Chapter 7

Social Language Learning

by Luc Steels

Luc Steels is director of the Sony Computer Science Laboratory in Paris and professor of Artificial Intelligence at the University of Brussels (VUB). Steels researches the origins and learning of language through computational and robotic models. He has developed the framework of evolutionary language games as a way to study the acquisition of concepts and language grounded and situated in experience. The framework allows a systematic study of transactional learning and could lead to novel learning environments for language acquisition.

This paper explores a theory of learning which emphasises social interaction and cultural context. The theory contrasts with individualistic theories of learning, where the learner is either seen as passively receiving large sets of examples and performing some sort of induction to arrive at abstract concepts and skills, or as a genetically pre-programmed organism where the role of the environment is restricted to setting some parameters. I focus the discussion on the question how meaning is constructed, in other words how people go from information to knowp ofledge. I am particularly interested how grounded meaning arises, i.e. meaning anchored in sensori-motor experiences. This is the question raised earlier (in Chapter 1) in the discussion on how information turns into knowledge. I am also interested in how shared meanings can be developed through communication and negotiation. 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?

The paper explores social and cultural learning using a novel methodology, namely the construction of artificial systems, i.e. robots, that implement certain theoretical assumptions and hence allow us to examine with great precision how certain learning mechanisms work and what they can achieve or not achieve. Some implications for education are presented towards the end.

7.1 The Origins of Meaning

It has been argued that one of the main objectives for children, particularly in primary school, is learning how to create meaning. This starts from a first magical moment around the agent of 18 months [16], but continues throughout childhood. As pointed out by Loris Malaguzzi (see Chapter 3 by Carla Rinaldi), the child is a semiologist who invents new meanings, and negotiates words to talk about these meanings with friends or parents and teachers. Also later in secondary school and higher education, the acquisition of new meanings and ways to communicate them is a crucial skill. For example, it could be argued that all significant advances in science arise from conceptual breakthroughs. My objective is to understand this creative semiological process in much more detail so that we can help children to acquire crucial semiological skills.

Let me start by summarising the intense debates that have taken place in the cognitive science literature over past decades on this question of the origins of meaning. These debates echo the nature/nurture debate and the discussion of a constructivist synthesis introduced by Johnson in this contribution (see Chapter 5).

Labelling versus Social Grounding

There are basically two main lines of thinking on the question of how language and meaning are bootstrapped: individualistic learning and social learning. In the case of individualistic learning, the child is assumed to have experienced, as input, a large number of example cases where speech is paired with specific situations. They either already mastering the necessary concepts or are able to extract through an inductive learning process what is essential to and recurrent in these situations, in other words to learn the appropriate categories underlying language, and then associate these categories with words. This is known as cross-situational learning [13]. Others have proposed a form of contrastive learning based on the same sort of data, driven by the hypothesis that different words have different meanings [8]. 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.

Figure 7.1: Children can solve problems through social interactions which none of them can solve individually. Interacting together with the physical world is part of play. It is as necessary and insightful as instructional learning. (mw)

I 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 [4]. It is known that induction by itself is a weak learning method, in the sense that it does not give identical results from the same data and may yield irrelevant clustering compared to human categories. To counter this argument, it is usually proposed that innate constraints help the learner zoom in on the important aspects of the environment.

In the case of social learning, interaction with other human beings is considered crucial ([5], [35], [30]). 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 henceforth 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, I also call social learning mediated learning. The goal of the interaction is not really teaching, which is why I use the term mediator as opposed to teacher. The goal is rather something practical in the world, for example, to identify an object or an action. The mediator helps the goal to be achieved and is often the one who wants to see the goal achieved.

The mediator has various roles: They set constraints on the situation to make it more manageable (scaffolding), give encouragement on the way, provide feedback, and act upon the

consequences of the learner's actions. Even though language is being learned, the feedback is not directly about language, and is certainly not about the conceptualisations implicitly underlying language. The latter are never visible. The learner cannot telepathically inspect the internal states of the speaker and the mediator cannot know which concepts are already known to 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 of which there are 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 [20]. Typically the learners use themselves 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 fill in missing holes. 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 Relation between Language and Meaning

The debate between individualistic versus social learning is related to the equally hotly debated question as to whether or not there is a causal role for language in concept learning. From the viewpoint of the labelling theory, the acquisition of concepts occurs independently of and prior to language acquisition, either because concepts are innate (nativism) or because they are acquired by an inductive learning process (empiricism) [17]. 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 [14]. The acquisition of words is seen as a problem of learning labels for already existing concepts.

Concerning then the issue of how the concepts themselves are acquired, two opposing schools of thought can be found: nativism and empiricism. Nativists like Fodor [14] 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 [6], the fundamental weakness of inductive learning, 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 [10]. Thus a large number of situations in which a red coloured object appears are seen by the learner, and clustered into 'natural categories'. These natural categories then form the basis for learning word meaning. An intermediate position is found with constructivists, who see a steady interplay between genetic constraints and learning processes (see Chapter 5 by Johnson and [11]), but still view the learner very much as individually bootstrapping their knowledge.

