In Wachsmuth, I., Lenzen, M., Knoblich, G. embodied communication in humans and machines. (2008) Oxford Publication

11

The emergence of embodied communication in artificial agents and humans

Bruno Galantucci and Luc Steels

11.1 Introduction

There has been a great deal of research on language, but usually it dissects an existing language and treats it as a static set of rules that is used more or less accurately and successfully to convey meaning. Here we are interested in the emergence of new communication systems and in their expansion and adaptation in usage. We seek a theory of the kind of cognitive mechanisms and interaction patterns necessary to bootstrap, maintain, and adapt a communication system that has similar properties as those found in human languages, such as those identified by Hockett (1960): discreteness, displacement, productivity, duality of patterning, etc.

The question of the emergence and continuous adaptation of communication systems is obviously relevant to the question of the origins of human languages, which has been lately at the center of increasing attention (e.g. Larson *et al.* 2007). In fact, we believe that these questions should receive attention from every student of language, for two reasons. The first one is that language, as any other complex social or biological phenomenon, cannot be thoroughly understood in the absence of a theory about its origins. The second reason is that research on language use (e.g. Clark 1996) has shown that even "mature" languages like English and the conceptual repertoires they employ undergo constant change. Language and conceptual systems are continuously adapting to cope with the problems of communicating novel meanings in novel settings, and these systems are coordinated between speakers and listeners via a multilevel process of alignment (Pickering and Garrod 2004). This dynamic nature of language suggests that linguistics (both descriptive and computational) should pay more attention to the processes that give rise to language and constantly reshape it, and that language should be viewed as a complex, adaptive system rather than as the static, formal calculus definable by generative grammars (Steels 2000; Tomasello 2005).

This paper focuses on models that attempt to capture this dynamic nature of language. In addition, it focuses on issues related to communication between embodied agents. Embodying communication entails that the individuals engaged in communication have a body with which they are present in the world, and that they can only communicate

through this body, as opposed to through some kind of direct or indirect way to transfer meaning.[1] This has three important consequences.

Embodiment implies no telepathy. The partners in communication have limited knowledge of what each of them perceives or knows, and they have no direct control over the internal states of others. Embodiment makes it therefore impossible to have telepathic transmission of thought, or to introduce some sort of global coordinating device that ensures that communication systems are shared among individuals. This raises the crucial question of how languages and conceptual inventories can nevertheless become sufficiently coordinated to make communication possible.

Embodiment implies different perspectives. Embodiment enables the partners in communication to independently gather information about the world, and to act on it. It is therefore a precondition to have grounded communication, that is communication about the real world perceived through a sensory–motor apparatus. But this implies that partners in communication will have a different sensory experience of a scene (for example they see things from different perspectives), that they may focus on different features of the world, or on different features of the signs used in communication, or that they may have a different repertoire of actions, etc. This raises the difficult question of how communication is possible without absolutely shared common ground (either for the content of sentences or for the conceptual system on which the sentences are based), and with all the uncertainty associated with noisy, real world action, and perception.

However, embodiment is not merely a source of problems. It is also a source of opportunities: The body (through gestures, sound productions, etc.) is the fundamental basis for constructing a communication system, partly because it allows the perception and reproduction of gestures that make up signs, but also because it can be used to set up frames of joint attention for example by eye gaze, pointing, etc. The body can also be the source of metaphors for conceptualizing the world, as in the case of temporal metaphors which are derived from spatial ones (Lakoff and Johnson 1999).

The main goal of this paper is to illustrate how embodying communication is both a source of difficulty for establishing a communication system and a source of opportunities.

11.2 Possible approaches for the study of the emergence and development of communication systems

The emergence and the development of communication systems have been studied via a number of different approaches.

Field studies. First, new languages occasionally develop, particularly in situations of social stress when individuals are brought together who have to communicate but do not share a common language. This has been the case for the emergence of creole languages in colonial times (Mufwene 2001), or the emergence of new vernaculars developing

[1] We do not intend this as an exhaustive notion of embodiment. For other important aspects of embodiment, see the contributions of Barresi, Proust and Kopp *et al.* in this volume.

today in the inner cities of Europe as a consequence of intense migration, such as "Verlan" a French vernacular that originated in French suburbs based on word play and Arabic influences (Lefkowitz 1991). Another example is the emergence of new sign languages, such as the Nicaraguan sign language (Kegl 1994) or, at a more individual level, the sign systems that are developed between deaf children and their hearing parents (Goldin-Meadow 2003). The data that are obtained in these natural experiments are highly valuable for many reasons, including that they tell us a lot about the dynamic aspects of natural communication systems. However, these data are not obtained in strictly controlled experiments and are therefore difficult to use as a foundation for scientific theories of the emergence of communication (but see Goldin-Meadow *et al.* 1996; Hudson Kam and Newport 2005, for interesting exceptions).

Experiments on human communication. In the last 40 years, students of language pioneered a second approach based on the experimental study of natural dialogue (Clark and Wilkes-Gibbs 1986; Garrod and Anderson 1987; Krauss and Weinheimer 1964). These students created challenging communication tasks for pairs of participants (for example the joint traversal of a maze) and carefully recorded the verbal interactions that took place in the dialogues of the participants that performed the tasks. Three important findings came out of this research. (a) Even though these researchers looked at an existing communication system (natural language), they observed that partners in dialogue occasionally introduce significant innovations. These innovations concern all levels of language: new ways of conceptualizing the situation (e.g. Garrod and Anderson 1987), new meanings for existing words (e.g. Krauss and Weinheimer 1964), and extensions of existing grammatical constructions (Traugott and Heine 1991). (b) Dialogue partners align their verbal behavior at all levels (Garrod and Pickering 2004). Their speech sounds and body gestures become similar (for the latter see, e.g. Kimbara 2006; for the former see, e.g. Pardo 2006). They quickly adopt words and word meanings used by others (e.g. Brennan and Clark 1996). They tend to echo the same grammatical constructions (Pickering and Branigan 1999). (c) There is often remarkable variation in how different pairs tackle the same task (e.g. Garrod and Anderson 1987). But when pairs are selected consecutively from the same group, alignment leads to "sublanguages", with much more sharing and therefore higher communicative success among the group members than across groups (Garrod and Doherty 1994).

More recently, this paradigm was extended further. Healey and coworkers introduced a graphical medium for communication, with essentially the same results (Healey *et al.* 2002, 2007). They asked participants to graphically describe a piece of music so that their partners could decide whether they were listening to the same or to a different piece. A graphical medium brings us closer to the emergence of a new communication system, because it is less constrained by prior conventions and so the degree of innovation is higher. Besides innovation, Healey and colleagues observed again alignment, variation, and the formation of shared subsystems in groups.

In the same line of research, one of us (Galantucci 2005; see also Stephan, this volume) designed a method in which participants play a videogame in which they can succeed

only when they communicate effectively. The game world consists of a set of rooms located on a grid and marked with icons. Players have to move to the same room with the minimum number of room changes, but they only have a local view and they cannot see where the other player is located. As they need to know this in order to decide their next moves, players are encouraged to develop ways for describing their own positions, where they intend to move next, or what they suggest the other player should do. A key element of the method is the introduction of an unusual graphical medium by which players can communicate. Each player is provided with a digital scratchpad that moves vertically as one draws on it, so that drawing a horizontal line results in a diagonal line with a slant that reflects the velocity profile of the drawing motion.

Because of this novel medium, players are forced to totally invent new forms for communicating. There is no prior inventory, not even a prior set of signs to build from. Nonetheless, most pairs of players manage to get a communication system started and we observe even more sharply the same findings as seen in natural dialogue: innovation, alignment, and variation. We also observe how emergent communication systems are tightly embedded within behavioral procedures that coordinate the actions of the partners. Because successful pairs are then faced with new challenges by increasing the number of rooms and by introducing additional tasks, the method allows us to study the further evolution of communication systems once they have emerged, showing that communication systems continue to be adapted by players while retaining the earlier solutions as much as possible.

However, not all pairs in the study manage to bootstrap a communication system. Besides obvious requirements such as pattern-recognition abilities, memory, enculturation, etc., the challenge seems to require a cooperative attitude, a particular type of social intelligence. Some players behave like Humpty Dumpty. They just assume that others see the world in their way and use symbols the way they decide. They fail to realize that their communication is ambiguous and do not have the social inclination to negotiate repairs. Frustration can run very high. A task that some pairs manage in 10 minutes, takes others 3 hours before they give up. We will give some detailed examples of this later in the paper.

