AIBO's first words. The social learning of language and meaning

Luc Steels[1,2] and Frederic Kaplan[1]
(1) SONY Computer Science Laboratory - Paris
(2) VUB Artificial Intelligence Laboratory - Brussels
steels@arti.vub.ac.be

Abstract

This paper explores the hypothesis that language communication in its very first stage is bootstrapped in a social learning process under the strong influence of culture. A concrete framework for social learning has been developed based on the notion of a language game. Autonomous robots have been programmed to behave according to this framework. We show experiments that demonstrate why there has to be a causal role of language on category acquisition; partly by showing that it leads effectively to the bootstrapping of communication and partly by showing that other forms of learning do not generate categories usable in communication or make information assumptions which cannot be satisfied.

1 Introduction

How children acquire the meaning of words is a fascinating, still unresolved problem but a key towards understanding how human-level language communication could ever have developed. This paper addresses two basic puzzles concerning this process:

1. *How does the bootstrapping into communication take place?* What are necessary prerequisites to enable the magic moment when the child learns 'how to mean' [Halliday, 1987]?

2. *How is meaning acquired?* The meanings used by a speaker cannot directly be observed by the listener, so how can a listener who does not know the meaning of words ever learn them?

There are two main lines of thinking on these questions: individualistic learning and social learning. In the case of individualistic learning, the child is assumed to receive as input a large number of example cases where speech is paired with specific situations. She is either already mastering the necessary concepts or able to extract through an inductive learning process what is essential and recurrent of these situations, in other words learn the appropriate categories underlying language, and then associate these categories with words. This is known as cross-situational learning [Fischer et al., 1994]. Others have proposed a form of contrastive learning on the same sort of data, driven by the hypothesis that different words have different meanings [Clark, 1987]. This type of individualistic learning assumes a rather passive role of the language learner and little feedback given by the speaker. It assumes no causal influence of language on concept formation. We call it the labelling theory because the language learner is assumed to associate labels with existing categories. The labelling theory is remarkably widespread among researchers studying the acquisition of communication and recently various attempts have been made to model it with neural networks or symbolic learning algorithms [Broeder and Murre, 2000]. It is known that induction by itself is a weak learning method, in the sense that it does not give identical results on the same data and may yield irrelevant clustering compared to human categories. This will indeed be demonstrated to be the case later in this paper. To counter this argument it is usually proposed that innate constraints help the learner zoom in on the important aspects of the environment [Bloom, 2000], [Smith, 2001] , [Markman, 1994].

In the case of social learning, interaction with other human beings is considered crucial ([Tomasello, 2000], [Steels, 2001c]). Learning is not only grounded in reality through a sensori-motor apparatus but also socially grounded through interactions with others. The learning event involves an interaction between at least two individuals in a shared environment. They will further be called the learner and the mediator. The mediator could be a parent and the learner a child, but children (or adults) can and do teach each other just as well. Given the crucial role of the mediator, we call social learning also mediated learning. The goal of the interaction is not really teaching, which is why we use the term mediator as opposed teacher. The goal is rather something practical in the world, for example,

to identify an object or an action. The mediator helps to achieve the goal and is often the one who wants to see the goal achieved.

The mediator has various roles: She sets constraints on the situation to make it more manageable (scaffolding), gives encouragement on the way, provides feedback, and acts upon the consequences of the learner's actions. The feedback is not directly about language and certainly not about the concepts underlying language. The latter are never visible. The learner cannot inspect telepathically the internal states of the speaker and the mediator cannot know which concepts are already known by the learner. Instead feedback is pragmatic, that means in terms of whether the goal has been realised or not. Consider a situation where the mediator says: "Give me that pen", and the learner picks up a piece of paper instead of the pen. The mediator might say: "No, not the paper, the pen", and point to the pen. This is an example of pragmatic feedback. It is not only relevant to succeed subsequently in the task but supplies the learner with information relevant for acquiring new knowledge. The learner can grasp the referent from the context and situation, hypothesise a classification of the referent, and store an association between the classification and the word for future use. While doing all this, the learner actively tries to guess the intentions of the mediator. The intentions are of two sorts. The learner must guess what the goal is that the mediator wants to see realised (like 'pick up the pen on the table') and the learner must guess the way that the mediator has construed the world [Langacker, 1991]. Typically the learner uses herself as a model of how the mediator would make a decision and adapts this model when a discrepancy arises.

Social learning enables active learning. The learner can initiate a kind of experiment to test knowledge that is uncertain or to fill in missing wholes. The mediator is available to give direct concrete feedback for the specific experiment done by the learner. This obviously speeds up the learning, compared to a passive learning situation where the learner simply has to wait until examples arise that would push the learning forward.

The debate between individualistic versus social learning is related to the equally hotly debated question whether there is a causal role for language in category acquisition or not. From the viewpoint of the labelling theory the acquisition of concepts occurs independently off and prior to language acquisition [Harnad, 1990]. So there is no causal role of language. Conceptualisation and verbalisation are viewed as operating in independent modules which have no influence on each other [Fodor, 1983]. The acquisition of language is seen as a problem of learning labels for already existing concepts. Concerning then

the issue how the concepts themselves are acquired, two opposing schools are found: nativism and empiricism. Nativists like Fodor [Fodor, 1999] claim that concepts, particularly basic perceptually grounded concepts, are innate and so there is no learning process necessary. They base their arguments on the poverty of the stimulus [Chomsky, 1975], the fundamental weakness of inductive learning [Popper, 1968], and the lack of clear categorial or linguistic feedback. Empiricists claim that concepts *are* learned, for example by statistical learning methods implemented as neural networks [Ellman, 1993]. Thus a large number of situations in which a red object appears are seen by the learner, and clustered into 'natural categories'. These natural categories then form the basis for learning word-meaning.

The alternative line of thinking, which is often adopted by proponents of social learning, claims that there *is* a causal role for culture in concept acquisition and this role is particularly (but not exclusively) played through language. This has been argued both by linguists and philosophers. In linguistics, the position is known as the Sapir-Whorf thesis. It is based on evidence that different languages in the world not only use different word forms and syntactic constructions but that the conceptualisations underlying language are profoundly different as well [Talmy, 2000]. Language acquisition therefore goes hand in hand with concept acquisition [Bowerman, 2001]. Language-specific conceptualisations change over time in a cultural evolution process which in turn causes grammatical evolution that may again induce conceptual change [Heine, 1997]. Note that a causal influence of language acquisition on concept formation does not imply that all concepts undergo this influence or that there are no concepts prior to the beginning of language acquisition. In fact, there are probably millions of concepts used in sensori-motor control, social interaction, emotion, etc. which are never lexicalised. The main point of the paper is that for those concepts underlying natural language communication this causal influence not only exists but is necessary.

Ludwig Wittgenstein [Wittgenstein, 1953] is the best known philosophical proponent of a causal influence of language on meaning. His position is in a sense even more radical than the Sapir-Whorf thesis. He has argued that meanings are an integrated part of the situated context of use. Thus the word "ball" not only includes a particular conceptualisation of reality in order to refer to a certain type of object but is also a move in a language game, indicating that the speaker wants to get a particular action be carried out. Moreover the meaning of "ball" is not abstract at all, i.e. something of the sort 'spherical shaped physical object of a uniform colour', but is very context-dependent, particularly in the first stages. This point has also been made by Quine [Quine, 1960] who has argued that basic

notions such as object-hood only gradually arise. Children do not start with the pre-given clean abstract categories that adults appear to employ.

1.1 Robots as models

This paper examines the hypothesis that communication is bootstrapped in a social learning process under the strong influence of language and that initially meanings are situated and context-dependent. It argues that individualistic observational learning and the labeling theory cannot explain the very difficult first steps into language-like communication. Many concrete and even formal models exist for individualistic learning [Broeder and Murre, 2000] but similar models for social learning are lacking. In the absence of such models it is difficult to seriously compare the different positions in the debate without sliding into rhethoric. The first goal of our work has therefore been to develop a concrete model for social learning and compare its behavior to individualistic learning. The second goal is to show that cultural influence and context-dependent meanings are indeed the most plausible and effective way for an individual to bootstrap herself into a language culture.

