivalence is a linguistic prerequisite (Sidman 1986, p.226), and is a determining variable that accounts "both for what people say and for their reactions to what other people say […] [i]n particular the existence of equivalence relations can account for such utterances as 'meaning', 'symbol', 'referent', and 'rule-governed'" (Sidman, 1992, p.20). Similarly, Hayes and colleagues have maintained that "stimulus equivalence transforms nonlinguistic conditional discriminations into semantic processes" (Wulfert & Hayes, 1988, p.126) and is "a kind of working empirical model of semantic relations" (Hayes & Hayes, 1992, p.1387). It was proposed that equivalence classes define symbolic behavior and that the stimulus equivalence paradigm provides the basis for an experimental analysis of symbolic behavior (e.g., Catania, 1992, p.156; Dugdale & Lowe, 1990, p.115).

A number of theories have been proposed in order to explain stimulus equivalence: Sidman takes it to be a primitive function that has to be accepted as a given innate mechanism (Sidman, 1990). The proponents of the so-called *Relational Frame Theory (RFT)* view equivalence as just one out of several relations arising from a history of arbitrarily applicable relational responding (Hayes et al., 2001). On the other hand, Horne & Lowe doubt that new learning mechanisms have to be assumed. Rather, they propose that the ability of human beings not only to passively listen to words, but to actively produce (and then hear) their own utterances, in combination with known learning mechanisms, can account for the phenomena of *naming*, which they suppose to be the basic unit of verbal behavior (Horne & Lowe, 2000).

In my Ph.D. thesis work, I will try to use studies on (the mechanisms of) stimulus equivalence as an empirical basis for extending a pattern-based model of categorization, to arrive at an autonomous agent model of symbol acquisition and communication working in a situated context. As mentioned above, I will use an exemplar-based approach to model categorization, which is based on a very simple similarity measure and a "categorization without categories" approach. This way the categories that are formed will be a result of the experience of the agent, and no external knowledge about "relevant features" or "the right kind of similarity" for a specific task will be brought in by the designer. When the agent engages in communicative interaction with other agents, it will ground the symbols it and the other agents use in the stored experiences of interaction with their common environment it made before. Since both the acquisition of the symbols and the content of the sensory categories depend only on the experience of the agent, the symbol acquisition will be autonomous. As a small step towards language, I will try to develop a mechanism to utilize symbols acquired by an agent to instruct it in new tasks. Using instructions as an additional resource for controlling one's behavior (apart from one's own reflections, the changing situational context, etc.) is a key feature of human cognition. Verbal instruction is a much better metaphor for understanding the role of symbols and rules in controlling behavior than the notion of a plan in classical AI, which takes the form of a computer program that must be followed and that leaves no room for interpretation (see the notion of the "plan-as-communication" vs. the "plan-as-program" view analyzed in Agre & Chapman, 1990). The classical AI view on planning also informed the notion of "rule" that is dominant in current AI and cognitive science research. While the standard view on rules is to see them as lines of program code that will be carried out, I will take a view on rules that is based on our everyday understanding of the term. While we can learn patterns of behavior on our own, just like animals, we can also profit from formulated rules that tell us how something can be done. Rules, whether formulated by an agent itself or by another agent, provide a means to avoid a

long trial-and-error process to find a successful pattern of interaction with the environment. A rule "stands in" for many patterns that would otherwise have to be learned separately. But to do so, a rule must employ symbols, and these symbols must be grounded in patterns that the agent has acquired before. My model of how symbols as means of communication can be acquired by an agent that learns patterns of interaction with its environment will thus provide my answer to the problem how patterns (the kind of representation of experience we share with all animals) and rules (the kind of representation that only humans and maybe some higher animals use, after they acquired symbols) can be integrated.

It should be noted, however, that I will surely not be able to arrive at a model of the full linguistic complexity as exhibited by human beings. I will focus on the emergence of some first symbols and their use to instruct and improve behavior in simple situations. The study of more complex linguistic phenomena with autonomous agent models will have to be left for future work.

Previous and current own work

In WS 2002/2003 and SS 2003, I took part in the EROSAL (Empirical RObot Study on Animal Learning) Master student project. The aim of the project was to build a robot model of the performance of chickens in categorization experiments. With our robot model we were able to successfully replicate experiments done with real chickens in the C. und O. Vogt Institute of Brain Research at the Heinrich-Heine-University of Düsseldorf. As explained above, the theory behind our model was that no feature extraction and matching to separate representations of categories was required in order to account for the performance of the animal subjects. Our model used an exemplar-based decision mechanism that operated by matching current input patterns with ones encountered earlier. Pattern similarity was defined by a very simple, pixel-based measure. No visual feature extraction was assumed. The robot model was tested in the same controlled environment as the chickens, i.e. in a conditioning chamber, using the same means of evaluation. Instead of matching the input patterns to a unit representing a category, every input pattern was directly associated with a behavior. Our robot is an example of a successful model of the autonomous acquisition of categories that serve an agent in the interaction with its environment.