Experiments on artificial communication. There is yet a third approach to study the emergence of a communication system *de novo*, which is to engage in experiments with artificial agents that are "language-ready" in the sense that they have all the cognitive machinery for expressing meaning in language utterances, and for parsing utterances back into meaning. The agents also have interaction scripts to play specific language games, such as drawing attention to an object in the shared situation. The agents come with a battery of strategies for repairing a failed communication and for consolidating their inventories based on the outcome of a game. What the agents do not have is a communication system, in the sense of a set of conventions relating meaning with form. In this respect, they are in the same boat as the agents in the experiments by Galantucci. It is possible to change systematically the learning mechanisms or the parsing and production mechanisms in the agents and thus study their effectiveness for bootstrapping a communication system. In this way, we can establish experimentally which mechanisms are required for successful communication to emerge and which mechanisms are better

adapted for the task. Experiments of this type started in the 1990s (Steels 1997) and have since flourished (Briscoe 2002; Cangelosi and Parisi 2002; Steels and Belpaeme 2005; Minett and Wang 2005; Wagner *et al.* 2003). This research does not tell us how the mechanisms for bootstrapping communication systems might have evolved or how they develop in the child, but rather what mechanisms are necessary and sufficient.

Many of these experiments focus exclusively on the mapping from meaning to form and from form to meaning, but some researchers have extended the methodology to consider the whole system involved in successful communication: from perception to language and from language to real world action in embodied agents, that is agents which have a physical body and the necessary sensors and actuators to engage with the real world (Steels 2003). Of course, this makes the problem much more complicated, partly because the robotic agents have to develop not only the language system, but also the conceptual system that they use to structure the sensory experiences that their language will express. However, in order to capture the essence of communication in a real world, and address properly the issues that arise in the embodiment of communication, the move from computational simulations to robotic experiments is unavoidable. This move forces us to no longer consider language as an abstract symbolic system but as a system for grounded communication. The communication is grounded in the sense that it is about the world as experienced by the embodiment of the agents.

This paper compares this work on the robotic modeling of emergent embodied communication, specifically the work of Luc Steels and his team (Steels 2003; Steels *et al.* 2002; Steels and Loetzsch 2007) with the empirical data coming from experiments with human subjects, specifically the studies conducted by Galantucci and his colleagues (Galantucci 2005; Galantucci *et al.* 2003; Galantucci *et al.* 2006). We begin by looking at the computational and robotic experiments and then examine how far the major conclusions of these experiments carry over to the human experiments.

11.3 Robotic experiments on emergent communication

Here we present a typical example experiment which illustrates the state of the art in orchestrating emergent communication in artificial robotic agents (for more details, see Steels and Loetzsch 2007). The experiment explores how a population of autonomous robots can develop a spatial lexicon for communicating about aspects of the situation in which they find themselves. Specifically, it shows how spatial categories, such as "left" versus "right" or "far" versus "near", can emerge as lexical concepts and become coordinated through repeated use. It shows furthermore how perspective reversals—that is the fact that an agent takes into account the different perspective of the partner—can be recruited for enhancing communication effectiveness and correspondingly marked in the language (as in "to my left" versus "to your right"). Finally, it shows how some forms of grammar can arise.

The perspective reversal experiment. The experiment uses physical robotic "agents" (the Sony ERS7 AIBO), which roam around freely in an unconstrained in-door environment containing balls and boxes (Figure 11.1). The robots have no direct way of communicating

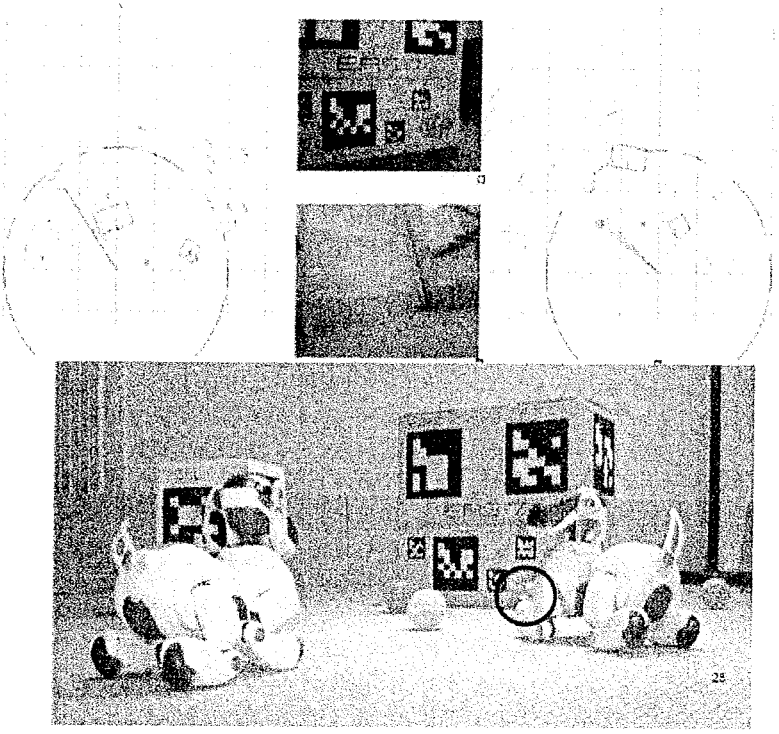

Figure 11.1 Experimental set-up for the perspective reversal experiment which features balls and boxes and two AIBO robots. The speaker (robot A) and the hearer (robot B) focus on the ball and track its movement. The bottom pane shows the ongoing interaction between the robots. The top left pane shows parts of the world model to the right. The trajectory of the ball is marked by an empty circle to a full circle and the position and orientation of speaker and hearer is shown by the arrows. The boxes are shown with rectangles.

except through visual or auditory means and they have no way to read or set each others' internal states. Even though the experimenters can track the complete internal state of each robot based on wireless communication between the robot and a base station, there is no central control, neither of the physical behavior nor of the cognitive operations that a robot performs. The robots are completely autonomous. In other words, once the experiment starts, the situation becomes similar to observational experiments with animals. Moreover, although the experiment employs only two robot bodies, it is relatively straightforward (and is now routinely done) to carry out experiments with a much larger population of agents: The state of an agent (its perceptual, conceptual, and linguistic inventory at a particular point in time) is after all a software state, and so it can be "downloaded" into a specific robot body before interaction starts and "uploaded" to another robot body at the end of an interaction. So we can have as many agents as we want even with a small number of robots.

The robotic agents engage in language games. A language game is an interaction between two agents which has a particular communicative goal, such as drawing attention to an object in the world, describing a situation, or requesting an action, and is situated in a common physical setting so that the participants share to some extent their experience of the environment. By design, the game is sufficiently constrained so that agents share the communicative goal, can establish joint attention independently of language, and are able to provide feedback on success or failure. These constraints are implemented by constraining the environment (for example there is only one orange ball and it is the focus of attention) and by programming quite specific behavioral interaction scripts in the agents. We do not consider here the problem of how agents could negotiate their communicative goals.

The language game used in the perspective reversal experiment is a description game. The speaker describes to the hearer what is novel about the present scene compared to the previous one. It works in the following manner. Two robots walk around randomly. As soon as one detects the ball, it comes to a stop and searches for the other robot, which also looks for the ball and stops when it sees it. Then the human experimenter pushes the ball with a stick so that it rolls a short distance, for example from the left of one robot to its right. This movement is tracked and analyzed by both robots and each uses the resulting perception as the basis for playing a language game, in which one of the two (acting as the "speaker") describes the ball-moving event to the other (the "hearer"). To do this, the speaker must first conceptualize the event in terms of categories like "left" and "right" that distinguish the latest event from the previous one, for example, that the ball rolled "away from the speaker and to the right", as opposed to "towards the speaker", or, "away from the speaker but to the left" as opposed to "away from the speaker but to the right". The available categories are perceptually grounded in the sense that they are processes operating over the sensory data. They are built up by the agents stimulated by the need to conceptualize a scene and aligned based on the outcome of a language game. Agents can perform perspective reversal, in the sense that they can geometrically transform their own image of the scene to compute what the scene looks like from the perspective of the hearer. They can then apply their perceptual categories to this transformed image instead of their own.

Next the speaker expresses this conceptualization using whatever linguistic resources in its inventory express it best and have been most successful in the past, and transmits the resulting utterance as an acoustic signal to the hearer. The hearer parses the utterance to reconstruct its possible meanings and applies them to the current scene.

The game is a success if, according to the hearer, one of the meanings not only fits with the current scene as it is perceived by him but is also distinctive with respect to the previous scene. For example, if a ball was to the left of the box in the previous scene and in the current scene, then a description "the ball is to the left of the box" is not considered to be appropriate even though it fits with the scene, because it does not describe a novel property of the current scene. The hearer then signals success or failure and both agents use this feedback to update their internal states. Note that there is no human intervention involved. The robot agent playing the role of hearer autonomously decides whether the game was a success or not.

The main point of the experiment is that neither a prior language nor a prior set of perceptually grounded categories (properties, relations, prototypes, etc.) are programmed into the agents. Indeed, the purpose is that of seeing what kinds of categories and linguistic constructions will emerge and, more specifically, whether they involve perspective marking and grammatical constructions to express them. Agents therefore need their cognitive machinery not only for playing the game and utilizing their available conceptual and linguistic inventories, but also for expanding these inventories by creating (as speaker) or adopting (as hearer) new categories, new words, and new grammatical constructions as the need for them arises. Agents take turns playing speaker and hearer, so that they each develop the competence to speak as well as that to understand, and all of them have equal rights to invent new bits of language or decide whether to re-use constructions introduced by somebody else.