The method we use to validate our claims is perhaps unusual from a social science perspective. First of all we completely operationalise and formalise the steps necessary in language acquisition. Validation of this formal model by hand is however completely excluded given the enormous complexity of the cognitive processing required for grounded language, even for handling single words. So it is at least necessary to do computer simulations. Here we go one step further and do experiments with autonomous mobile robots, in line of similar work reported in ([Steels and Vogt, 1997], [Billard et al., 1998], [Vogt, 2000]).

Robotic experiments are motivated as follows:

1. They force us to make every claim or hypothesis about assumed internal structures and processes very concrete and so it is clear how the theoretical assumptions have been operationalised.

2. We can use real-world situations, i.e. physical objects, human interactions, etc. to get realistic presuppositions and realistic sources of input. This is particularly important when studying social learning, which relies heavily on the intervention of the mediator, grounded in reality.

3. We can extract data about internal states of the learning process, which is not possible with human beings. Internal states of children going through a developmental or learning process cannot be observed at all.

4. We can easily examine alternative hypotheses. For example, we can compare what an individualistic inductive learning process would achieve with the same data as a social learning process.

But there are obviously also important limits to this methodology:

1. We do not pretend at all that robotic experiments model in any realistic way children nor the environments in which they typically operate. Our goal here is to examine specific assumptions about the emergence of communication by building artificial systems, so realism is not at issue [Steels, 2001b].

2. It is an extraordinary challenge to build and maintain physical robots of the required complexity. For practical reasons (limitations of camera resolution, memory and processing power available on board) we cannot always use the best known algorithm available today. This puts limits on what can be technically achieved today and so experiments need to be designed within these limits.

By using real world autonomous robots, our experiments differ from other computational experiments in word learning (such as [Siskind, 1995]) in which situation-word pairs are prepared in advance by the human experimenter, and even more from more traditional connectionist word learning experiments where meanings are explicitly given by a human [Regier, 1996]. Here we approach much more closely the conditions of a one year old child who is moving around freely with no preconception of what might be the meaning of a word. In fact, we make the situation even more difficult than that of a child which presumably has already acquired many more concepts that could potentially be used or adapted for language communication.

For the experiments reported in this paper, we use an enhanced version of the Sony AIBO^{TM} robot (see figure 1) further called the robot. The robot is fully autonomous and mobile with more than a thousand behaviors, coordinated through a complex behavior-based motivational system [Fujita and Kitano, 1998]. It features 4-legged locomotion, a camera for visual input, two microphones, a wide variety of body sensors, as well as on board batteries and the necessary computing

power. We have chosen this platform because the AIBO is one of the most complex autonomous robots currently in existence but nevertheless reliable enough for systematic experiments due to the industrial standards with which it has been designed and built. Moreover the AIBO is designed to entice interaction with humans, which is what we need for experiments in social human-robot interaction. It comes with a very wide range of capabilities which are necessary to establish the conditions for social interaction, such as the ability to look at the object as a way to draw attention of the speaker to the object.

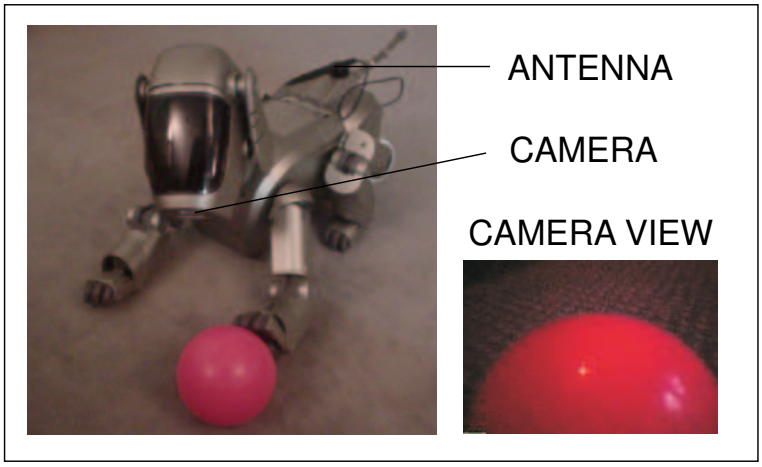

Figure 1: Our robot is an enhanced version of commercially available AIBOs. It is linked to an additional computer through a radio connection

The experiments discussed further in this paper work on an enhanced version of the AIBO because there is not enough computing power on board to do them. We have decided to keep the original autonomous behavior of the robot and build additional functionality on top of it. Our system thus acts as a cognitive layer which interferes with the current autonomous behavior, without controlling it completely. A second computer implements speech recognition facilities which enables interactions using spoken words. In order to avoid recognition problems linked with noise, the mediator uses an external microphone to interact with the robot. The computer also implements a protocol for sending and receiving data between the computer and the robot through a radio link. The mediator must take into account the global ”mood” of the robot as generated by the autonomous motivational system. For example, it is possible that a session becomes very ineffective

because the robot is in a ”lethargical” mood.

Even though the robot is extraordinarily complex, it does not necessarily use state of the art algorithms for every aspect because that would require vastly more computational resources. For example, 3d depth recognition is now feasible [Beymer and Konolige, 1999] but would require more hardware on the robot (such as two cameras instead of one) and much more processing. These technological constraints limit necessarily the potential levels of intelligence. We feel however that in the present context these weaknesses are not a drawback, because we want to focus on the very first words. This should not require a complex structural analysis of visual scenes, nor very sophisticated world models, nor complex grammatical constructions or intricate dialogs.

1.2 Overview

We have done three types of experiments, all focusing on naming three objects in its environment: a red ball, a yellow puppet called ”Smiley”, and a small AIBO imitation called Poo-chi^{TM}. In the first experiment the robot has been programmed to be capable of some form of social learning. The learning takes place through intense interaction with a human mediator (figure 2a).

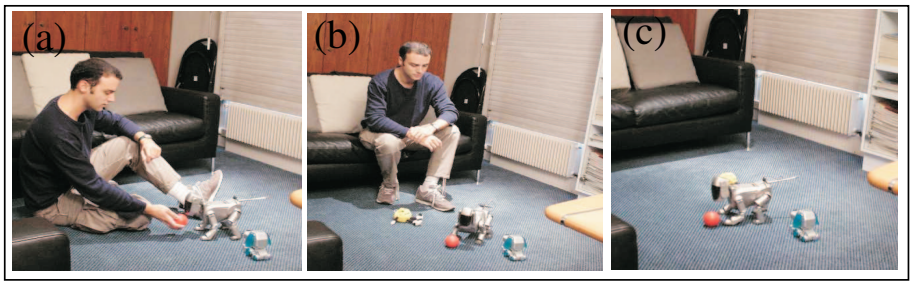

Figure 2: Three types of experiments with different degrees of social interaction: strong interaction (a), observational learning with supervision (b), unsupervised learning (c)

In the second experiment the role of the mediator is strongly reduced to create a learning situation comparable to supervised observational learning (figure 2 b). The robot gets examples pairing a word with a view of one of the objects. The human mediator supplies the words but there is not the intense interaction characteristic of social learning.

In the third experiment we examine unsupervised observational learning (figure 2 c). The robot gets a series of images and uses unsupervised clustering methods to detect the natural categories present in the data. The question here is whether these natural categories have any relation to the categories that underly the words normally used in English for referring to the objects.

The rest of the paper contains two main parts. The first part is about the functionalities that need to be in place for establishing forms of interaction in which early language could sprout through social learning. We will argue that the notion of a language game [Steels, 2001a] is an appropriate framework for setting up such interactions and introduce the example of a classification game. The second part of the paper focuses on the issue of meaning, and particularly on the debate between observational vs. social learning. We conclude that there are strong reasons to insist on social learning to explain how verbal communication might bootstrap.