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 that this role is particularly (but not exclusively) played through language. This has been argued both by lin-

Figure 7.2: From the viewpoint of a social approach to learning, interaction beween learners and mediators is absolutely crucial, particularly for the acquisition of language. (mw)

guists 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 [34]. Language acquisition therefore is believed to go hand in hand with concept acquisition [3]. Moreover language-specific conceptualisations change over time in a cultural evolution process, which in turn causes grammatical evolution that may induce further conceptual change [18].

This does not mean that there are no similarities between the underlying conceptualisations of different languages. For example, the distinction between objects (things, people) on the one hand and events (actions, state changes) on the other appears universal [37]. Similarly many categorial dimensions like space, time, aspect, countability, kinship relations, etc., are lexicalised in almost all languages of the world, even though there may be differences in how this is done. Thus, some languages lexicalise kinship relations as nouns (like English: father) and others as verbs [12]. But profound conceptual differences in the way different languages conceptualise reality are not hard to find ([34], [3]) and they also show up in other cognitive tasks such as memory tests [9]. For example, the conceptualisation of the position of the car in “the car is behind the tree” is just the opposite in most African languages. The front of the tree is viewed as being in the same direction as the face of the speaker and hence the car is conceptualised as in front of the tree as opposed to behind the tree [18]. These examples suggest that different human cultures invent their own ways to conceptualise reality and propagate it through language, implying a strong causal influence of language on concept formation. 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 here is that for those concepts underlying natural language communication, this causal influence not only exists but is necessary, i.e. these concepts necessarily have a cultural dimension.

Ludwig Wittgenstein 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 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 a particular action to 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 who 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.

7.2 How to Study Learning?

There is a long tradition in psychology that studies learning by observing learning behaviour, particularly that of children, or by performing experiments in which human subjects have to learn something and their performance is monitored. More recently, research in Artificial Intelligence has advanced sufficiently or it to become possible to use an alternative approach. It is now possible to take a particular learning method claimed to be effective for a certain task, turn it into an artificial system (typically a computer program), feed it with the data that a human learner is supposed to have, and see whether the learning method is up to the task. Often the learning method or its implementation attempt to be faithful to human behavioral data or compatible with what is known about the brain.

The Methodology of the Artificial

This methodology has also been applied to the question of meaning and language acquisition (see [4]). However, these modelling efforts so far mainly use an individualistic approach with passive, observational learning. Precise models for social learning are lacking. In the absence of such models, it is difficult to compare the different positions in the debate seriously without sliding into rhetoric. The first goal of my work has therefore been to develop concrete models of social learning and compare their behavior to individualistic learning. What is also lacking are experiments to test the cultural influence of language on meaning creation and propagation. And so my second goal has been to develop precise models showing that cultural influence and context-dependent meaning creation are indeed the most plausible and effective way for individuals to bootstrap themselves into a language culture.

Previous work in the computational modelling of language learning has been entirely based on software simulations. Given the enormous complexity of the cognitive processing required for language, even for handling single words, computer simulations are the only way one can test formal models. But if one believes in the importance of embodiment, social interaction, and the necessity of grounding language in the world, we must go one step further and use autonomous mobile robots. We can then try to set up experiments where autonomous robots, in strong interaction with humans and grounded in the world through a physical sensori-motor apparatus, develop language-like communication systems. If the

robots manage to do this, then we have discovered a plausible way in which language is bootstrapped. Recent dramatic advances in robotics technology have made this approach feasible, even though of course it requires building very complex systems, and the remainder of this paper is based on experience gained with such experiments.

To many social scientists, the idea of using autonomous robots for testing theories of cognition and communication is very unusual and they are very sceptical that anything could come out of it. But there are important advantages:

1. The experiments 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 cannot begin to pretend that robotic experiments model children in any realistic way, nor the environments in which they typically operate. But our goal is to compare theories of how language and communication develop, so realism is not an issue.

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 algorithms available today. This puts limits on what can be technically achieved today and so experiments need to be designed within these limits.

AIBO's first words

Together with a number of collaborators, in particular Frederic Kaplan, I have been using various kinds of robots in experiments that try to reconstruct the very beginning of language and meaning. The experiments discussed further in this paper are based on an enhanced version of the Sony AIBO^{TM} robot (see figure 7.3). This robot is fully autonomous and mobile with more than a thousand behaviors, coordinated through a complex behavior-based motivational system. The AIBO features 4-legged locomotion, a camera for visual input, two microphones, and 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 to which it has been designed and built. Moreover the AIBO is designed to encourage interaction with humans, which is what we need for experiments in social human-robot interaction. It comes with a very wide range