Figure 11.2 shows the results of the experiments. The left panel shows an experiment without perspective reversal and the right panel with perspective reversal. We see that in the experiment illustrated in the left panel, agents are not successful. This is entirely due to the fact that they are embodied. If they would have exactly the same sensory experience of the world, for example if they would share the same bird's eye view of the world, they would not need perspective reversal. In the experiment presented on the right panel, successful communication systems invariably emerge and these systems exhibit many of the properties of human languages as identified by Hockett (1960), including: "arbitrariness": there is no specific reason why something is called in a particular way except convention; "productivity": the capacity to say or understand things that have never said before; and "displacement": because they implicitly talk about what is novel with respect to a situation which is no longer before them "here and now". Successful communication here means that the agents consistently agree that the descriptions they communicate to each other describe a novel aspect of the situations they perceive. This is only possible when

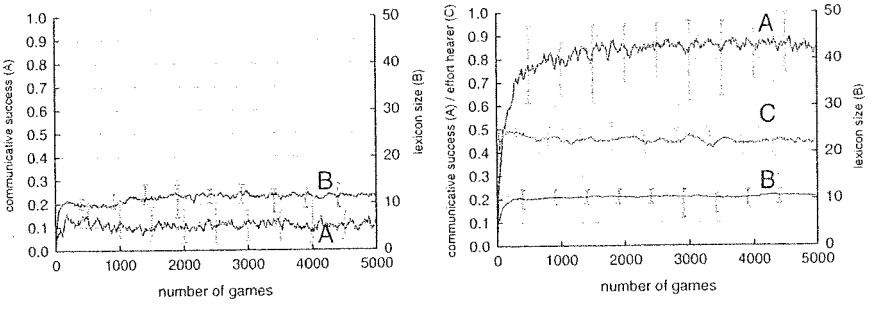

Figure 11.2 Results from five experimental runs of 5000 language games in a population of 10 embodied agents. A is communicative success and B the size of the lexicon. Left: Robots are unable to perform perspective reversal and their communication system does not get off the ground. Right: Robots have recruited the egocentric perspective transformation into their language faculty. Success is now close to 90 % and the lexicon is stable. Cognitive effort (C) is quite high and can be diminished by grammatically marking the perspective transform.

their conceptual and lexical inventories have become sufficiently coordinated. Figure 11.2 (right) plots communicative success, lexical inventory size, and cognitive effort characterized as the additional processing necessary for perspective reversal for five successful experimental runs of 5000 language games in a population of ten physically embodied agents. The agents start initially with empty conceptual and linguistic inventories but reach a high level of communicative success (close to 90%) despite the severe challenges posed by real world interactions, perceptual processes, and embodied communication.

Cognitive mechanisms used by agents. What kind of theory could account for the results presented in the preceding section? There are basically three main theoretical approaches that are discussed in the literature for explaining how communication systems may originate. These approaches differ depending on their choice for a driving mechanism: One approach focuses on genetic evolution, the second on intergeneration cultural evolution, and the third on intrageneration problem solving.

The theory based on genetic evolution, defended for example by Pinker and Jackendoff (2005) or Bickerton (1984), argues that humans are equipped with a special neural circuitry for language. Consequently, when they have to establish a new communication system (for example, as in the case of Nicaraguan sign language) they apply their innate predispositions and arrive automatically at language systems and conceptual systems that are largely determined *a priori*, reproducing the universal characteristics of human languages. We can apply this idea to the artificial agents, by endowing each of them with an artificial genome that lays down in great detail the circuitry with which they can communicate, in other words what perceptual primitives they use for segmenting the world and identifying objects and features, what concepts they have for structuring their world, what words they can use to express these concepts, what types of grammatical constructions they can employ, etc. Innovation takes place when the genome is transmitted from parents to children through copying, crossover, and partial mutation. Natural selection, operationalized as success in communication, acts then as a way to ensure that similar genes appear in the population and that the genes which lead to the most effective communication system survive. Such experiments in artificial genetic evolution of language are possible and have indeed been carried out (see, for example, Briscoe 2000; Cangelosi and Parisi 1998).

A second approach is based on models of intergeneration cultural evolution, such as those of Boyd and Richerson (1985). These models are similar to genetic models in the sense that innovation takes place in the transmission from one generation to the next, but now the language and conceptual system is considered to be transmitted culturally instead of genetically. Children learn the language from their parents and then use it, largely unchanged, throughout the rest of their life. The learning process introduces generalizations and variations because children are never exposed to the full set of possible linguistic material and hence innovation may take place as children acquire language. This innovation appears in the linguistic material they generate for the next generation. This framework has not only been applied to the study of the emergence of human languages (specifically the Nicaraguan case with different cohorts assumed to be responsible for progressively pushing the language to a grammatical streamlined system (Polich 2005)) but has been also used in experiments with artificial agents (Kirby and Hurford 2002).

In particular, it has been demonstrated that the learning bottleneck (which implies that children experience only part of the linguistic data that their parents can produce) may induce compositionality (Kirby 2000).

A third approach views the task of building up a communication system as a kind of problem-solving process within a generation of users. This process is conceived of as an intuitive process that is often inaccessible to conscious inspection, rather than as a rational, conscious problem-solving process (like the one a computer designer, for example, engages in). Moreover, this process is not conceived of as an individualistic problem-solving process, but rather as a collective effort in which different individuals participate in a peer to peer fashion. According to this view (developed for example by Tomasello 1999) a communication system is built up in a step by step fashion driven by the needs for coordinated interactions. It employs a large battery of strategies and cognitive mechanisms which are not specific to language but appear in many other kinds of cognitive tasks, such as tool design or tool use (Hutchins 1995). When children or second language learners acquire the communication system already well established in a language community, they reconstruct in a step-wise fashion that communication system, by successive inventions and adjustments, but obviously their own inventions have almost no chance to be accepted by the rest of the community (Mufwene 2001).

There is a second dimension to this problem-solving approach, namely that if different individuals each invent their own communication system, a competition arises in the population as a whole among concepts, words, and grammatical constructions. So language emerges as a complex adaptive system like a biological ecosystem or an economy. The selection process does not take place at the level of genetics or intergeneration cultural transmission but at the level of language itself (Croft 2000; Mufwene 2001). This collective dynamics is similar to the models and processes studied in opinion dynamics or collective economic decision making (Axelrod 2005).

There are many levels of competition: between perceptual categories; between synonyms for becoming dominant in expressing a particular meaning; between holistic and compositional expressions of combinations of meanings; between idiomatic patterns that group a number of words and the words themselves, which may still occur as individual units; between different syntactic and semantic categories that are competing for a role in the emergent grammar; etc. Often there is no particular reason why one solution would be preferred over another one, except that it is more frequent in the population.

The problem-solving approach and complex adaptive systems (CAS) view of language underlies the perspective reversal experiment and other experiments on embodied communication carried out by Luc Steels and his team (such as the "Talking Heads Experiment" (Steels 2003)), or the experiments on the coevolution of color terms and color categories (Steels and Belpaeme 2005). The results of these experiments converge with the results of a fast-growing body of mathematical work that provides formal mechanisms to explain the emergence of coherence in multiagent systems and that examines the impact of system size or network structure (Baronchelli *et al.* 2005, 2007). In contrast to the genetic evolution models, there is no genetic coding of the specific conceptual or language inventories used by the agents and hence no genetic transmission nor natural

selection based on communicative success. In contrast to the intergeneration cultural evolution models, generational change is not considered a necessary condition for explaining the origins of linguistic structure. Although the composition of the population of agents may change, each agent is both speaker and listener and therefore teacher or learner, depending on context. Anyone can at any time change aspects of language, similarly to what is observed in the empirical research on natural dialogue.

In the problem solving/CAS approach, individuals are considered to be endowed with a series of strategies to cope with the tasks and the problems they encounter. In the present experiment, this is directly implemented in artificial agents, leaving the problem open as to how these strategies might be recruited (Steels and Wellens 2006). Agents are endowed with a number of strategies which operationalize aspects of the problem solving that they need to engage in.

First of all, there are strategies for setting up a situation in which negotiations can take place to establish a communication system. Specifically the robotic agents are programmed to have ways for setting up a frame of joint attention with enough common ground and shared knowledge to enable them to guess the meanings that might be expressed by unknown words or constructions. In more sophisticated experiments with humanoid robots, this is achieved with pointing gestures, eye gaze following, movement towards objects that are going to be the subject of the conversation, etc.

Second, there are strategies for detecting that something is going wrong in the communication, and for finding out the exact cause of the problem. The main feedback signal is of course that the communication does not achieve its desired effect. But agents need also more fine-grained analysis to diagnose what went wrong. For example, a word may have been misunderstood, a perceptual category used by the speaker may have been broader or more restricted compared to that of the hearer, the speaker may have adopted another perspective on the scene than the hearer without signaling this explicitly, etc.