2 Language Games

In previous work we have found that the notion of a language game is a very effective way to frame social and cultural learning [Steels, 2001a]. A game is a routinised sequence of interactions between two agents involving a shared situation in the world. The players have different roles. There are typically various objects involved and participants need to maintain the relevant representations during the game, e.g. what has been mentioned or implied earlier. The possible steps in a game are called moves. Each move is appropriate in circumstances determined by motivations and long term objectives and the opportunities in the concrete situation, just like a move in a game of chess. Games are much more encompassing than behaviors in the sense of behavior-based robots [Steels and Brooks, 1994]. They may run for several minutes and invoke many behaviors and cognitive activities on the way. They may be interrupted to be resumed later.

Here is an example of a game played with a child while showing pictures of animals:

```
Father: What does the cow say? [points to cow] Mooooh.
Child: [just observes]
Father: What does the dog say? [points to dog] Waf.
Child: [observes]
```

```
Father: What does the cow say?
    [points to cow again and then waits ... ]
Child: Mooh
Father: Yeah!
```

The learner learns to reproduce and recognise the sounds of the various animals and to associate a certain sound with a particular image and a particular word. The example is very typical, in the sense that (1) it involves many sensory modalities and abilities (sound, image, language), (2) it contains a routinised set of interactions which is well entrenched after a while so that it is clear what is expected, (3) the learner plays along and guesses what the mediator wants and the mediator sets up the context, constrains the difficulties, and gives feedback on success or failure. (4) The meaning of words like 'cow' and 'dog' or 'moooh' and 'waf' involves both a conceptual aspect (classification of the animals and imitations of the sound they make) and a game aspect (moves at the right moment). Every parent plays thousands of such games with their children and, equally important, after a while children play such games among themselves, particularly symbolic games.

Games like the one above are typical for children around the age between 2 and 3. This example focuses exclusively on language learning. Normally games try to achieve a specific cooperative goal through communication where language plays an auxiliary role, such as:

- Get the listener to perform a physical action, for example move an object.
- Draw attention of the listener to an element in the context, for example, an object that she wants to see moved.
- Restrict the context, which is helpful for drawing attention to an element in it.
- Transmit information about one's internal state, for example to signal the degree of willingness to cooperate.
- Transmit information about the state of the world, for example as relevant for future action.

For all these games there must be a number of prerequisites for social interaction like the following:

1. Become aware that there is a person in the environment, by recognising that there is a human voice or a human bodily shape.

2. Recognise the person by face recognition or speaker identification.

3. Try to figure out what object the speaker is focusing attention to, independently of language, by gaze following and eye tracking.

4. Use the present situation to restrict the context, predict possible actions, and predict possible goals of the speaker.

5. Give feedback at all times on which object you are focusing, for example by touching the object or looking at it intently.

6. Indicate that you are attending to the speaker, by looking up to the speaker.

These various activities are often associated with having a 'theory of mind' [Baron-Cohen, 1997]. It is clear that these prerequisites as well as the ones specifically required for the language aspects of a game require many cognitive capabilities: vision, gesturing, pattern recognition, speech analysis and synthesis, conceptualisation, verbalisation, interpretation, behavioral recognition, action, etc. This paper will not go into any technical detail how these capabilities have been achieved for the robot (in most cases by adopting state of the art AI techniques) nor how they are integrated. It suffices to know that we have a large library of components and a scripting language COALA that handles the integration and scheduling in real-time of behaviors to implement interactive dialogs. The agent's scripts for playing a game should not only invoke the necessary components to achieve success in the game but also trigger the learning algorithms that can help to fill in missing concepts or learn the meaning of new stretches of natural language. We do not pretend that any of these components achieves human level performance, far from it. But they are enough to do experiments addressing the issues raised in this paper and observers are usually stunned about the level of performance already achieved.

2.1 The Classification Game

In earlier work, we have been experimenting with various kinds of language games, most notably a guessing game [Steels and Kaplan, 1998], in which the listener must guess an object in a particular context through a verbal description

that expresses a property of the object which is not true for any of the other objects in the context. The robots in these experiments were static: pan-tilt units mounted on a fixed tripod. These experiments have demonstrated beyond doubt that the language game approach is effective, given appropriate scripts, not only for sustaining a dialog but also for acquiring the necessary concepts and words. The emergence and evolution of a lexicon in a population of agents has been experimentally shown to arise [Steels et al., 2002].

In the present paper we will use a classification game. The classification game is similar to the guessing game, except that there is only a single object to be classified. Because the robot's camera does not have a very wide view angle, only one object is generally in view, and so the classification game is more natural for this robot than the guessing game. It is similar to the interactions studied in [Roy, 1999] and [Fujita et al., 2001] and so it makes comparison with other work easier. The playful objects to be classified include a ball, a small puppet that looks like a Smiley, a small AIBO imitation marketed as poo-chi. English words like "ball", "smiley", or "poo-chi" are used by the human mediator in interactions with the robot. The main goal for the robot is to acquire the proper use of the word in relation to visual images.

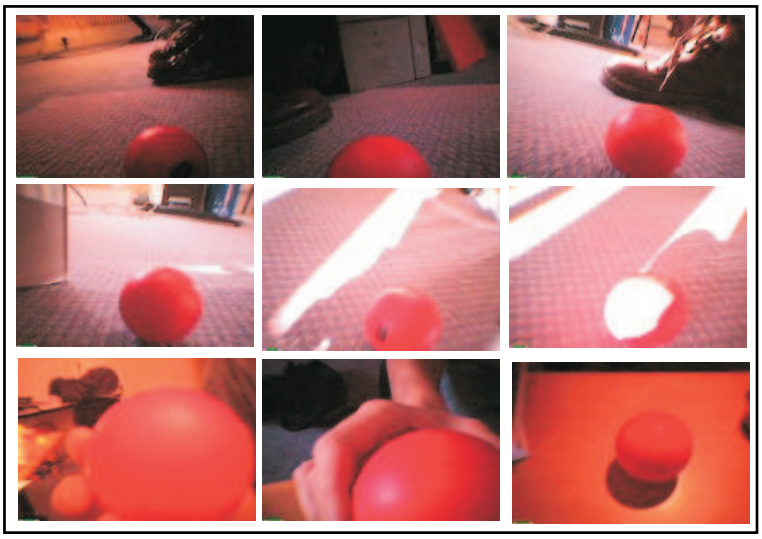

Figure 3: Different views of a red ball as captured by the robot's camera.

Figure 3 gives an idea of the difficulties involved. All these images have been

captured with the robot's camera. Different ambient light conditions may change completely the colour reflection of an object. An object is almost never seen in its entirety. It can have a complex structure so that different sides are totally different (for example back and front of poo-chi). Consequently segmentation and subsequent classification is extremely difficult. For example, the red ball may sometimes have a light patch which looks like a second object or fuse so much with the background that it is hardly recognisable. We feel that it is of extreme importance to start from realistic images taken during a real world interaction with the robot and a human. By taking artificial images (for example pre-segmented images under identical light conditions) many of the real world problems that must be solved bootstrapping communication would disappear, diminishing the strength of the conclusions that can be drawn.