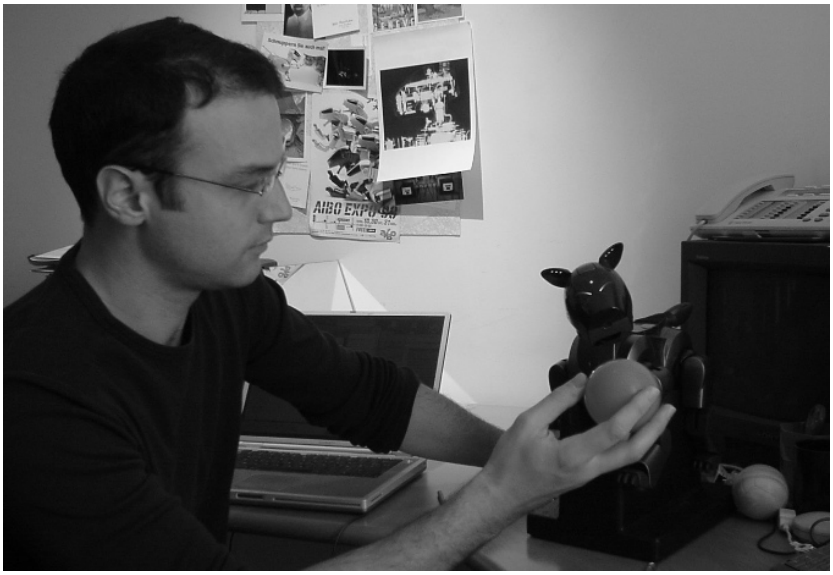

Figure 7.3: Our robot is an enhanced version of the commercially available AIBO. It is linked to an additional computer through a radio connection.

of capabilities which are necessary to establish the conditions for social interaction, such as the ability to look at an object as a way to draw attention of the speaker to the object.

The experiments discussed further in this paper, described in more detail in [31] work on an enhanced version of the AIBO because there is not enough computing power on-board to do them. We 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 already available autonomous behavior, without controlling it completely. A second computer implements speech recognition facilities which enable 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 "lethargic" mood.

By using real-world autonomous robots, our experiments differ from other computational experiments in word learning (such as [24]) 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. Here we approach much more closely the conditions of a one year old child who is moving around freely with no preconception of what the meaning of a word might be. In fact, we tackle a situation which is even more difficult than that of a child, because we assume that the robot has not yet acquired any concepts that could potentially be used or adapted for language communication.

7.3 Language Games

In previous work we found that the notion of a game, and more specifically a language game, is a very effective way to frame social and cultural learning [30]. A game is a routinised sequence of interactions between two agents involving a shared situation in the world. Psychological research into the transition from pre-linguistic to linguistic communication has found clear evidence for an important role of game-like interactions, which initially are purely based on gestures, before including vocalisations that then become words [15]. The players in a language game 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 all-encompassing than behaviors in the sense of behavior-based robots [32]. They may run for several minutes and invoke many behaviors and cognitive activities on the way. They may be interrupted to be resumed later.

Competences in a language game

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] Moooo.
Child: [just observes]
Father: What does the dog say? [points to dog] Woof.
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 ‘mooo’ and ‘woof’ 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 of two. 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 the 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.

Figure 7.4: Language games rest on a rich substrate of competences concerned with turn-taking, face identification, and sharing attention. These are acquired through play in the first years of life. (Picture by Jan Belgrado)

- 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 on, 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 at the speaker.

These various activities are often associated with having a 'theory of mind' [1]. It is clear that these prerequisites as well as those 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 in our robots (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 that handles the integration and scheduling in real-time of behaviors to implement interactive dialogs. We do not pretend that any of these components achieves human level performance, far from it. But they are enough to carry out experiments addressing the issues raised in this paper and observers are usually stunned by the level of performance already achieved.

The Classification Game

We have been experimenting with various kinds of language games, most notably a guessing game [29], 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. Another game we have studied intensely is the classification game, and that game will be used as a source of illustration in the rest of the paper. The classification game is similar to the guessing game, except that there is only a single object in the visual image to be classified.

Figure 7.5 gives an idea of the difficulties involved in playing a classification game and they dramatically illustrate the difficulties that children must encounter when constructing their first meanings. All these images have been captured with AIBO's camera. Different ambient lighting conditions may completely change 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

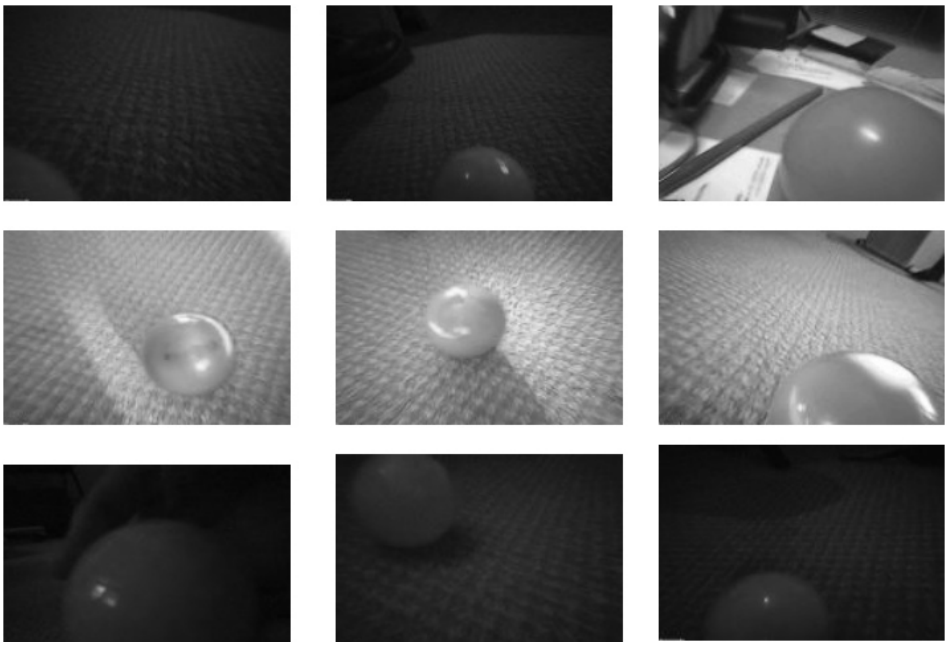

Figure 7.5: Different views of a red ball as captured by the robot's camera.