Third, there are strategies for fixing a problem. For example, agents may introduce a new word or change the definition of a word they have in their lexicon, they may shift a perceptual category to slightly align it with the way that category is used by the speaker, they may start to mark perspective explicitly, or they may introduce more syntax to curtail combinatorial explosions in the search space or ambiguities in semantic interpretations. One of the main points of the perspective reversal experiment is that one strategy for fixing problems due to embodiment is to introduce a way to shift perspective and mark this explicitly.

This experiment, along with more and more sophisticated robotic experiments that are currently being carried out (Steels and Wellens 2006), demonstrates that the problem solving/CAS approach is a source of valuable insights. Although there are still a large number of unsolved problems, both in identifying and operationalizing problem-solving strategies and in understanding the competition dynamics, there is now enough evidence to consider this approach as a viable option to understand how communication systems may emerge. The skeptic will argue that all of this may be true for robotic agents, but does not necessarily hold for humans bootstrapping an embodied communication system, and so the question addressed in the rest of the paper is whether there is empirical evidence in the data coming from experiments with emergent communication

in humans showing whether humans have similar strategies for initiating and repairing communicative failures. In the next section we provide evidence that this is indeed the case.

11.4 Human experiments on emergent communication

This section presents data collected with the method developed by Galantucci (2004). The key elements of the method have been introduced in Section 11.1. Here we present a slightly more detailed description of the method in order to provide a context for the interpretation of the data that will be discussed later in the section (for a more detailed exposition of the method, see Galantucci 2005).

The basic idea behind the method is that of creating a context within which two adults need to communicate, but cannot use a pre-established way to do so. A simple implementation of the idea is Game 1.

Game 1 set-up. Two adults participate in a real time videogame with interconnected computers located at different locations. Each player controls the movements of an agent in a shared virtual environment composed of four intercommunicating rooms (Figure 11.3A). However, players do not see the full environment but only the room in which their agent is currently located (Figure 11.3B). That is, players do not share a bird's eye view of the overall environment of the game but must rely only on individual local

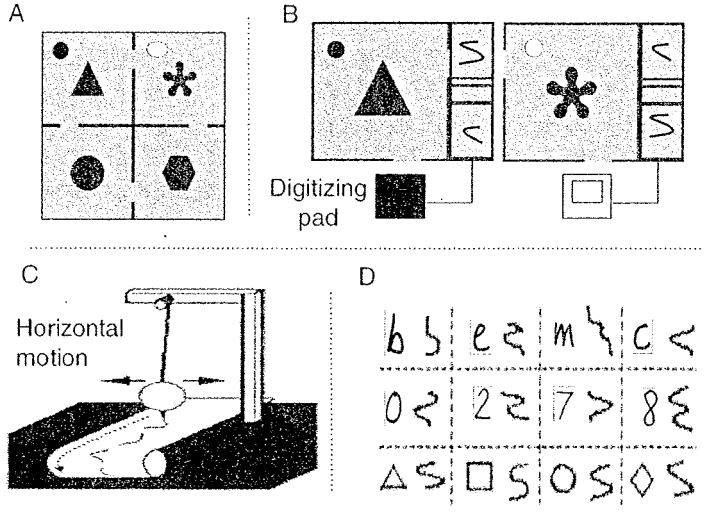

Figure 11.3 Method. (A) Game 1 map. The agents are represented by the blue dot and the white dot. Each room is marked by an icon, the location of which does not change over the course of the game. (B) Game set up. Players' individual views of the game environment and of the communication medium. (C) The graphic signal was similar to the output of a seismograph but quickly faded and allowed discontinuities. (D) How common graphic symbols looked on the screen when traced via the communication medium.

perspectives, exactly as in the robotic experiments presented in the previous section. Moreover, players are not told what the layout of the game environment is. In consequence, before players can successfully communicate about the environment, they must acquire some sharable understanding of its layout.

Task. Players engage in a cooperative game. At the beginning of each round of the game, the agents are located in two different rooms at random, and the players' goal is to bring the agents into the same room without making more than a single room change per agent. Chance-level performance in the game is 50% and can be improved only if information about location and intended movement of the agents is communicated. Once this occurs, however, the game reduces to an easy 100% win.

Communication medium. Players cannot see or hear each other but can communicate by using a magnetic stylus on a small digitizing pad. The horizontal component of the stylus' movements on the pad directly controls the horizontal movements of a trace that is relayed to the screens of both players (Figure 11.3B). The trace's vertical component is independent from the player's movements and has a constant downward drift which causes the tracings to disappear from the screen quickly (Figure 11.3C). Under these conditions, the use of common symbols such as letters or numerals is practically impossible and the use of pictorial representations is severely reduced (Figure 11.3D); players must converge onto a non-obvious way of using the graphic medium in order to set up a communication system extemporaneously. Moreover, since the communication medium can be used simultaneously by both players throughout the entire duration of the experiment, players have to set up procedures to coherently organize their signaling activity.

Procedure. Thirty pairs of participants were recruited to play Game 1. Before playing the game, players were briefly instructed and informed that their partners received the same instructions. During the game, players were encouraged to focus on the score as their primary goal. (The score consisted of a numerical index that increased only when the pair won consistently in the game.) Upon reaching a score that reflected performance above chance, each player explained in detail to the experimenter the communication system developed by the pair and described how the system was used to solve different scenarios of the game.

Game dynamics. During the game, there were three distinct kinds of interactions in which players could exchange signals[2] in order to set up a communication system. The first one occurred when a round of the game was ongoing. Players were always in different rooms, and their views of the task environment had no overlap. In this context (which we will refer to as *online disjointed view interaction*), hypotheses about the meaning of players' signals had to be tested by trial-and-error, keeping track of the

[2] Throughout the section, we will distinguish between *signals*, that is the perceivable products of the physical activity on the digitizing pad, and *signs*, that is the meaningful units of functional communication systems.

successes and the failures at achieving the goal to find each other. The other two kinds of interaction occurred when a round was over. At that stage, agents could no longer leave their rooms, until both players decided to terminate the round by moving the agents into one of four marked locations in the room (henceforth reset zone). As soon as both agents entered a reset zone a new round of the game resumed; agents were instantly relocated in two different rooms at random and players returned to an online, disjointed view interaction. In other words, at the end of each round players gained control of the pace of the game and could decide to interact in absence of a direct pressure to win a round of the game. These interactions (which we will refer to as *offline interactions*) differed depending on whether or not the pair won or lost the round. If the pair won the round, players completely shared their views of the task environment and could see each other's agents, which were in the same room. In this context (which we will refer to as *offline same view interaction*), hypotheses about the meaning of players' signals could be tested through the parallel communication channel provided by the movements of the agents in the room. (Not only players saw each other's agents location in the room but also their orientation in the game environment, given that the agents had human-like animated bodies). These movements were publicly visible and could be used to ground the meaning of the signals. This option was never available during online interactions because players were always in different rooms. If the pair lost the round, there was no overlap between the players' views of the task environment and players could not see each other's agents, which were in different rooms. In this context (which we will refer to as *offline disjointed view interaction*), hypotheses about the meaning of players' signals could not be tested at all.

Results. The data collected with the 30 pairs are described and analyzed in details in Galantucci (2005) and Galantucci *et al.* (2006). Here we present the results from a general point of view, focusing on the overall successes and the failures of the pairs at the game. Then, we will analyze in detail two specific examples of a very successful pair and a very unsuccessful pair.

Eighteen pairs of the 30 pairs that played Game 1 attempted to establish their communication systems primarily via online disjointed view interactions.[3] We will not consider these pairs here.

Twelve pairs of the 30 pairs that played Game 1 attempted to establish their communication systems primarily via offline interactions.[4] None of these pairs failed at establishing a communication system for Game 1. Four of the 12 pairs performed very well at Game 1 and reached the last stage of the game (a version of Game 1 played on a 4×4 grid) within 6 hours of playing. Three of the 12 pairs performed very poorly at Game 1 and, in spite of extensive playing, never went beyond the next stage in

[3] Two of the 18 pairs failed at establishing a communication system within the first 4 hours of playing.

[4] To a lesser or greater extent, most of these pairs relied also on online disjointed interactions.

the game (a version of Game 1 played on a 2×3 or 3×3 grid). The remaining pairs were in between these extremes.

In what follows, we contrast two pairs that come from the two extreme groups of pairs. One of them, Pair A, comes from the most successful group. The other, Pair B, comes from the least successful group. In other words, we contrast a pair that greatly benefited from offline interactions with a pair that benefited much less from them. The contrast will provide information about the mechanisms through which players established effective communication procedures. We will analyze these mechanisms with three goals in mind. The first one is that of identifying whether humans indeed engage in problem solving and what problem-solving strategies they appear to use. The second goal is that of understanding which behaviors facilitate the establishment of the frame of joint attention that is necessary for benefiting from the opportunities offered by offline same view interactions. The third goal is that of understanding in which way establishing a frame of joint attention helps players in solving the problem of communicating.