Currently I am working on my Master's thesis entitled "*Categorization in sensory-motor coordination: Experiments in functional category acquisition*". Building on the experiences gained in the project, I aim at applying a similar categorization mechanism in more complex experiments performed outside the conditioning chamber. The experiments are inspired by the ones proposed in Pfeifer & Scheier (1997) and are to be carried out with the Khepera™ robot available in our institute.

With the extended model of categorization, I hope to lay the foundation for my future Ph.D. work on the use of symbols denoting categories. I believe that a sound account of categorization, symbol grounding and reference has a lot to offer to both cognitive science and artificial intelligence. Proper bottom-up models of these phenomena can create updated and more ambitious frames for cognitive modeling. Embodied cognitive science offers the best tools and perspectives to model the mentioned phenomena, since it forces one to take account of the agent as a complete system interacting with the environment. In this framework, it is impossible to just assume the grounding of representations as given. However, the problem with most work using robots is that its relation to cognitive phenomena in natural agents is unclear. By building on empirical results on categorization, stimulus equivalence and naming, I will try to avoid this danger and develop an autonomous agent as a model in a cognitive science context. Equipping robots with means for symbol use and communication might later also prove useful in purely technical applications.

References

Agre, P. E. (1995). Computational research on interaction and agency. In: Agre, P. E., & Rosenschein, S. J. (eds.) (1995). *Computational Theories of Interaction and Agency*, Cambridge, MA; London, UK: MIT Press.

Agre, P. E., & Chapman, D. R. (1990). What are Plans for ?. In: Maes, Pattie (ed.) (1990). *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*. Cambridge, MA; London, UK: MIT Press, pp. 17-34.

Agre, P. E., & Rosenschein, S. J. (eds.) (1995). *Computational Theories of Interaction and Agency*. Cambridge, MA; London, UK: MIT Press.

Arkin, Ronald C. (1998). *Behavior-Based Robotics*. Cambridge, MA; London, UK: MIT Press.

Barsalou, L. W. (2000). Being there conceptually: simulating categories in preparation for situated action. In: Stein, N. L., Bauer, P. J., & Rabinowitz, M. (2002). *Representation, memory, and development: Essays in honor of Jean Mandler*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brooks, R. A. (1991a). Intelligence without representation. *Artificial Intelligence* 47, 139-160.

Brooks, R. A. (1991b). New Approaches to Robotics. *Science* 253, 1227–1232

Catania, A.C. (1992). *Learning* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Chalmers, D.J. (1992). Subsymbolic computation and the Chinese room. In: Dinsmore, J. (ed.) (1990). *The symbolic and connectionist paradigms: closing the gap.* Hillsdale, NJ: Lawrence Erlbaum.

Dean, J., Kindermann, T., Schmitz, J., Schumm, M., & Cruse, H. (1999). Control of walking in the stick insect: from behavior and physiology to modeling. *Autonomous Robots* 7, 271-288.

Deiwiks, C., Gergou, A., Läer, L., Land, R., Lange, S., Plate, J. & Türkmen, U. (2003): *EROSAL: Empirical Robot Study on Animal Learning*. Project Report, Institute of Cognitive Science, University of Osnabrück.

Dreyfus, H. L. (1972). *What computers can't do*. Cambridge, MA; London, UK: MIT Press.

Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Hahn, U., & Chater, N. (1998). Similarity and Rules: Distinct? Exhaustive? Empirically Distinguishable?. *Cognition* 65, 197-230.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica* D 42, 335-346.

Harnad, S. (1993). Symbol Grounding is an Empirical Problem: Neural Nets are Just a Candidate Component. In: Kintsch, W. (ed.). *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, pp. 794-799.

Haugeland, J. (1981). Semantic engines: an introduction to mind design. In: Haugeland, J. (ed.). *Mind Design*. Cambridge, MA; London, UK: MIT Press.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (eds.) (2001). *Relational Frame Theory: A Post-Skinnerian Account of Human Language and Cognition*. New York: Kluwer Academic / Plenum Publishers.

Hayes, S. C. (1989). Nonhumans have not yet shown stimulus equivalence. *Journal of the experimental analysis of behavior* 51*,* 385-392.

Hayes, S.C., & Hayes, L.J. (1992). Verbal relations, cognition, and the evolution of behavior analysis. *American Psychologist* 47, 1383-1395.

Horne, P.J., & Lowe, C.F. (1996). On the origins of naming and other symbolic behavior. *Journal of the experimental analysis of behavior* 65, 185-241.

Horne, P. J. & Lowe, C. F. (2000). Putting the naming account to the test: Preview of an experimental programme. In: Leslie, J. C. & Blackman, D. (eds.). *Experimental and Applied Analysis of Human Behavior*. Reno, NV: Context Press, pp.127-148.

Maes, P. (ed.) (1994). *Designing autonomous agents: theory and practice from biology to engineering and back*. Cambridge, MA; London, UK: MIT Press.