are totally different. 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 extremely important 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 lighting conditions) many of the real-world problems that must be solved in bootstrapping communication would disappear, diminishing the strength of the conclusions that can be drawn.

Learning Classes

Obviously the classification game must rely on a cognitive subsystem that is able to classify objects. There are many possible ways to implement such a system and many techniques are known in Artificial Intelligence literature as to how to learn the required classes. I believe that it is not so important which learning technique is used, rather how the method is integrated within the total behavior of the learner.

The first question to be addressed in building a classificatory system is how to segment objects. Twenty years of research in computer vision have shown that object segmentation is notoriously difficult. It is even believed to be impossible, unless there is already a relatively 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 foolproof. 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 initially concepts for objects will be highly context-sensitive, as opposed to clear Platonist abstractions. 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.

The second question concerns the method itself. We have used an instance-based method of classification ([2]), which means that many different 'views' are stored of an object situated in a particular context, and classification takes place by a nearest neighbor algorithm: the view with the shortest distance in pair-wise comparison to the input image is considered to be the 'winning' view. We cannnot really say that the memory "represents" objects, because the robot has no notion yet of what an object is, and its memory always stores an object within a certain context.

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. (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 stored as well, performance quickly goes up again. This type of learning 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.

Word learning

To play the classification game, the robot must also have a cognitive subsystem for storing and retrieving the relation between object views and words. Each association has an associated score, which represents past success in using that association as speaker or listener. When speaking, the robot always chooses the association with the highest score so that there is a positive feedback loop between the success of a word and its subsequent use. Word learning takes place by reinforcement learning [33]: when the classification conforms to that expected by the human mediator, there is positive feedback, and the score of the association that was used goes up. At the same time, there is lateral inhibition of alternative hypotheses. When there is a negative outcome of the game, there is negative feedback. The score of the association that was used is decreased. If there is a correction from the mediator, the robot stores a new association between the view and the correcting word, but only if the association did not already exist.

Scripts

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. 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 7.6 a).

```
3. Human: Look
```

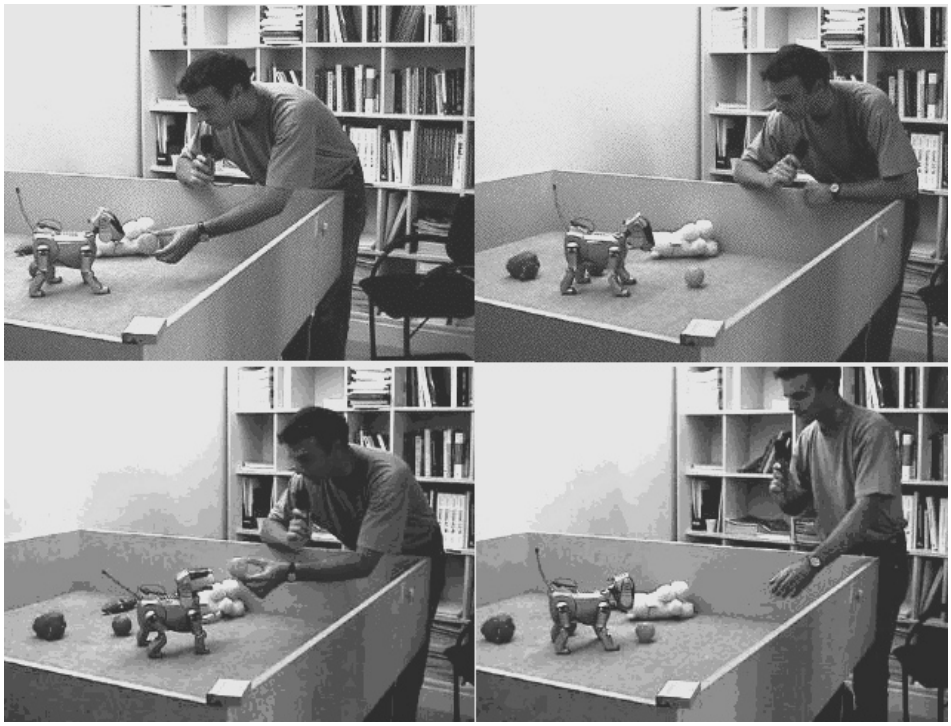

Figure 7.6: Different steps in a language game.