Pair A. Pair A was one of the most successful pairs of the 30 pairs that played Game 1 and smoothly continued the game until completion of the last stage of the game. The pair solved Game 1 in 20 minutes of playing. During this time, the pair played 33 rounds, losing only six of them (18%), four of which came during the first six rounds of the game. The two top plots of Figure 11.4 illustrate the success of Pair A in Game 1. The first plot (Figure 11.4A) illustrates the steady rise in score over time. The second plot (Figure 11.4B) illustrates the steady decline in the time that it took players to make the first move in a round, an indication that players were ever more confident in how to play the game. Their confidence was justified, as indicated by the steady raise of the score. Pair A's communication system comprised four signs, one for each of the rooms in the task environment (Figure 11.4C). The signs were fairly concise: For 72% of the whole Game 1 time, none of the two players used the digitizing pad. The bulk of Pair A's communication system was established early on in Game 1, during 10 crucial rounds. As detailed below, during these rounds, the pair extensively exploited the opportunities offered by offline same view interactions. At the same time, as illustrated in Figure 11.4D, Pair A never pursued offline disjointed view interactions.

- *Round 1.* At the beginning of the first round, both players drew repeatedly on the digitizing pad the icons they saw on the floor (for a graphical synopsis of the rounds, see Figure 11.5). The White player (henceforth W; see Figure 11.6 for a complete list of abbreviations) drew a circle; the Blue player (henceforth B) drew a flower-like shape. However, after changing room, the White player produced six dots, probably to indicate the hexagon that was on the floor of the new room (the six dots were repeated once). The round was lost and there was no offline interaction.

  This round highlights how W quickly changed his signaling strategy from drawing shapes to counting vertices. W never drew shapes again.

- *Round 2.* In absence of any sign exchange, B moved to the room with a circle on the floor (henceforth CR) and, fortuitously, found the partner there.

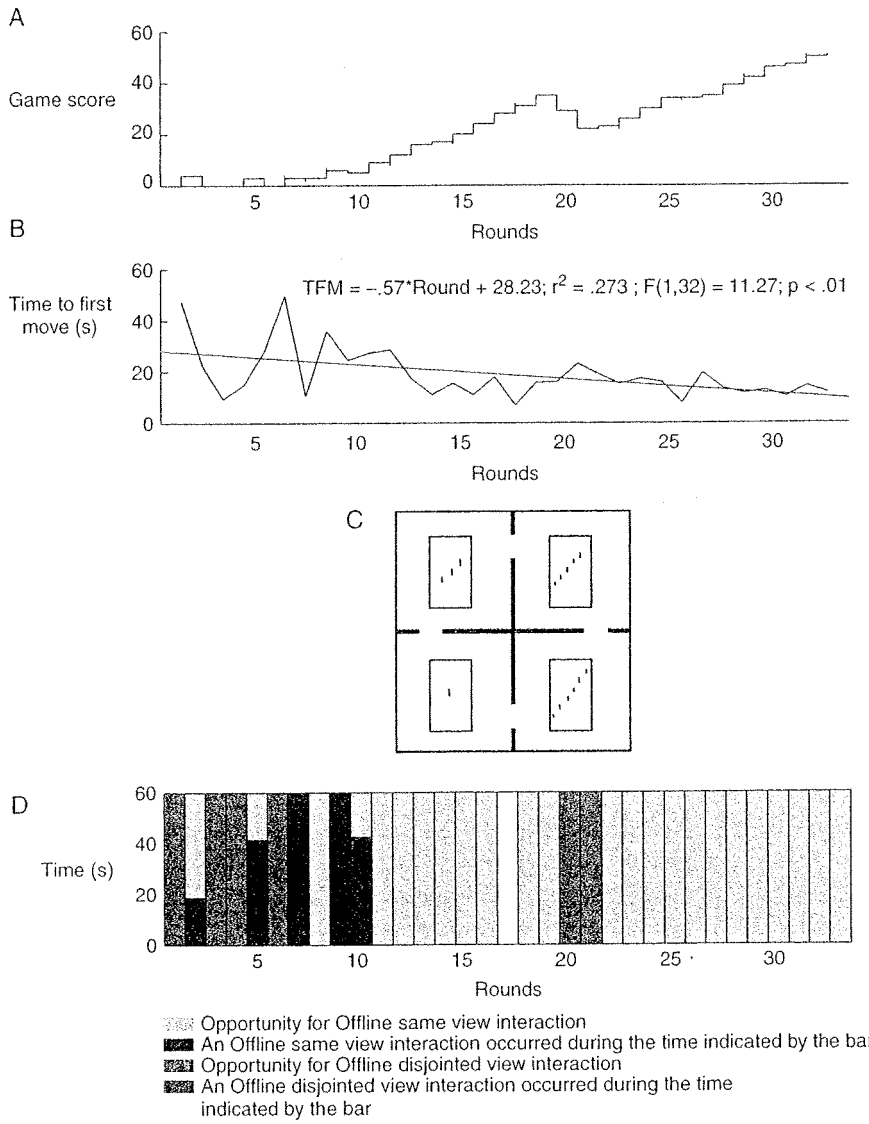

Figure 11.4 Pair A's basics. (A) Score during the first 33 rounds. (B) Time it took players to make the first move over the first 33 rounds. (C) Sign system developed by Pair A to solve Game 1. (D) Time spent in offline interactions over the first 33 rounds.

There was an offline interaction, during which a first crucial event occurred. While B was moving toward a reset zone—getting further away from W—W produced a dot on the digitizing pad. B immediately stopped the movement of the agent and then backtracked briefly toward W. W moved near the door that would lead to the room with a triangle on the floor (henceforth TR) and produced two more dots. Only when W had finished signaling and moved away from the door, B resumed the movement toward the reset zone.

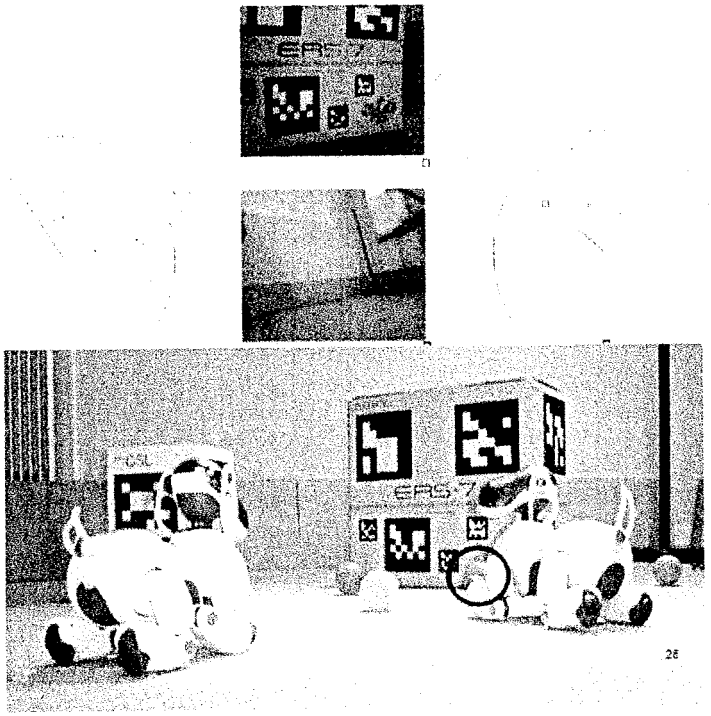

Plate 1 Experimental set-up for the perspective reversal experiment which features balls and boxes and two AIBO robots. The speaker (robot A) and the hearer (robot B) focus on the ball and track its movement. The bottom pane shows the ongoing interaction between the robots. The top left pane shows parts of the world model to the right. The trajectory of the ball is marked by an empty circle to a full circle and the position and orientation of speaker and hearer is shown by the arrows. The boxes are shown with rectangles.

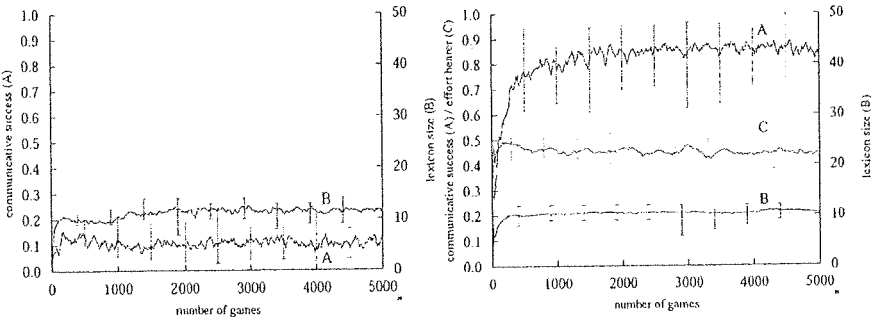

Plate 2 Results from five experimental runs of 5000 language games in a population of 10 embodied agents. A is communicative success and B the size of the lexicon. Left: Robots are unable to perform perspective reversal and their communication system does not get off the ground. Right: Robots have recruited the egocentric perspective transformation into their language faculty. Success is now close to 90 % and the lexicon is stable. Cognitive effort (C) is quite high and can be diminished by grammatically marking the perspective transform.