Newell, A., & Simon, H. (1976). Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM* 19(3),113-126.

Pfeifer, R., & Scheier, C. (1997). Sensory-motor coordination: the metaphor and beyond. *Robotics and Autonomous Systems* 20, 157-178.

Pfeifer, R., & Scheier, C. (1999). *Understanding Intelligence*. Cambridge, MA; London, UK: MIT Press.

Russel, S. J., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice-Hall.

Scheier, C., & Lambrinos, D. (1996). Categorization in a real-world agent using haptic exploration and active perception. In: Maes, P., Mataric, M., Pollack, J., Meyer, J.-A.,& Wilson, S. (eds.): *From Animals to Animats 4, Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. Cambridge, MA; London, UK: MIT Press.

Searle, J.R. (1980). Minds, brains and programs. *The Behavioral and Brain Sciences* 3, 417-457.

Sidman, M. (1971). Reading and auditory-visual equivalences. *Journal of speech and hearing research* 14, 5-13.

Sidman, M. (1986). Functional analysis of emergent stimulus classes. In: Thompson, T., & Zeiler, M.D. (eds.) (1986). *Analysis and integration of behavioral units*. Hillsdale, NJ: Erlbaum, pp. 213-245.

Sidman, M. (1990). Equivalence relations: Where do they come from? In: Blackman, D.E., & Lejeune, H. (eds.) (1990). *Behavioral analysis in theory and practice: Contributions and controversies*, Hillsdale, NJ: Erlbaum, pp. 93-114.

Sidman, M. (1992). Equivalence relations: some basic considerations. In: Hayes, S.C., & Hayes, L.J. (eds.) (1992). *Understanding verbal relations*. Reno, NV: Context Press, pp. 15-27.

Sidman, M., & Cresson, O., Jr. (1973). Reading and crossmodal transfer of stimulus equivalences in severe retardation. *American journal of mental deficiency* 77, 515-523.

Sidman, M., Rauzin, R., Lazar, R., Cunningham, S., Tailby, W., & Carrigan, P. (1982). A search for symmetry in the conditional discriminations of rhesus monkeys, baboons and children. *Journal of the experimental analysis of behavior* 3, 23-44.

Sidman, M., & Tailby, W. (1982). Conditional discrimination vs. matching to sample: an expansion of the testing paradigm. *Journal of the experimental analysis of behavior* 37, 5-22.

Smith, E., & Medin, D. (1981). The exemplar view. In: Margolis, E., & Laurence, S. (eds.) (1981). *Concepts: Core Readings*. Cambridge, MA; London, UK: MIT Press, pp. 207-221.

Steels, L. (1996a). Synthesising the origins of language and meaning using co-evolution and self-organisation. In: Hurford, J. (ed.) (1996). *Evolution of Human Language*. Edinburgh: Edinburgh University Press.

Steels, L. (1996b). Emergent Adaptive Lexicons. In: Maes, P., Mataric, M.J., Meyer, J.-A., Pollack, J., & Wilson, S.W. (eds.). *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. Cambridge, MA; London, UK: MIT Press.

Steels, L. (1996c). Perceptually grounded meaning creation. In: Tokoro, M. (ed.). *Proceedings of the International Conference on Multi-Agent Systems*. Cambridge, MA; London, UK: MIT Press.

Steels, L. (1997). Constructing and Sharing Perceptual Distinctions. In: van Someren, M., & Widmer, G. (eds.). *Proceedings of the European Conference on Machine Learning, ECML'97*. Berlin et al.: Springer, pp. 4-13.

Steels, L. & Vogt, P. (1997). Grounding adaptive language games in robotic agents. In: Husbands, C., & Harvey, I. (eds.). *Proceedings of the Fourth European Conference on Artificial Life (ECAL' 97)*. Cambridge, MA; London, UK: MIT Press.

Steels, L., & Kaplan, F. (2001). AIBO's first words: The social learning of language and meaning. *Evolution of Communication* 4(1).

Webb, Barbara (2001). Can robots make good models of biological behaviour?. *The Behavioral and Brain Sciences* 24 (6), 1033-1050.

Werner, C.W.; Gravemeier, B. & Rehkämper, G. (no year). A mathematical model of configural processing of integral and separable compounds by chickens, unpublished manuscript.

Wulfert, E., & Hayes, S.C. (1988). Transfer of a conditional ordering response through conditional equivalence classes. *Journal of the experimental analysis of behavior* 50, 125-141.

Wittgenstein, L. (1953). Philosophical Investigations. Oxford: Basil Blackwell. [translated from German by G.E.M. Anscombe]

Zentall, T.R., Galizio, M., & Critchfield, T.S. (2002). Categorization, concept learning, and behavior analysis: an introduction. *Journal of the experimental analysis of behavior* 78, 237-248.

Ziemke, T. (1999). Rethinking grounding. In: Riegler, A., Peschl, M., & von Stein, A. (eds.) (1999). *Understanding Representation in the Cognitive Sciences*. New York: Kluwer Academic / Plenum Publishers.