The word “look” helps to focus attention and signals the beginning of a language game. The robot now concentrates on the ball, starts tracking it, and signals focus by looking at the ball (figure 7.6 a) and trying to touch it (figure 7.6 b). It further signals attention by looking first at the speaker (figure 7.6 c) and then back at the ball (figure 7.6 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 yet know a word for this object, so a learning activity starts. The robot first asks 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 could go gone wrong in this episode and the human mediator would typically spontaneously provide additional feedback. For example, the wrong word might be heard due to problems with speech recognition, the robot might not be 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.

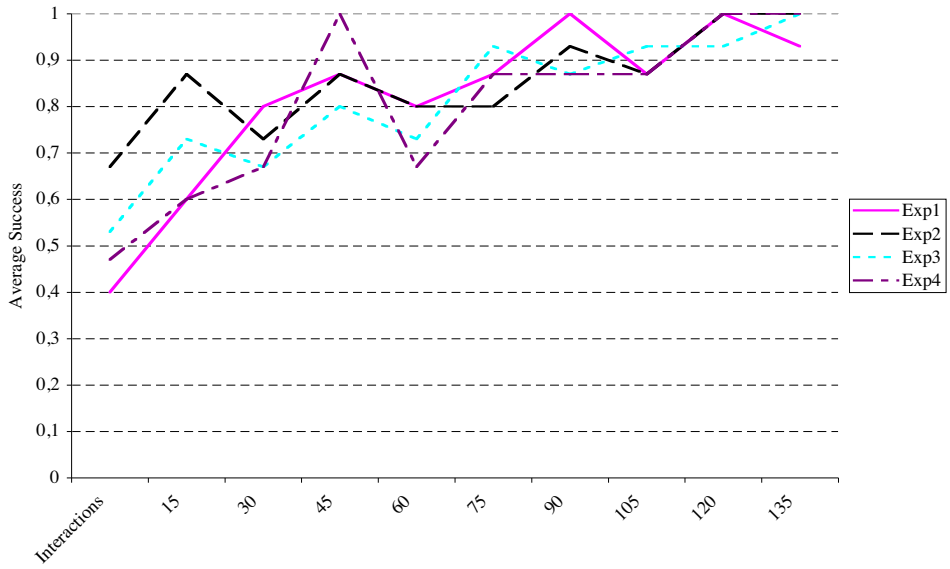

Figure 7.7: Evolution of the classification success for four different training sessions.

7.4 Experimental Results

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. The objects we used were an AIBO imitation called Poo-Chi, a yellow object, Smiley, a red ball, etc. The experiments were performed on successive days, under very different lighting conditions, and against different backgrounds in order to obtain realistic data. 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.

Successful Learning of 'the first words'

Figure 7.7 presents the evolution of the average success for four sessions, each starting from zero knowledge (no words and no concepts). 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 7.1: Average success during the training sessions

If we average the classification success over the whole training session, we obtain 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 however not cause a failure to learn them.

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

Table 7.2: This table shows the word/meaning success rate for one of the sessions.

It is obviously possible to make the perception and categorisation in these experiments more complex. 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 had been adopted they would not fit on the available hardware and the dialog would no longer have a real-time character.

Comparison with non-social learning

Two counter-arguments have been advanced against the need for strong social interaction on first word learning: (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) some researchers have proposed that innate constraints guide the learner to the acquisition of the appropriate concepts.

To examine the first counter-argument Frédéric Kaplan has 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 of using one of the best available unsupervised clustering method (the EM method) in order to see whether any natural categories are hidden in the data. 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 would give the same, or worse, results than the EM algorithm.

As the 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 under many different lighting 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 7.3: 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 7.4: Clusters and names that best correspond with them.

An additional point is that the EM clustering methods, as any other clustering method, arrive 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 individuals, that all use 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.

I now turn to 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 of properly segmenting the image (which would require some sort of template and hence already an idea of what the object is). It therefore seems much more plausible that the social interaction helps the learner zoom in on what needs to be learned.

These experiments only give a glimpse of our methodology at work and we, as well as other colleagues in this field, have done much more. What is important is that we can examine the implications of different theoretical assumptions by varying components of the artificial system, feed it different data, put it into another environment, etc. Each time we can carefully monitor the results of the experiments and examine the causal relation between a theoretical assumption and observed behavior. This approach therefore introduces an experimental methodology into the study of learning which is not possible with human subjects.

7.5 Implications

What can we learn from this kind of experiment that is relevant for the future of learning? I believe there are three important take-home lessons.

The Nature of Early Word Learning

First, these experiments tell us something about the nature of learning, more concretely the learning of the very first words and their associated meanings. They show that there is not a single magical mechanism but that the key lies in the integration of many different skills, ranging from sharing attention, turn-taking, vision, and categorisation to word learning. They also show that it is not enough to look at the individual in isolation.

The interaction with a mediator is crucial and games are a good way to structure these interactions. I believe that mediation is important for the following reasons:

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 have been learned, more complex games become feasable. We have already experimented with games for learning names of actions or more complex descriptions of scenes.

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 have particularly used this for the acquisition of speech. The robot first asks for the confirmation of a wordform before incorporating a new association in its memory.

Figure 7.8: The emphasis on learner-centered learning and IT-based learning tools used on an individual basis, should not make us forget the important role of interaction as a motor and enhancer of learning processes.