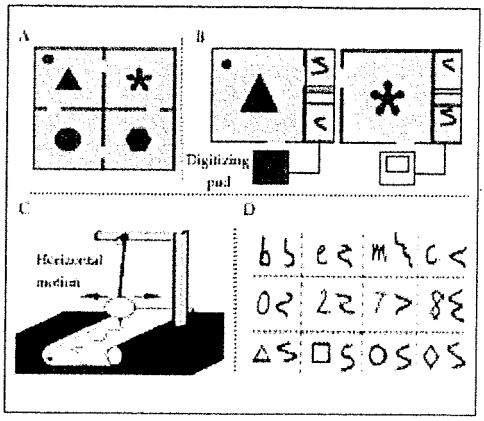

Plate 3 Method. (A) Game 1 map. The agents are represented by the blue dot and the white dot. Each room is marked by an icon, the location of which does not change over the course of the game. (B) Game set up. Players' individual views of the game environment and of the communication medium. (C) The graphic signal was similar to the output of a seismograph but quickly faded and allowed discontinuities. (D) How common graphic symbols looked on the screen when traced via the communication medium.

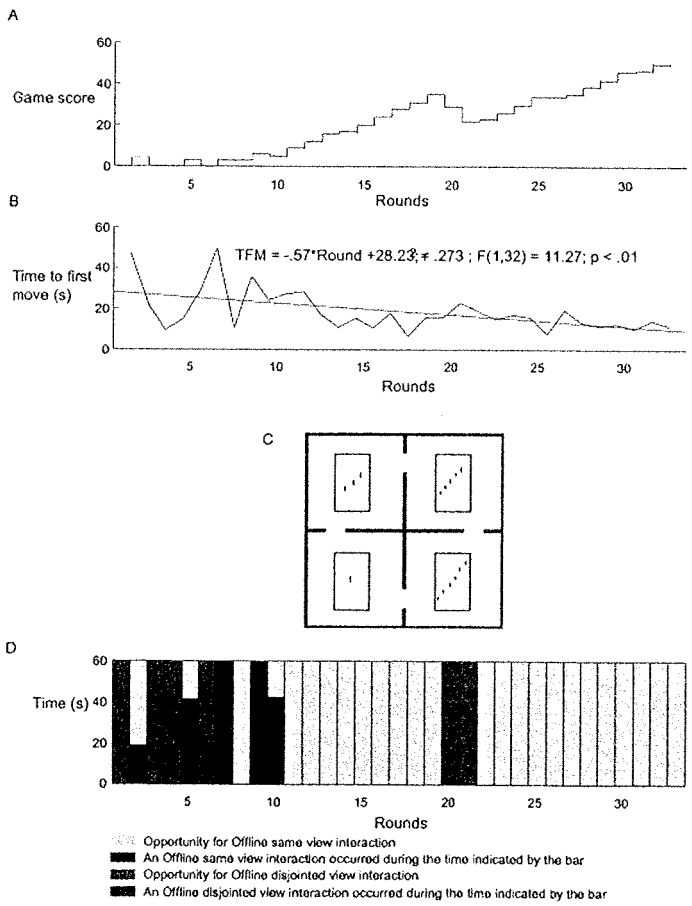

Plate 4 Pair A's basics. (A) Score during the first 33 rounds. (B) Time it took players to make the first move over the first 33 rounds. (C) Sign system developed by Pair A to solve Game 1. (D) Time spent in offline interactions over the first 33 rounds.

Round	Initial position	White player	Blue player	Move 1	White player	Blue player	Move 2	Outcome	Offline corrections	Notes
1		Draws circle twice	Draws flower		6 dots twice			Loss		
2					3 dots			Win	In same room, B is extremely attentive; W does 3 dots, B stops moving (2.18 - 2.22); 3 dots = TR	
3		5 dots; dot before moving				Draws circle		Loss		
4		3 dots			3 dots, angle, 3 dots	Draws circular shapes		Loss		
5		3 dots	After W, 3 dots then draws circles					Win	Reference by wall and by icon, dot becomes common currency. CR = 1 dot	
6		Dot, then dot and a line	Silent, moves after W signs			6-8 dots		Loss		B seems to have problem remembering location of CR, or W's line generates confusion.
7		Dot	Silent, moves after W signs					Win	Reference by wall and by icon, dot becomes common currency; HR = 5 dots	
8		5 dots, "going to", 1 dot	3 dots			1 dot		Win		W uses back-and-forth sign for "going to"
9		5 dots; "going to"	3 dots					Win	Joint attention on flower, players use pointing-by-bumping (back-and-forth sign may mean "?" as well as "Yes"); FR = 5 dots	
10		6 dots, "going to", 1 dot	3 dots, 1 dot					Win	Joint attention on wall toward HR, players use pointing-by-bumping, 6 dots, W uses back-and-forth to mean "going to" (1 dot, back-and-forth, 6 dots); HR = 6 dots	

Plate 5 Pair A's first ten rounds.

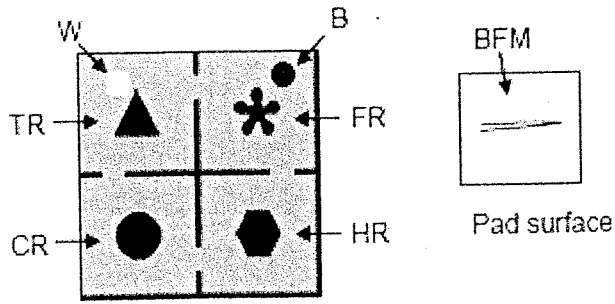

Plate 6 Abbreviations used in the round descriptions.

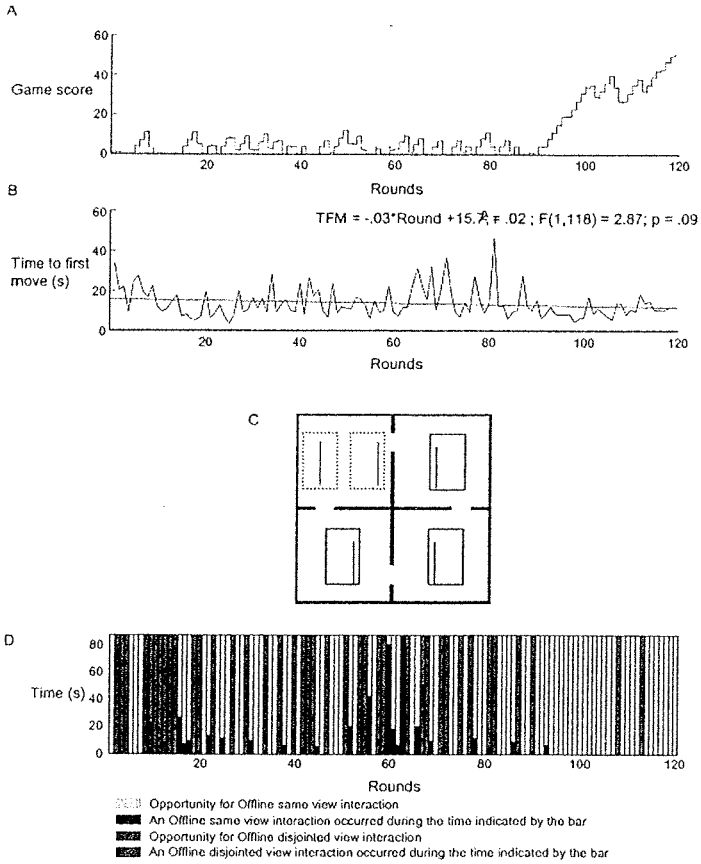

Plate 7 Pair B's basics. (A) Score during the first 119 rounds. (B) Time it took players to make the first move over the first 119 rounds. (C) Sign system developed by Pair B to solve Game 1. (D) Time spent in offline interactions over the first 119 rounds.

Round	Initial position	White player	Blue player	Move 1	White player	Blue player	Move 2	Outcome	Offline interactions	Notes
1		Draws triangle	Near door, H line C-L			Draws indistinct scribble		Win		
2		Draws portion of hexagon?	Draws indistinct scribble					Loss	B: Draws indistinct scribbles W: H line C-R, H line R-C (to quit round?)	Players moved almost simultaneously
3		Draws indistinct scribbles	Draws indistinct scribble					Loss	B: Draws indistinct scribble (perhaps the shape of the flower)	Players moved almost simultaneously
4								Loss	B: Draws indistinct scribble W: Draws indistinct scribble	Players moved almost simultaneously
5		Prolonged dot C	Draws indistinct scribble					Win		
6		Prolonged dot C	Draws indistinct scribble; H line C-R					Win		
7		Prolonged dot C	3 V lines B-U					Win		
8			Draws indistinct scribbles		Prolonged dot C			Loss	B: Draws indistinct scribbles W: BFM, (to quit round?)	Players moved almost simultaneously
9		Draws indistinct scribbles; prolonged dot R						Loss	W: Draws indistinct scribble B: BFM	
10		prolonged dot R						Loss	B: Draws indistinct scribbles	B moved twice, back and forth from FR

Plate 8 Pair B's first ten rounds.