2.2 Script

The robot has a script, implemented as a collection of loosely connected schemas, for playing the classification game. Here is a typical dialog based on this script, starting when the robot sits down.

```
1. Human: Stand.
2. Human: Stand up.
```

The robot has already acquired names of actions (as explained in [Kaplan et al., 2001]). It remains under the influence of its autonomous behavior controller. Forcing the robot to stand up is a way to make it concentrate on the language game. Because speech signals have been heard, the robot knows that there is someone in the environment talking to it. The human now shows the ball to the robot (figure 4 a).

```
3. Human: Look
```

The word "look" helps to focus attention and signals the begining of a language game. The robot now concentrates on the ball, starts tracking it, and signals focus by looking at the ball (figure 4 a) and trying to touch it (figure 4 b). It further signals attention by looking first at the speaker (figure 4 c) and then back at the ball (figure 4 d). In fact, these are all emergent behaviors of the object tracker. The other autonomous behaviors interact with the schemas steering the language game.

```
4. Human: ball
```

The robot does not know a word yet for this object, so a learning activity starts. The robot asks first for feedback of the word to make sure that the word has been heard correctly.

```
5. Aibo: Ball?
6. Human: Yes
```

Ball is the correct word and it is associated with a view of the object seen.

Note that several things might have gone wrong in this episode and correction from the human mediator would have been required. For example, the wrong word might have been heard due to problems with speech recognition, the robot might not have been paying attention to the ball but, because of its autonomous behaviors, might have started to look elsewhere, etc. By maintaining a tightly coupled interaction, the mediator can help the learner and this is the essence of social learning: constraining context, scaffolding (the human says ”ball” not ”this is the ball” which would be much more difficult), and pragmatic feedback on the way.

The dialog scripts implemented on the robot are sufficiently flexible to allow many variants of the classification game, and specifically enable the learner to test out knowledge. Here are some other example dialogs:

```
1. Human: What is it?
2. Aibo: Ball
3. Human: Good.
```

```
1. Human: What is it?
2. Aibo: Smiley
3. Human: No; listen; Ball.
4. Aibo: Ball?
5. Human: Yes.
```

```
1. Human: Is it ..  Smiley?
2. Aibo: No; ball
3. Human: Good.
```

In each case, the robot categorises and names the object and gets feedback whether the naming was correct from a pragmatic viewpoint.

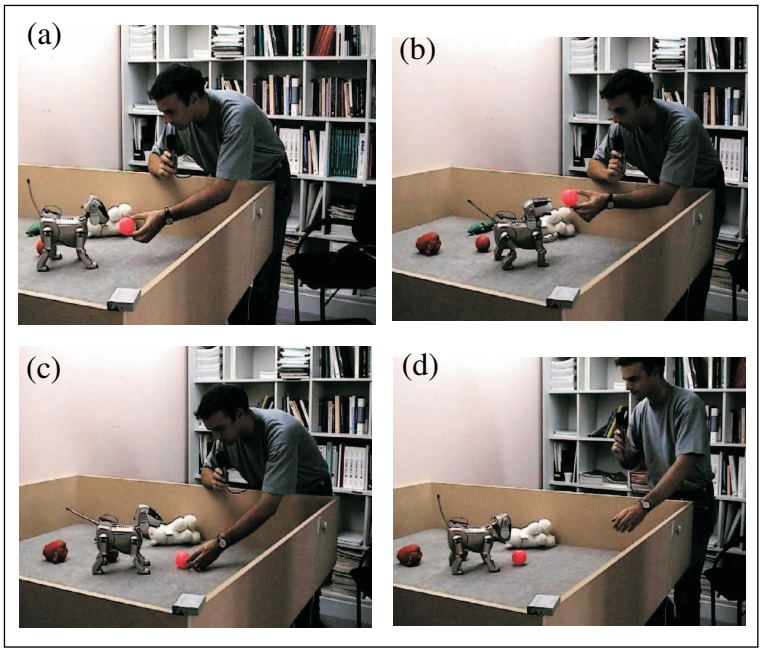

Figure 4: Different steps during an language game

We have implemented all the necessary components to have the robot play classification games of the sort shown in these examples and experimented for several months in human-robot interactions. These experiments have shown that the framework of language games is effective to enable the learning of the 'first words' and the classificatory concepts that go with it. Before discussing the results of these experiments, the next sections provide some more detail on the most important parts of the classification script.

2.3 Classification

The classification game relies on the ability to classify objects. There are many possible ways to classify objects and many techniques are known to acquire each type of classification. We have tried to use as much as possible well known, state of the art methods. The first important decision was not to segment objects. Object segmentation is notoriously difficult and generally believed to be impossible, unless there is a clear template of the object available. Edge detection, 3-d segmentation, colour segmentation, segmentation based on change from one image to the next, etc. all yield possible segments but none is failproof. So the learner is confronted with a chicken and egg problem. There is no way to know what counts as an object but without this knowledge it is virtually impossible to perform segmentation. By not relying on prior segmentation we resolve this paradox. It implies however that the initial concepts for objects are always highly context-sensitive. This situated, context-sensitive nature of object knowledge is in line with Wittgenstein's point of view and has also been argued on empirical grounds [Clancey, 1997].

The second decision was to use an instance-based method of classification ([Mel, 1997], [Witten and Eibe, 2000]). Many different 'views' are stored of an object in context and classification takes place by a nearest neighbor algorithm. Views are not stored in terms of full RGB bitmaps, which would require too much storage and would require very computation-intensive methods for comparison. Instead the image is first normalised in order to remove too much dependency on brightness [Finlayson et al., 1998], R(ed)G(reen)B(lue) data are normalised with respect to R+G+B and only two dimensions (G and B) are kept since R + G + B = 1.0, so the third dimension is no longer informative. Then a 16 x 16 2D color histogram is constructed of the normalised image and every image is represented by 256 values. These are the dimensions of the conceptual space used to represent objects in memory. Figure 5 shows an object and its corresponding histogram.

Note that we cannnot really say that the memory ”represents” objects because the robot has no notion yet of what an object is. The color histogram reflects both the perception of the object and its background.

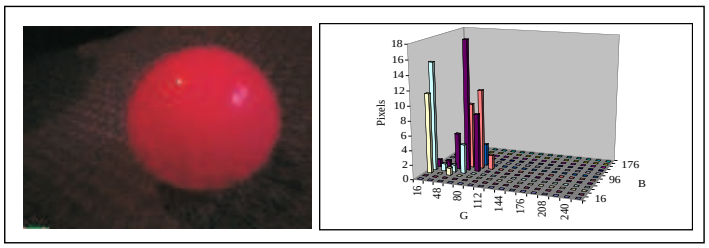

Figure 5: Example of an image and the associated histogram

To compare the perceived histogram with the ones previously stored in memory, we use a χ^2-divergence measure defined in the following way:

$$\chi^2(A,B) = \sum_i \frac{(a_i - b_i)^2}{a_i + b_i} \tag{1}$$

where a_i and b_i are the value of two histograms A and B indexed by i. The view with the shortest distance in pair-wise comparison to the input image is considered to be the 'winning' view. Several other methods for matching histograms have been compared with this measure by Schiele and Crowley [Schiele and Crowley, 1996] and it appeared to be the best one. It is used also by Roy [Roy, 1999].

Instance-based learning was used for two reasons: (1) It supports incremental learning. There is no strict separation between a learning phase and a usage phase which would be very unrealistic with respect to human language learning. (2) It exhibits very quick acquisition (one instance learning) which is also observed in children. Acquisition can of course be followed by performance degradation when new situations arise that require the storage of new views. Once these views have been seen, performance quickly goes up again. This type of behavior is very different from that of inductive learning algorithms (such as the clustering algorithm discussed later) which show random performance for a long time until the right classes have been found.

2.4 Word learning

The present experiment does not focus on the recognition, synthesis and acquisition of speech itself. We use a state-of-the-art speech system for this purpose. It is capable of speaker independant recognition, with no need for training once the wordforms are known. We have supplied the system with a large list of words that might occur in the dialog. Although the recognition rate is high, it is not perfect and so provisions must be made in the language game script to overcome the problem of recognition error. The speech synthesis system is a state of the art text-to-speech synthesiser, similar to the one described in [Dutoit, 1997].