Teaching Practices

Our research results are too recent to have had any impact on educational practice. However I believe that they are potentially far reaching. The child needs to learn to become a semiologist, capable of inventing new meanings for dealing with reality and of externalising meanings through words or other representations, and so better understanding this process can help to enrich or give a theoretical foundation for teaching practices, particularly for first language teaching. The main message of this paper is that our vision of 'meaning creation' processes needs to shift from the traditional nativist or empiricist stance, which assumes a passive learner confronting in isolation reality, towards the view of an active being engaged in meaning construction in a social fashion.

It also becomes obvious why children (and adults) have such great difficulty to acquire a second language within a traditional school environment, whereas they learn it (seemingly) effortlessly in grounded social interaction with others. The 'school' style of learning language differs in three ways from the social learning of language modeled in our robotic experiments

1. Grounding: Traditional language teaching usually takes place in a de-contextualised setting, without any grounding in the real world. This makes it very difficult to form meanings, i.e. to map information to knowledge.

2. Mediation: When there is a single teacher in front of a classroom of 30 or more pupils, it is not possible to have the one-on-one interaction that is necessary for social learning.

3. Active learning: Our research shows that the learner must be able to take the initiative, testing out different hypotheses and filling in missing holes. In a classroom context it is very difficult to achieve this form of active learning.

This is not meant to be a criticism of teachers. It is remarkable what a teacher can achieve given the 'unnatural' circumstances of a classroom.

Learning Tools

The methodology of building artificial systems for investigating issues in the theory of learning has an additional benefit. It can potentially lead to new new educational technologies, through robots or artificial animated agents in virtual environments that are fun to interact with, but at the same time induce moments of social learning. Robots like the AIBO excite children enormously and induce them to various forms of social play (see also the next chapter by Kerstin Dautenhahn). This suggests that robots could potentially be a platform in which grounded, situated learning may be embedded, including activities in meaning construction. Robots could either play the role of mediator for helping to learn certain concepts and the language that goes with them, or the learner could play the role of mediator and thus be forced to reflect on their own concepts or language. Experiments have not yet been done in this direction, and they would in any case require many more advances in modeling learning processes on robots. But the potential seems undeniable.

Acknowledgements

This research was conducted at the Sony Computer Science Laboratory in Paris. The experiments with the AIBO were conducted in a collaboration with Frédéric Kaplan and Angus McIntyre. I am indebted to the members of the Bagnols workshop for much feedback on the issues discussed in this paper.

Bibliography

[1] Baron-Cohen, S. (1995) Mindblindness: An Essay on Autism and Theory of Mind. MIT Press.

[2] Mel, B. (1997) SEEMORE: Combining Color, Shape, and Texture Histogramming in a Neurally-Inspired Approach to Visual Object Recognition Neural Computation, 1997, 9 , 777-804

[3] Bowerman, M. and S. C. Levinson (2001) Language acquisition and conceptual development. Cambridge Univ. Press, Cambridge.

[4] Broeder, P. and J. Murre (2000) Models of Language Acquisition. Inductive and Deductive Approaches. Oxford University Press, Oxford.

[5] Bruner, J.S. (1990) Acts of Meaning. Harvard University Press, Cambridge Ma.

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

[7] Chomsky, N. and H. Lasnik (1993) The Theory of Principles and Parameters. In: J. Jacbos, A. von Stechow, W. Sternefeld and T. Vennemann (eds) Syntax: An International Handbook of Contemporary Research. Walter de Gruyter, Berlin. p. 506-569.

[8] Clark, E.V. (1987) The Principle of Contrast: A constraint on language acquisition. In: MacWhinney, B. (ed.) Mechanisms of Language Acquisition. L. Erlbaum Hillsdale NJ.

[9] Davidoff, J., I. Davies, J. Roberson (1999) Color categories in a stone-age tribe. Nature, vol 398. 230-231.

[10] Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48, 71-99.

[11] Elman, J., Bates, E., Johnson, M.H., Karmiloff-Smith, A., Parisi, D. and Plunkett, K. (1996). Rethinking Innateness: A connectionist perspective on development. Cambridge, Ma: MIT Press.

[12] Evans, N. (2000) Kinship verbs. p 103-172. in: In: Vogel, P.M. and B. Comrie (eds.) (2000) Approaches to the Typology of Word Classes. Mouton de Gruyter, Berlin.

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

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

[15] Golifkoff, R. (1983) (ed.) The Transition from Prelinguistic to Linguistic Communication. Lawrence Erlbaum Ass. Hilssdale NJ.

[16] Halliday, M.A.K. (1984) Learning How to Mean. Cambridge Univ. Press, Cambridge.

[17] Harnad, S. (1990) The Symbol Grounding Problem. Physica D 42: 335-346.

[18] Heine, B. (1997) Cognitive Foundations of Grammar. Oxford University Press, New York.

[19] Labov, W. (1994) Principles of Linguistic Change. Volume 1: Internal Factors. Basil Blackwell, Oxford.

[20] Langacker, R. (1987). Foundations of cognitive grammar, vol.1. Stanford: Stanford University Press.

[21] Lightfoot, D. (1991). How to set parameters. MIT Press, Cambridge Ma.

[22] Lightfoot, D. W. (1998). The Development of Language: Acquisition, Change, and Evolution. Blackwell, Oxford.