Round	Initial position	White player	Blue player	Move 1	White player	Blue player	Move 2	Outcome	Offline interactions	Notes
1		Draws circle twice	Draws flower		6 dots twice			Loss		
2					3 dots			Win	In same room, B is extremely attentive: W does 3 dots, B stops moving (2:18 -2:22). 3 dots = TR	
3		5 dots, dot before moving				Draws circle		Loss		
4		3 dots			3 dots, angle, 3 dots	Draws circular shapes		Loss		
5		3 dots	After W, 3 dots then draws circles					Win	Reference by wall and by icon, dot becomes common currency, CR = 1 dot	
6		Dot, then dot and a line	Silent, moves after W signs			6-8 dots		Loss		B seems to have problem remembering location of CR, or W's line generates confusion.
7		Dot	Silent, moves after W signs					Win	Reference by wall and by icon, dot becomes common currency; HR = 5 dots	
8		5 dots, "going to", 1 dot	3 dots			1 dot		Win		W uses back-and-forth sign for "going to"
9		5 dots; "going to"	3 dots					Win	Joint attention on flower, players use pointing-by-bumping (back-and-forth sign may mean "?" as well as "Yes"); FR - 5 dots	
10		6 dots, "going to", 1 dot	3 dots, 1 dot					Win	Joint attention on wall toward HR, players use pointing-by-bumping, 6 dots, W uses back-and-forth to mean "going to" (1 dot, back-and-forth, 6 dots); HR = 6 dots	

Figure 11.5 Pair A's first ten rounds.

This event is crucial for three reasons. First, it demonstrates that B is highly attentive to W's behavior. Second, it demonstrates that W is prone to initiate offline interactions. Third, it demonstrates that W has developed the idea of referring to a room by going close to the door that would lead to it (henceforth, we will refer to this as "pointing").

- *Round 3.* In absence of any sign exchange, B moved from the room with a hexagon on the floor (henceforth HR) to CR. W, in the room with a flower icon on the floor (henceforth FR), produced five dots. B, in CR, drew circles on the digitizing pad a number of times. W moved to HR and the round was lost. There was no offline interaction.

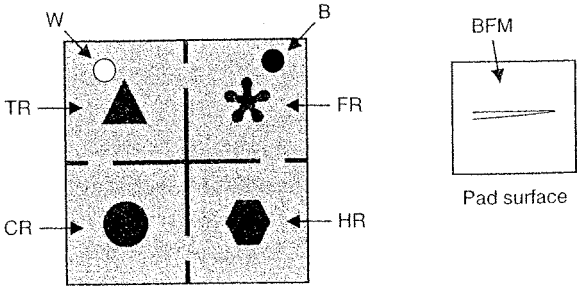

Figure 11.6 Abbreviations used in the round descriptions.

- *Round 4.* In TR, W produced three dots. B moved from HR to FR. Still in TR, W produced two series of three dots, separated by a wide back-and-forth movement of the pen on the pad (henceforth BFM). Once in FR, B drew the shape of the flower a number of times on the pad. W moved to CR and the round was lost. There was no offline interaction.
- *Round 5.* From this moment onward, B stopped moving the agent in absence of sign exchanges; signing had become the first activity of a round for both players. In TR, W produced three dots. In CR, B produced three dots (adopting W's signal units), then drew a circle on the pad. As soon as that happened, W moved to CR.

  There was an offline interaction during which three events occurred. First, W went near the door that would lead to TR. B moved very close to W, and then W produced three dots. Then W moved toward the door that would lead to HR, followed by B. Once B arrived near the door, W produced four dots. Then B moved very close to the circle, followed by W. B produced one dot, W repeated the dot, then both players moved toward reset squares.

  This interaction, not only confirms that the players agree on the use of the doors as proxies to refer to the room the door would lead to, but it also highlights four important points. The first one is that the players have now established the habit of following each other and entering into frames of joint attention. That is, not only players are aware of each other's positions, orientations and signaling activities, but they also modulate their behaviors according to their understanding of their respective states. The second point is that the players' roles in this frame of joint attention are interchangeable. Although the interaction was initially led by W, later on B took the initiative of moving and pointing to the circle icon. The third point is that the pointing routine is used flexibly; also the icon in the room, and not only the doors, may serve as an object to refer to. The fourth point is that W starts providing feedback, by repeating the sign produced by the partner.
- *Round 6.* In CR, W first produced a dot, then another dot and finally a very prolonged dot. B, first move toward CR then, after a long hesitation which seems to indicate that B understood the sign, moved from HR to FR. Once there, B did 11 dots. W moved to HR and the round was lost.

  There was no offline interaction. There are two possible explanations for B's mistake. Either B understood W's sign for CR but had forgotten where CR was with respect to HR or W's very prolonged dot confused B.
- *Round 7.* In CR, W produced a dot. With no hesitation, B moved from HR to CR.

  There was a long offline interaction (58 s) during which two events occurred. First, B and W moved close to the circle, W produced a dot and B repeated it. Second, B and W moved close to the door that would lead to HR. Once there, W produced five dots, which B repeated. Then W produced three dots. B "responded" with five dots, which W repeated.

This interaction highlights two important points. The first one is that both players confirm signs; from this moment onward the dot for CR is a completely stable sign. The second point is that players repair each other's signs. To indicate HR, W produced first five dots then, after B had repeated them, W produced three dots. B "corrected" this by producing five dots and W agreed on the correction, by repeating the five dots.

- *Round 8.* In TR, B produced three dots. Soon after W, in HR, produced five dots. After a while, W produced one dot. B immediately "responded" with three dots. W produced three dots followed by the BFM (suggesting that the BFM may indicate movement) and then one dot. In the mean time, B moved to CR, and produced one dot. W moved to CR soon after. There was no offline interaction.

  This round demonstrates that players are not only able to communicate at this stage of the game, but they are also looking for a procedure to coordinate their moves. This is particularly helpful when both agents have to move in order to find each other.

- *Round 9.* In TR, B produced three dots. W, in FR, responded with five dots, followed by the BFM. B, who was closer to the door that would lead to FR, moved to FR.

  There was a long offline interaction (58 s), during which a number of events occurred. W moved near the icon on the floor, followed by B. Once there, W produced five dots, then moved toward the door that would lead to HR. B followed W. However, before W had time to do anything, B went back to the icon on the floor. Once there, B waited for W to reach the icon as well. Once W was near the icon, B produced five dots, followed by signal composed of large amplitude oscillations (probably a request of agreement, since the five dots had been used before for HR). B bumped repeatedly against the icon on the floor (probably a request of signals from W). W produced five dots, followed by the large amplitude oscillations, which were immediately reproduced by B. Then B and W moved toward reset zones.
  This interaction highlights three important points. First, B initiates a repair. Five dots had been used before to indicate the HR. Now B wants to clarify that they are to be used for the FR. Second, B is not only aware that W's attention is needed to establish a frame of joint attention (and waits for W to reach the icon), but he also introduces the concept of "signal request", by bumping on the icon after having produced a sign that refers to it. Third, B introduces a signal (the large amplitude oscillation) to confirm that the sign is understood. Remarkably, all of B's new behaviors are understood and properly reciprocated by W. At this stage, the frame of joint attention established by B and W is providing rich scaffolding for communication.

- *Round 10.* In TR, B produced three dots. In HR, W produced six dots. B produced one dot. W produced a BFM, a dot and then moved to CR. B moved to CR soon after.

  There was an offline interaction. While W was moving toward the reset square, B moved toward the door that would lead to HR. W immediately reached B at the door. B bumped a few time against the door and then produced six dots. W produced one dot, a BFM and six dots. As soon as W finished producing the sixth dots, both players moved toward reset zones.