An associative memory stores relations between object views and words. The different views of an object form an implicit category [Kaplan, 1998], based on the fact that they are named the same way. Word learning takes place by reinforcement learning [Sutton and Barto, 1998]. When the classification conforms to the one expected by the human mediator, there is positive feedback (”good”). When there is a negative outcome of the game, as in the second example above, there is negative feedback (”no”). If there is a correction from the mediator as in the second example (”listen; Smiley”), the agent stores a new association between the view and the correcting word (i.e. between the view and Smiley) but only if the association did not exist already.

2.5 Performance data

We have done experiments with these mechanisms on the recognition and naming of the three objects mentioned earlier. For only one object, namely the red ball, a focus of attention mechanism was available (using dedicated hardware on the robot so that it is fast enough). This mechanism is designed to recognise quickly patches of red in an image, to control the head so that this patch gets into the center of the visual field, and to keep tracking the patch when either the object moves or the head moves. During each session the mediator plays a number of classification games with each of these objects. Each game includes a move for sharing attention (e.g. by holding the object in front of the robot), a question like ”what is it”, and approval or correction depending on the answer of the robot. In the case of a bad classification, the right name was uttered by the mediator. The experiments were performed on successive days, under very different lighting conditions, and against different backgrounds so as to get realistic data.

Figure 6 presents the evolution of the average success for four training ses-

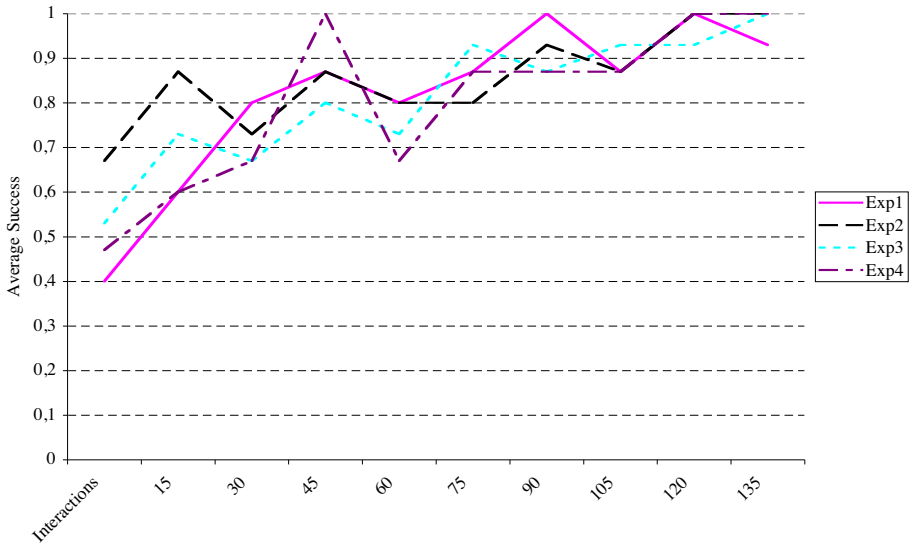

Figure 6: Evolution of the classification success for four different training sessions

sions, each starting from zero knowledge. The success of a game is recorded by the mediator based on the answer of the robot. We see that for all the runs the success climbs regularly to successful communication. It is interesting to note that from the very first games the classification performance is very high. It only takes a few examples to be able to discriminate successfully the three objects in a given environment. But as the environment changes, confusion may arise and new learning takes place, pushing up performance again. This is a property of the instance-based learning algorithm.

	Exp1	Exp2	Exp3	Exp4
Average success	0.81	0.85	0.81	0.80

Table 1: Average success during the training sessions

If we average the classification success over the whole training session, we obtaine an average performance between 0.80 and 0.85 (table 1), which means that on average the robot uses an appropriate name 8 times out of 10. This includes the period of training, so the learning is extraordinarily fast. A closer look at the errors that the robot makes (table 2), shows that the robot makes fewer classification errors for the red ball than for the other two objects. This is due to the focus

of attention mechanism available for tracking red objects. It eases the process of sharing attention on the topic of the game and as a consequence provides the robot with data of better quality. The lack of this capability for the other objects does not cause a failure to learn them however.

word/meaning	Poo-chi	Red Ball	Smiley	Classification success
Poo-chi	34	8	9	0.66
Red Ball	0	52	4	0.92
Smiley	6	2	49	0.86

Table 2: This table shows the word/meaning success rate for one of the sessions.

2.6 Complexity and realism

It is obviously possible to make the perception and categorisation in these experiments more complex. It is probably better to use psychologically more realistic sensory dimensions, such as the Hue Saturation Value space or the L*a*b space, which abstracts the Lightness dimension and uses the opponent channels (red-green; yellow-blue) [Wyszecki and Stiles, 1982]. The L*a*b space moreover maps better on the human experience of colour distance. There are many more sophisticated ways to use instance-based classification as well, for example by using k-nearest neighbor, population coding, radial basis functions, etc. [Witten and Eibe, 2000]. A more sophisticated model of word learning could be used based on maintaining a score between word-meaning associations that reflects the success in using a word [Steels, 1996].

Instead we have adopted the simplest possible solutions in order to make the experiments - which involve real-time interaction with humans - possible. If more complex methods would have been adopted they would not fit on the available hardware and the dialog would no longer have a real time character. We use the same solutions in the other experiments described shortly, so the difference in performance, and hence the conclusions drawn, do not hinge on which choices have been made for perception, categorisation and naming.

3 Comparisons

In the previous section, we have presented a framework that enables the acquisition of a set of first words using the framework of social learning. The framework is effective in the sense that the robot is indeed capable to acquire the meaning of a set of first words, without prior knowledge of the concepts involved nor unrealistic constraints on its movements. We see a number of explanations why communication could successfully be bootstrapped:

1. The language game constrains what needs to be learned. In the specific example developed here, this is knowledge for classifying objects. So, rather than assuming prior innate constraints on the kinds of concepts that should be learned or assuming that unsupervised clustering generates 'natural categories', the social learning hypothesis suggests that constraints are provided by the language games initiated by mediators.

2. The language game guarantees a certain quality of the data available to the learner. It constrains the context, for example with words like "listen" or through pointing gestures. This helps to focus the attention of the learner. Adequate data acquisition is crucial for any learning method and the more mobile and autonomous the learner, the less obvious this becomes.

3. The language game induces a structure for pragmatic feedback. Pragmatic feedback is in terms of success in achieving the goal of the interaction, not in terms of conceptual or linguistic feedback.

4. The language game allows the scaffolding of complexity. The game used in this paper uses a single word like "ball" for identifying the referent. Once single words are learned more complex games become feasable.

5. Social learning enables active learning. The learner does not need to wait until a situation presents itself that provides good learning data but can actively provoke such a situation. We use this particularly for the acquisition of speech. The robot first asks for the confirmation of a wordform before incorporating a new association in memory.

We have shown experimentally that all these conditions are sufficient for the learning of the first word. But the question is now whether they are necessary. This can be examined by changing the experimental conditions We will focus here on two points only: The claim that social learning is necessary to constrain what needs to be learned (point 1 above) and to ensure a sufficient quality of the data (point 2).

3.1 Constraining what needs to be learned

Language games are required to constrain what needs to be learned. For example, the classification game implies that there is a focus on objects even though the learner remains free to employ the specific method used for identifying objects. In another game, the agent might be solicited to perform a certain action and so this would push towards the acquisition of action concepts and conceptualisations of the roles that objects play in the action. Two counterarguments have been advanced against the need for social constraints on the learning situation: (1) Unsupervised learning has been claimed to generate natural categories which can then simply be labelled with the words heard when the same situation occurs (the labelling theory), and (2) Innate constraints can guide the learner to the acquisition of the appropriate concepts.