[23] Pinker, S. (1994) The Language Instinct. The New Science of Language and Mind. Penguin, Harmondsworth.

[24] Siskind, J. (2000) Visual Event Classification Through Force Dynamics. AAAI Conference 2000. AAAI Press, Anaheim Ca. pp. 159-155.

[25] Smith, L. (2001) How Domain-General Processes may create Domain-Specific Biases. In: Bowerman, M. and S. C. Levinson (2001) Language acquisition and conceptual development. Cambridge Univ. Press, Cambridge. p. 101-131

[26] Steels, L. (1996) Self-Organizing Vocabularies. In: Langton, C. and T. Shimohara (ed) (1997) Proceedings of the Artificial Life V. The MIT Press, Cambridge, Ma. pp. 179-184.

[27] Steels, L. (1997a) The Synthetic Modeling of Language Origins. Evolution of Communication Journal 1(1), 1-35. (1997)

[28] Steels, L. (1997b) Constructing and Sharing Perceptual Distinctions. In: van Someren, M. and G. Widmer (ed.) (1997) Proceedings of the European Conference on Machine Learning, Springer-Verlag, Berlin. pp. 4-13.

[29] Steels, L. (1998) The origins of syntax in visually grounded robotic agents. Artificial Intelligence 103 (1,2) p. 133-156.

[30] Steels, L. (2001) Language Games for Autonomous Robots. IEEE Intelligent systems, September/October 2001, p. 16-22.

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

[32] Steels, L. and R. Brooks (1995) The Artificial Life Route to Artificial Intelligence. Building Embodied, Situated Agents. Lawrence Erlbaum Ass, New Haven.

[33] Sutton, R. and A. Barto (1998) Reinforcement Learning. The MIT Press, Cambridge, Ma.

[34] Talmy, L. (2000) Toward a Cognitive Semantics: Concept Structuring Systems (Language, Speech, and Communication) The MIT Press, Cambridge, Ma.

[35] Tomasello, M. (1999). The Cultural Origins of Human Cognition. Harvard University Press, Cambridge, Ma.

[36] Traugott, E. and Heine, B. (1991) Approaches to Grammaticalization. Volume I and II. John Benjamins Publishing Company, Amsterdam, 1991.

[37] Wierzbicka, A. (1992) Semantics, Culture and Cognition. Oxford University Press, Oxford.

.

.

Group Discussion 7
The Social Learning of Language

The Problem of Grammar

Johnson: You showed us how robots could acquire basic perceptually grounded concepts and the words for them, but what about grammar?

Steels: Well I didn't have time to talk about grammar, but that is another direction in which I am working very intensely. We are setting up experiments in which robots play language games with each other and build up not only concepts and names for these concepts but also grammatical conventions for expressing things like predicate-argument structure, determination, time, aspect, modality, topic-comment structure, etc. I do not believe in the Chomskyan idea that there is a kind of abstract system that exists prior to the first data and in which parameters are then set based on linguistic evidence. Instead I believe that grammar is something that arises gradually, that is constructed by the child. Consider the distinction between nouns and verbs. Even assuming they were given innately, how would you know whether a word like "fiets" (bike) in Dutch was a noun or a verb, unless you already know the language? Moreover the language data is confusing because even though "fiets" is basically a noun, it can also be used as a verb (as in "ik fiets naar huis" (I bike home)). And how would you explain that there are languages such as Mundari, an Austro-Asiatic language, which do not make a distinction between nouns and verbs. Any lexicalised predicate can be used as a verb in the sense that it can be used as predication and takes tense and aspect markers, agreement, voice, etc., *and* as a noun, in which case it takes case markers and is used referentially. Words therefore denote both things and events. For example *lutur* means both ear and listen, and *kumRu* a thief and to steal.

So I believe that grammar, even the basic grammatical notions like nouns and verbs, are constructed by language users and that children have to perform this constructive effort themselves, using their generic cognitive apparatus. They have to use mechanisms like analogy, generalisation, etc. I don't believe there is a strong innate system that specifies that there is something like nouns and verbs in a language. I realise that we have an enormous amount of work ahead of us to actually prove all this through experiments of the sort I have shown you, but this is my plan anyway.

Rinaldi: The development of one language can be influenced by the development of another language, of the interaction between different languages. In the Reggio approach we talk about the hundred languages of children, which are all different representations they use to cope with reality. And we know from our practice that children 'discover' and construct these languages. Are you working on the interaction between different modes of representation?

Steels: This is a very interesting idea but we have not done anything on it. Your point is very well taken because if you look at the many languages of expression and the feedback

relationship between them, one could bootstrap the other, and in this way they could all gradually be constructed. This is a fantastic idea to expand our research program.

Rinaldi: We would be very grateful for a deeper understanding of all this in Reggio.

Nature versus Nurture

Ken Mogi: This whole question of nature versus nurture doesn't seem very productive to me. In this particular experiment that you did, the robot is programmed to play the game. This is in some sense genetically determined. Without this kind of determination, there would be no game.

Steels: Yes, of course. This is true. But the claim of nativists is much stronger than that. Consider colour. Psychologists like Sheppard would say that the focal points, the prototypes, for the basic colours like blue, green, red, etc., are innately given. Also for phonetics, they would say that phonetic categories, like whether a consonant is voiced (as /d/) or unvoiced (as /t/), which allow a distinction between different speech sounds, are innately given. The argument is not whether you need a genetic determinate structure of some sort, clearly you do, the argument is about "What are they?", "How language-specific are they? And then you have those taking a strong stance that concepts, grammatical conventions, etc., are innate. And I am saying: they are not.

Mogi: I'm not a linguist. But I see Chomsky and his colleagues as strawmen in research nothwithstanding that strawmen are fun, I think it is fair to say that genes and experience work together.

Steels: Of course. But you must agree that one's theory of learning is going to be very different if you assume that colour categories are innate, that they're fixed and universally shared, what I called the labelling theory, or if you assume that colour concepts are shaped by ecology, culture, and language. As long as we do not go into detail about what is exactly innate, we can argue until we fall asleep.

Mogi: My question is: can you state again what is innate in this system?

Steels: There are a number of generic abilities, e.g. the ability to categorize, to use an associative memory, to visually extract features from the environment, to share attention, etc. What is not innate is which categories there are, what words will be used, which word is linked to which meaning. It is like neural networks. They have the capability to arrive at some generalisation but you don't put in what the generalisation is going to be.

Punset: The most convincing argument of Steven Pinker's theory is that it is impossible for children to learn a language in less than three years, and it has convinced lots of people. Children at that age can only handle very simple calculation problems. But language seems extraordinarily complex and children can do it.

Steels: I believe that children take much more time than three years to learn a language! If you interact with three year olds you have the impression they master a lot of language but if you probe deeper you find they use a lot of canned phrases without really having meanings well established or without mastering the proper generalisations. For example, if you ask a four year old about the colours of their clothes they can answer perfectly because they have been playing games to name these colours. But if you ask colours of unknown objects they can't do it. Colour words, just to take this example, only stabilise around the age of eight or so. For grammar it is much more dramatic. Many children cannot produce well-formed complex sentences (with dependent clauses, etc.) until sixteen years. Many adults cannot properly write out their thoughts. I strongly resist the idea that language learning is finished at three years old. Instead I believe that we learn and invent language all our life

Figure 7.9: Group discussion. How far can we learn anything from building artificial systems for the practice of education? From left to right Caroline Nevejan, Bernard Allien, Luc Steels, and Marleen Wynants.

and consequently that teaching language, writing, etc., must be based on this assumption. We must teach children to be active participants in the language community instead of assuming that all they have to do is set a number of parameters in an innate language acquisition device.

Punset: But would a theory based on learning not take too much time?

Steels: No, we have done many simulations and things go often surprisingly fast. For example, in the domain of speech there are many constraints that naturally limit what behaviors the learner can acquire. For example constraints coming from the articulatory system (tongue, vocal chords, shape of mouth, etc.) restrict the kinds of sounds you can make. These constraints don't have to be innately given in your brain. They don't have to be learned. Some sounds simply cannot be made. There are many other constraints which push learners naturally towards certain types of solutions. The arguments usually given for innateness are arguments from ignorance, we don't understand how children can learn it so it must be innate. But if we work harder at it, I believe that we can understand the learning process.

Hedegaard : In learning language, we find that children have two strategies, one for learning names of objects and another one for learning names for action. Have you done anything on actions?

Steels: Yes, I did not talk about it but we have experiments on this. The basic idea is that the robot comes with a repertoire of behaviors organised in a network and tries different actions to find out which one is to be connected with a certain name.

Hedegaard: I don't mean learning the action. I mean naming the action.

Steels: Yes. But the learning of the action, and the conceptualisation and the naming of it, they all go together. Of course all this is very difficult, but we're trying to do it.

Collective Learning

Nevejan: Can you say more about your experiments in which groups of robots together developed a shared set of words and meanings, without human intervention?

Steels: Yes, this is our way to investigate the impact of culture on meaning creation and language. We created a set-up with a limited number of 'robotic bodies' able to make contact with the world through a pan-tilt camera, and a large open-ended growing set of

'agents', software entities, that could use these bodies to interact with each other about a shared situation in the world. The agents played a guessing game, another kind of language game, in which the speaker had to draw the attention of the hearer to an object before them by using words. In the experiment, the agents did not have any *a priori* concepts or any pre-given lexicon. They constructed new meanings to distinguish the objects in their environment, named these meaning, and above all negotiated shared meanings for the words. There is a lot to say about this experiment, which ran for several months and involved thousands of agents, but the key point is that it showed that language can emerge through a collective self-organising process when each of the participants in the community has the ability to construct their own meanings and words and at the same time adapt their constructions to that used by others.

Wynants: Collective learning, sharing a vision in order to learn and to develop together is fundamental to the development of cognition. Justine Cassell accentuates the importance of self-efficacious storytelling about the self, through collaboration with other children and through the development of real projects. It helps children see their own power and possibility, to establish their belief that they can have an effect on the world around them. Or in the words of psychologist John Dewey: "I think that every kind of learning takes place through the participation of the individual in the social consciousness of humankind."