To examine the first counterargument we have done an experiment using a database of images recorded from 164 interactions between a human and a robot drawn from the same dialogs as those used in social learning. The experiment consisted in using one of the best available unsupervised clustering method in order to see whether any natural categories are hidden in the data. The method is known as the EM algorithm and discussed in the appendix. We have used an implementation from the publically-available data mining software called Weka [Witten and Eibe, 2000]. The EM algorithm does not assume that the learner knows in advance the number of categories that are hidden in the data, because this would indeed be an unrealistic bias which the learner cannot know. Unsupervised neural networks such as the Kohonen map [Kohonen, 2001] would give the same or worse results than the EM algorithm.

The technique of cross validation has been used to guarantee quality of learning. The total data set is divided randomly into 10 sections. Each section contains approximatively the same number of instances for each class. The learning scheme is applied to 9 sections and then tested on the remaining one to obtain the success rate. The learning procedure is executed a total of ten times, each time changing the training and testing sections. The overall success is estimated by the average of the ten experiments. In order to diminish the effect on the initial division into sections, the whole procedure is repeated ten times, and the results are averaged.

As results in table 3 show, the algorithm indeed finds a set of clusters in the data, eight to be precise. But the clusters that are found are unrelated to the classification needed for learning the words in the language. The objects are viewed

in many different light conditions and background situations and the clustering reflects these different conditions more than the specific objects themselves.

Clusters	C0	C1	C2	C3	C4	C5	C6	C7
Poo-chi	9	0	2	11	6	20	0	3
Red Ball	6	2	13	6	0	24	3	2
Smiley	5	2	5	2	12	25	3	3

Table 3: Objects and their clusters, obtained from unsupervised learning.

If we had to assign a name to a single cluster, Poo-Chi would be assigned to C3, the red ball to C2 and Smiley to C5. With this scheme only 30% of the instances are correctly clustered. If we associate each cluster with its best name (as shown in table 4), it would not be much better. Only 47% would be correctly clustered. We suspect that the clustering is more sensitive to contextual dimensions, such as the light conditions or background of the object rather than the object itself.

Cluster	Best name	
C0	(Poo-chi)	9
C1	(Red Ball, Smiley)	2
C2	Red Ball	13
C3	Poo-chi	11
C4	(Smiley)	12
C5	Smiley	25
C6	(Red Ball, Smiley)	3
C7	(Poo-chi, Smiley)	3

Table 4: Clusters and names that best correspond with them.

An additional point is that the EM clustering methods, as any other clustering method, arrives at different clusters depending on the initial conditions (random seeds). There is not necessarily a single solution and the algorithm might get stuck into a local minimum. This implies that different agents all using unsupervised learning to acquire categories are unlikely to end up with the same categories which makes the establishment of a shared communication system impossible.

The conclusion of this experiment is clear. Without the causal influence of language, a learning algorithm cannot learn the concepts that are required to be

successful in language communication. Note that the clustering experiment makes use of very good data (because they were acquired in a social interaction). If an agent is presented with a series of images taken while it is simply roaming around in the world, a clustering algorithm produces even more irrelevant classifications.

What about the second counterargument, namely that innate constraints could guide the learning process. The question here is what these constraints could be. There is nothing in the observed visual data that gives any indication whatsoever that we are dealing with objects. As mentioned earlier, the robot is not even capable to properly segment the image (which would require some sort of template and hence already an idea what the object is). It seems much more plausible that the social interaction helps the learner zoom in on what needs to be learned.

3.2 Constraining data acquisition

The second point is that language games are necessary to set up the right context for the acquisition of the sensory data. If the influence of the mediator weakens, these data become less and less reliable and as a consequence learning becomes less successful. Many machine learning and neural network experiments do not address this question because the data is carefully prepared by the human experimenter. For example, light conditions are kept constant, prior segmentation is performed, the background eliminated, only good examples are kept in the data set, etc.

To examine this point, we did another experiment in which the role of the mediator is reduced, but not entirely otherwise we would end up in unsupervised clustering which was just shown to be inadequate. The robot is now freely moving around in a place where there are three objects: Poo-chi, Ball and Smiley, the same as used in earlier experiments. When the mediator sees the robot looking at one of these objects, he or she supplies the corresponding name (figure 2b). We take this situation to capture the essence of supervised observational learning. Due to the autonomous behavior, it is very well possible, and often the case, that the robot is already looking somewhere else when the word has been processed. Moreover the mediator cannot always know precisely where the robot is looking and so might mention a name for an object when the object is not in view. For one of the objects (ball), the robot has the capability to identify and track the object (more precisely, track something of a red color). This implements a shared attention mechanism which was also used in the social learning experiment. This should again make it easier to learn the word for ball.

The dataset of 150 images (50 for each object) resulting from this interaction was then tested with an instance-based learning method (the same as used in the first experiment) and the results compared to what the same algorithm produces for the data set obtained through more intense interaction with a human mediator. Observational supervised learning reaches an average of 59% success which is intermediary between the results of unsupervised classification and social learning. A closer look at the classificiation errors that the robots makes (table 5 and figure 7) shows that the Poo-chi and the Smiley are very often confused. For the Red Ball better results are obtained (although not as good as in the case of social learning), which is explained by the fact that the robot is spontaneously attracted to red objects and thus naturally focuses its gaze on them.

word/meaning	Poo-chi	Red Ball	Smiley	Classification success
Poo-chi	20	13	17	0.4
Red Ball	3	42	5	0.84
Smiley	13	10	27	0.54

Table 5: Word/meaning success rate for a session without social embedding

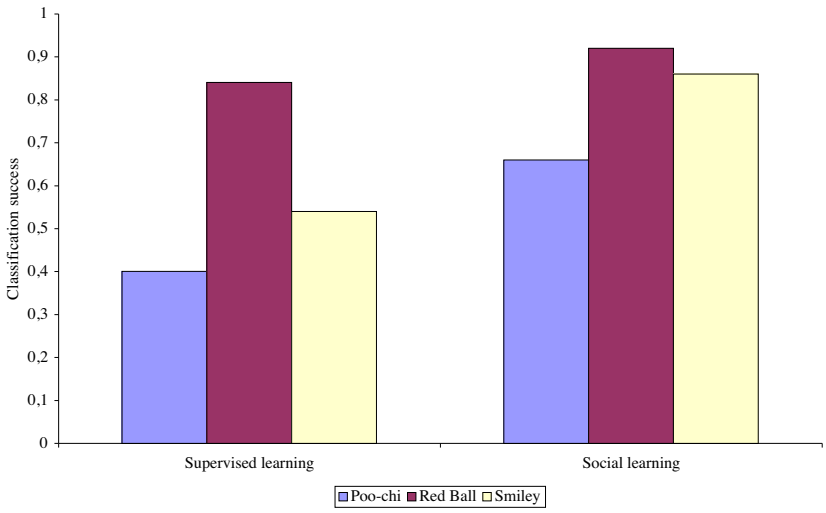

Figure 7: Comparison of classification success for observational and social learning

Two conclusions can be drawn: (1) When the role of the mediator is reduced,

the learning data becomes less reliable and hence classification success deteriorates. Instead of an overall 82 % successrate, we have a 59 % success rate. The recognition of Poo-chi has an average success rate of 40% and that of Smiley 54%, which contrasts with a success-rate of 66% and 86% respectively based on the 'better' learning data. If the role of the mediator is reduced still further (for example if the mediator is less careful in supplying a word for an object), these results aggrevate further. (2) When there is sharing of attention results are better. Thus the success-rate for identification of the ball is consistently better than that of other objects (around 84% in both cases).

The conclusions of this second experiment are therefore clear: When the interactivity characteristic for social learning is reduced, the quality of the data available for learning is reduced and hence the effectiveness of the outcome.

3.3 Scaling Up

We have done additional experiments to examine the effect of scaling up the set of objects and consequently the set of words. Reporting on these experiments is beyond the scope of the present paper, but some conclusions can be briefly mentioned: (1) Not surprisingly, increasing the set of features helps. For example, we have used other colour spaces in addition to the RGB space and found that this increases the reliability of object recognition. (2) Very soon (with half a dozen objects) instance-based learning reaches limits which start to degrade performance. The problems are twofold: (i) As the number of views stored in memory reaches a critical point, the time needed to recognise an object is too long to sustain fluent real-time interaction. (ii) As all sensory dimensions are indiscriminately taken into account, the distance measure used becomes less and less effective.

We have therefore experimented with other concept learning strategies - which can only be used once the initial bootstrapping as reported in this paper has taken effect. The first strategy is to learn the most significant visual dimensions, which can be done by statistical methods that examine the predictive value of each dimension. Statistical correlations of each dimension with the object classes and the intercorrelations among dimensions can be computed and the dimensions with the highest class correlation and the lowest intercorrelations retained. This collapses the space into a more compact and hence more efficient and more reliable conceptual space and immediately improves the efficacity of instance-based learning. The second strategy is to gradually complement instance-based learning with rule induction or the induction of decision-trees, operating over the data obtained in a

social learning framework.

4 Conclusion

This paper examined what it would take to re-enact a situation in which an autonomous physical being can begin to acquire 'first words'. We have carried out a realistic robotic experiment in the sense that the robot is not only ignorant about the words in the language but also about the perceptually grounded concepts underlying these words. Moreover the robot is mobile and fully autonomous. As a consequence we have been forced to confront a situation in which the data available for learning is not given by the human experimenter but must be acquired by the robot as it is interacting with a human in a complex real world environment.

The paper argued in favor of social learning as opposed to individualistic learning. This conclusion has also been defended by students of child language acquisition [Tomasello, 2000] and researchers engaged in teaching words to animals [Pepperberg, 1991]. In social learning, the mediator plays a crucial role to constrain the situation, scaffold complexity, and provide pragmatic feedback. Social learning makes it easier to introduce a causal influence of language on category formation which was shown to be necessary if categories learned by the robot are to be similar enough to those already used in an existing human culture. We have also argued in favor of a gradual bootstrapping process

We now return to the question posed in the beginning of the paper: What are the crucial prerequisites for the acquisition of 'the first words'. We have argued these to include the following:

1. The ability to acquire and engage in structured social interactions, i.e. interactions that follow a routinised pattern. This requires abilities like turn taking, recognition of others, focus of attention, and other capabilities associated with a 'theory of mind'.

2. The presence of a mediator. The mediator is already part of a culture and therefore influences concept acquisition so that it conforms to what is needed for a specific language.

3. Incremental learning algorithms for the acquisition of concepts, such as the instance-based learning schema used in this paper.

4. An associative memory for storing the relation between words and meanings and reinforcement learning methods for the acquisition of these associations.

This paper has not addressed many other issues that can be raised in the present context. The types of words that are learned are not uncommon for the first words also used by children but we did not discuss the acquisition of words for action, or any other conceptual domain. The issue of grammar has not been addressed. In any case, it arises in children only after an initial lexicon is in place. Some hypotheses and robotic experiments of the transition to grammar can be found in [Steels, 1998]. We did not address the issue how language games themselves are learned or invented. This is clearly a very difficult problem and will be addressed in other papers. Finally, we did not address scale-up. Our additional experiments not reported in this paper have already shown however that instance-based learning is adequate for the initial phases of bootstrapping but has to be complemented with other learning methods to scale up concept acquisition and hence word learning.

We believe that there is great value in carrying out robotic experiments of the sort shown in this paper because they force us to deal with realistic assumptions. Many learning methods achieve quite reasonable performance in the supervised learning of words and meanings, but they sidestep the problem where the learning data comes from. Social learning does address this issue by providing a framework for helping the learner to focus on what needs to be learned and to gather high quality data critical for learning.

5 Acknowledgement

Research presented in this paper was conducted at the Sony Computer Science Laboratory in Paris. We are grateful to members of Sony's Digital Creatures Lab in Tokyo who have designed the AIBO and provided us with invaluable information to do the experiments. We also thank Nicolas Neubauer for help in the gathering of data.

References

[Baron-Cohen, 1997] Baron-Cohen, S. (1997). *Mindblindness: an essay on autism and theory of mind*. MIT Press, Boston, MA, USA.

[Beymer and Konolige, 1999] Beymer, D. and Konolige, K. (1999). Real-time tracking of multiple people using continuous detection. In *IEEE Frame Rate Workshop*. (www.eecs.lehigh.edu/ tboult/FRAME/Beymer).

[Billard et al., 1998] Billard, A., Dautenhahn, K., and Hayes, G. (1998). Experiments on human-robot communication with robota, an interactive learning and communicating doll robot. In Edmonds, B. and Dautenhan, K., editors, *Socially situated intelligence workshop (SAB 98)*, pages 4–16.

[Bloom, 2000] Bloom, P. (2000). *How children learn the meanings of words*. Mit Press, Cambridge, MA.

[Bowerman, 2001] Bowerman, M.and Levinson, S. C. (2001). *Language acquisition and conceptual development*. Cambridge U.P.

[Broeder and Murre, 2000] Broeder, P. and Murre, J. (2000). *Model of language acquisition. Inductive and deductive approaches*. Oxford university press, Oxford, UK.

[Chomsky, 1975] Chomsky, N. (1975). *Reflections on Language*. Pantheon, New York.

[Clancey, 1997] Clancey, W. (1997). *Situated cognition : On human knowledge and computer representations*. Cambridge University Press, Cambridge, UK.

[Clark, 1987] Clark, E. (1987). The principle of constrast : A constraint on language acquisition. In MacWhinney, B., editor, *Mechanisms of language acquisition*. L. Erlbaum Hillsdale NJ.

[Dutoit, 1997] Dutoit, T. (1997). *An introduction to Text-To-Speech Synthesis*. Kluwer Acamdec Publishers, Dordrecht.

[Ellman, 1993] Ellman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99.

[Finlayson et al., 1998] Finlayson, G., Schiele, B., and Crowley, J. (1998). Comprehensive colour image normalization. In *ECCV́98, European Conference on Computer Vision*.

[Fischer et al., 1994] Fischer, C., Hall, G., Rakowitz, S., and Gleitman, L. (1994). When it is better to receive than to give : syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92:333–375.

[Fodor, 1983] Fodor, J. (1983). *The modularity of mind*. MIT Press, Cambridge, MA.

[Fodor, 1999] Fodor, J. (1999). *Concepts - where cognitive science went wrong*. Oxford University Press.

[Fujita et al., 2001] Fujita, M., Costa, G., Takagi, T., Hasegawa, R., Yokono, J., and Shimomura, H. (2001). Experimental results of emotionally grounded symbol acquisition by four-legged robot. In Muller, J., editor, *Proceedings of Autonomous Agents 2001*.

[Fujita and Kitano, 1998] Fujita, M. and Kitano, H. (1998). Development of an autonomous quadruped robot for robot entertainment. *Autonomous Robots*, 5.

[Halliday, 1987] Halliday, M. (1987). *Learning how to mean*. Cambridge university press, Cambridge, UK.

[Harnad, 1990] Harnad, S. (1990). The symbol grounding problem. *Physica D*, 40:335–346.

[Heine, 1997] Heine, B. (1997). *Cognitive foundations of grammar*. Oxford University Press, Oxford, UK.

[Kaplan, 1998] Kaplan, F. (1998). A new approach to class formation in multi-agent simulations of language evolution. In Demazeau, Y., editor, *Proceedings of the third international conference on multi-agent systems (ICMAS 98)*, pages 158–165, Los Alamitos, CA. IEEE Computer Society.

[Kaplan et al., 2001] Kaplan, F., Oudeyer, P.-Y., Kubinyi, E., and Miklosi, A. (2001). Taming robots with clicker training : a solution for teaching complex behaviors. In Quoy, M., Gaussier, P., and Wyatt, J. L., editors, *Proceedings of the 9th European workshop on learning robots*, LNAI. Springer.

[Kohonen, 2001] Kohonen, T. (2001). *Self-Organizing Maps. Extended edition*. Springer, Berlin.

[Langacker, 1991] Langacker, R. (1991). *Foundations of cognitive grammar*. Stanford University Press, Stanford, CA.

[Markman, 1994] Markman, E. (1994). Constraints on word meaning in early language acquisition. *Lingua*, 92:199–227.

[Mel, 1997] Mel, B. (1997). Seemore : combining color, shape and texture histrogramming in a neurally inspired approach to visual object recognition. *Neural Comp.*, 9:777–804.

[Pepperberg, 1991] Pepperberg, I. (1991). Learning to communicate : the effects of social interaction. In Klopfer, P. and Bateson, P., editors, *Perspectives in Ethology*. Plenum, New York.

[Popper, 1968] Popper, K. (1968). *The logic of scientific discovery*. Hutchinson, London, revised edition.

[Quine, 1960] Quine, W. (1960). *Word and Object*. The MIT Press, Cambridge, MA.

[Regier, 1996] Regier, T. (1996). *The Human Semantic Potential: Spatial Language and Constrained Connectionism*. Neural Network Modelling and Connectionism. MIT Press, Boston, MA, USA.

[Roy, 1999] Roy, D. (1999). *Learning form sights and sounds : a computational model*. PhD thesis, MIT Media Laboratory.

[Schiele and Crowley, 1996] Schiele, B. and Crowley, J. (1996). Probabilistic object recognition using multidimensional receptive field histograms. In *ICPR 96 Proceedings of the 13th International Conference on Pattern Recognition, Volume B*, pages 50–54.

[Siskind, 1995] Siskind, J. (1995). Grounding language in perception. *Artificial Intelligence Review*, 8:371–391.

[Smith, 2001] Smith, L. (2001). How domain-general processes may create domain-specific biaises. In Bowerman, M. and Levison, S. C., editors, *Language acquisition and conceptual development*, pages 101–131. Cambridge university press, Cambridge, UK.

[Steels, 1996] Steels, L. (1996). Self-organizing vocabularies. In Langton, C. and Shimohara, T., editors, *Proceeding of Alife V*, Cambridge, MA. The MIT Press.

[Steels, 1998] Steels, L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103:1–24.

[Steels, 2001a] Steels, L. (2001a). Language games for autonomous robots. *IEEE Intelligent systems*, pages 17–22.

[Steels, 2001b] Steels, L. (2001b). The methodology of the artificial. *Behavioral and brain sciences*, 24(6).

[Steels, 2001c] Steels, L. (2001c). Social learning and language acquisition. In McFarland, D. and Holland, O., editors, *Social robots*. Oxford University Press, Oxford, UK.

[Steels and Brooks, 1994] Steels, L. and Brooks, R. (1994). *The 'artificial life' route to 'artificial intelligence'. Building Situated Embodied Agents.* Lawrence Erlbaum Ass, New Haven.

[Steels and Kaplan, 1998] Steels, L. and Kaplan, F. (1998). Stochasticity as a source of innovation in language games. In Adami, C., Belew, R., Kitano, H., and Taylor, C., editors, *Proceedings of Artificial Life VI*, pages 368–376, Cambridge, MA. The MIT Press.

[Steels et al., 2002] Steels, L., Kaplan, F., McIntyre, A., and Van Looveren, J. (2002). Crucial factors in the origins of word-meaning. In Wray, A., editor, *The Transition to Language*. Oxford University Press, Oxford, UK.

[Steels and Vogt, 1997] Steels, L. and Vogt, P. (1997). Grounding adaptive language games in robotic agents. In Harvey, I. and Husbands, P., editors, *Proceedings of the 4th European Conference on Artificial Life*, Cambridge, MA. The MIT Press.

[Sutton and Barto, 1998] Sutton, R. and Barto, A. (1998). *Reinforcement learning : an introduction*. MIT Press, Cambridge, MA.

[Talmy, 2000] Talmy, L. (2000). *Toward a cognitive semantics : concept structuring systems (language, speech and communication)*. The MIT Press, Cambridge, MA.

[Tomasello, 2000] Tomasello, M. (2000). *The cultural origins of human cognition*. Harvard U.P.

[Vogt, 2000] Vogt, P. (2000). *Lexicon grounding on mobile robots*. PhD thesis, Vrije Universiteit Brussel.

[Witten and Eibe, 2000] Witten, I. and Eibe, F. (2000). *Data mining*. Morgan Kaufmann Publishers.

[Wittgenstein, 1953] Wittgenstein, L. (1953). *Philosophical Investigations*. Macmillan, New York.

[Wyszecki and Stiles, 1982] Wyszecki, G. and Stiles, W. (1982). *Color science : concepts and methods, quantitative data and formulae*. John Wiley and sons, New York.

APPENDIX I.

This appendix briefly describes the EM clustering algorithm used for the experiments in unsupervised learning of classification concepts [Witten and Eibe, 2000]. The algorithm is based on a statistical model called finite mixture. A mixture is a set of k probability distributions, representing k clusters. In the case of Gaussian distributions, each distribution D is determined by two parameters its mean μ_D and its standard deviation σ_D. If we know that $x_1, x_2 \ldots x_n$ belong to the cluster D, μ_D and σ_D are very easy to compute. For instance in the simple case in which there is only one numeric attribute x :

$$\mu_D = \frac{x_1 + .. + x_n}{n} \tag{2}$$

$$\sigma_D^2 = \frac{(x_1 - \mu_D)^2 + .. + (x_n - \mu_D)^2}{n - 1} \tag{3}$$

If we know μ and σ for the different clusters, it is also easy to compute the probabilities that a given instance comes from each distribution. Given an instance x, the probability that it belongs to cluster D is :

$$Pr[D/x] = \frac{Pr[x/D].Pr[D])}{Pr[x]} = \frac{f(x; \mu_D, \sigma_D).Pr[D])}{Pr[x]} \tag{4}$$

where $f(x; \mu_D, \sigma_D)$ is the normal distribution function for cluster D

$$f(x; \mu_D, \sigma_D) = \frac{1}{\sqrt{2\pi}\sigma_D} \exp \frac{-(x - \mu_D)^2}{2.\sigma_D^2} \tag{5}$$

The EM algorithm stands for ”expectation-maximization”. Given an initial set of distributions, the first step is the calculation of the cluster probabilities (the ”expected” class values). The second step is the calculation of the distribution parameters by the ”maximization” of the likelihood of the distributions given the data. These two steps are iterated like in a *k-means* algorithm.

For the estimation of μ_D and σ_D, a slight adjustment must be made compared to equations 2 and 3 due to the fact that only cluster probabilities, not the clusters themselves, known for each instance. These probabilities act like weights.

$$\mu_D = \frac{w_1.x_1 + .. + w_n.x_n}{w_1 + .. + w_n} \tag{6}$$

$$\sigma_D^2 = \frac{w_1.(x_1 - \mu_D)^2 + .. + w_n.(x_n - \mu_D)^2}{w_1 + .. + w_n} \tag{7}$$

where the x_i are now all the instances and w_i is the probability that instance i belongs to cluster D.

The overall likelihood of a distribution set is obtained by multiplying the probabilities of the individual instances i:

$$Q = \prod_i \sum_D p_D.Pr[x_i/D] \tag{8}$$

where $Pr[x_i/D]$ is determined from $f(x_i; \mu_D, \sigma_D)$. Q is an indicator of the quality of the distribution. Q increases at each iteration. The algorithm stops when the the increase of the log-likelihood becomes negligible (e.g. less than 10^{-10} increase for ten successive iterations).

The EM algorithm is guaranteed to converge to a maximum but not necessary to the global maximum. The algorithm could be repeated several times, with a different initial configuration. By varying the number of clusters k, it is possible to determine the one which maximize Q and thus the "natural" number of clusters.