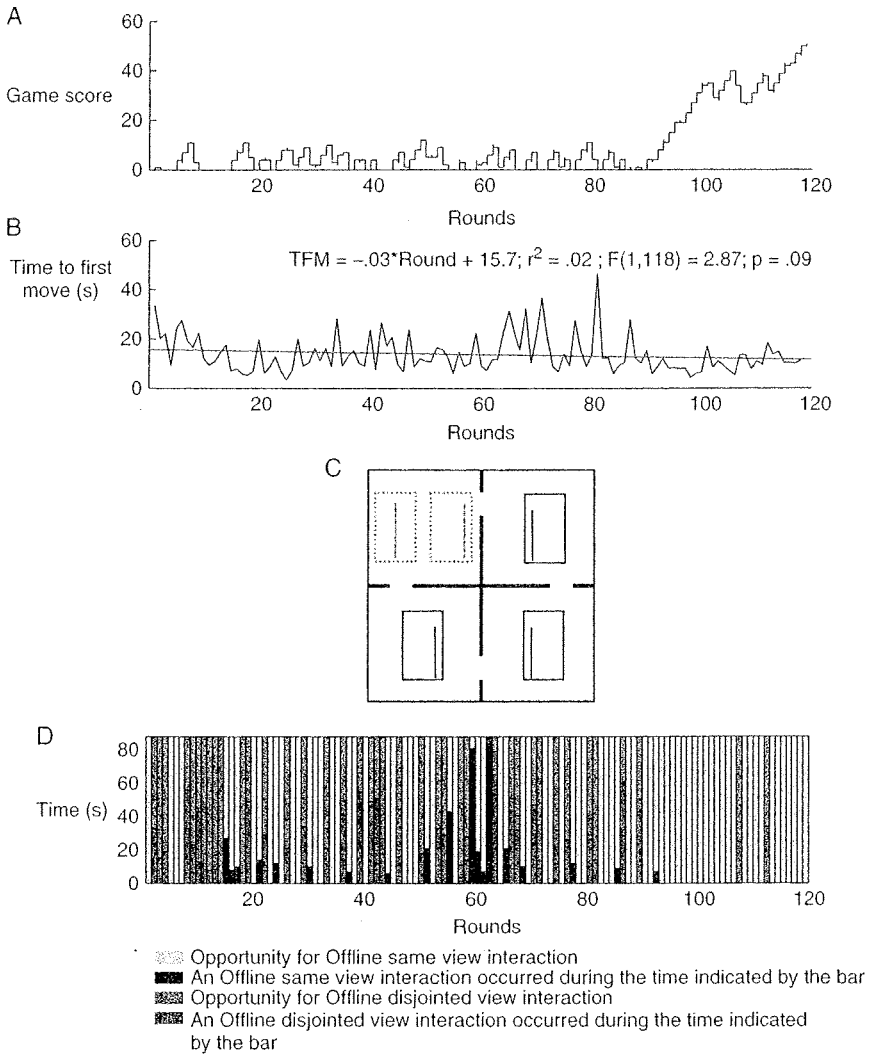

Figure 11.7 Pair B's basics. (A) Score during the first 119 rounds. (B) Time it took players to make the first move over the first 119 rounds. (C) Sign system developed by Pair B to solve Game 1. (D) Time spent in offline interactions over the first 119 rounds.

This interaction completes the sign system of the pair: three dots for TR, one dot for CR, five dots for FR and six dots for HR. The signs had been established in 10 rounds, within 12 minutes from the very beginning of the game. From this moment onward, the pair is extremely successful in playing the game, losing only two of the 23 rounds that would lead them to reach the threshold score for completing the game.[5]

[5] The two losses occurred because each player had to make a move and the moves were not coordinated properly.

Pair B. Pair B was one of the least successful pairs of the 30 pairs that played Game 1 and did not go beyond Game 1, after about 4 hours of playing. The pair completed Game 1 in 75 minutes of playing. During this time, Pair B played 119 rounds, losing 44 rounds (37%), four of which during the last 20 rounds of Game 1. The two top plots of Figure 11.7 illustrate the difficulties encountered by Pair B in Game 1. The first plot (Figure 11.7A) illustrates a prolonged hovering of the score at around zero, before the final rise. The second plot (Figure 11.7B) illustrates that there was only a slight decline in the time that it took players to make the first move in a round, an indication that players' confidence in how to play the game increased only slightly during the course of the game.

At the end of Game 1, Pair B's communication system comprised two agreed upon signs, one indicating the right side of the game environment and one indicating the bottom left room (Figure 11.7C). The two players used different signs for the top left room. The signs of Pair B were fairly verbose: For 63% of the whole Game 1 time, the digitizing pad was used by at least one of the two players.

Pair B's communication system was established later on in Game 1, after a large number of unsuccessful rounds. During these rounds, the pair recurred extensively to offline interactions. As illustrated in Figure 11.7D, Pair B recurred to offline disjointed view interactions as well as to offline same view interactions. Offline disjointed view interactions were not only completely irrelevant for establishing an effective communication system to play the game but, as detailed below, might have also been a significant source of confusion.

Round	Initial position	White player	Blue player	Move 1	White player	Blue player	Move 2	Outcome	Offline interactions	Notes
1		Draws triangle	Near door, H line C-L			Draws indistinct Scribble		Win		
2		Draws portion of hexagon?	Draws indistinct Scribble					Loss	B: Draws indistinct scribbles W: H line C-R, H line R-C (to quit round?)	Players moved almost simultaneously
3		Draws indistinct scribbles	Draws indistinct scribble					Loss	B: Draws indistinct scribble (perhaps the shape of the flower)	Players moved almost simultaneously
4								Loss	B: Draws indistinct scribble W: Draws indistinct scribble	Players moved almost simultaneously
5		Prolonged dot C	Draws indistinct scribble					Win		
6		Prolonged dot C	Draws indistinct scribble; H line C-R					Win		
7		Prolonged dot C	3 V lines B-U					Win		
8			Draws indistinct scribbles		Prolonged dot C			Loss	B: Draws indistinct scribbles W: BFM, (to quit round?)	Players moved almost simultaneously
9		Draws indistinct scribbles; prolonged dot R						Loss	W: Draws indistinct scribble B: BFM	
10		prolonged dot R						Loss	B: Draws indistinct scribbles	B moved twice, back and forth from FR

Figure 11.8 Pair B's first ten rounds.

- *Round 1.* In TR, W drew a triangle twice (for a graphical synopsis of the rounds, see Figure 11.8). In HR, B moved toward the left, near to the door that would lead to CR. Then, before crossing the door, B drew a series of horizontal lines from the center of the pad to the left of the pad. Then, B crossed the door and entered CR. W moved to CR and found the partner there.

  There was no offline interaction, except for a brief scribble produced by B on seeing the partner entering the room. The signal was ignored by W, who continued moving toward the reset zone.

- *Round 2.* In HR, W drew an indistinct scribble (perhaps half of the shape of a hexagon). In CR, B drew an indistinct scribble, then three vertical lines from the bottom of the pad to the top. Then, B moved (up) to TR, almost at the same time as W moved to FR. It is important to notice that, although B might have wanted to indicate the movements of the agent with the three vertical lines, this information was totally lost, since the pad could not reproduce vertical movements.

  The round was lost but there was an offline interaction. B moved toward the center of the room (away from the closest reset zone) and then drew an indistinct scribble. W, who remained in the reset zone all the time, "responded" with a horizontal line from the center of the pad to right of the pad, perhaps indicating to the partner to go to a reset zone.

  The interaction highlights two important points. The first one is that B is prone to perform behaviors irrelevant for the task at hand, such as moving around in the room while not seen (and while doors cannot be crossed) or emitting signals in a context in which the partner has no possibility to understand them. The second point is that W seems to understand the futility of the interaction, and seems to invite the partner to stop it.

- *Round 3.* In CR, W drew two indistinct scribbles. In HR, B drew two vertical lines from the top of the pad to the bottom. Then B moved (up) to FR, almost at the same time as W moved to HR.

  The round was lost and there was an offline interaction. B drew a series of indistinct scribbles and then moved to the reset zone. During this time, W remained in a reset zone.

- *Round 4.* W moved from FR to TR; B moved from FR to HR. The players moved almost simultaneously and without prior exchange of signals.

  The round was lost and there was an offline interaction. B drew an indistinct scribble. W, while in a reset zone, drew a horizontal line from the left side of the pad toward the center of the pad. B moved toward a reset zone which was not the closest available.

  The interaction highlights again the fact that B is prone to perform irrelevant behaviors. Moving toward a more distant reset zone when a closer one is available has no purpose in the game.

- *Round 5.* In FR, B drew indistinct scribbles. In TR, W drew a vertical line in the center of the pad. B moved to TR and found the partner there.

There was no offline interaction except that B went to the same reset zone as W, although the reset zone was not the closest available.

- *Round 6.* In TR, W drew a vertical line in the center of the screen. In CR, B drew indistinct scribbles. W moved to CR and found the partner there.

  There was no offline interaction.

- *Round 7.* In TR, W drew a vertical line in the center of the screen. In CR, B drew three vertical lines from the bottom of the pad to the top (in the center of the pad), then moved (up) to TR and found the partner there.

  There was no offline interaction.

- *Round 8.* In FR, B drew three vertical lines from the top of the pad to the bottom (in the center of the screen) then moved down to HR. In CR, W drew a vertical line at the center of the pad. Then, probably interpreting B's signal as the signal W adopted for TR, moved to TR.

  The round was lost and there was a prolonged offline interaction (56 s). B, while roaming around the room, drew a vertical line from the top of the pad to the bottom. W, while in a reset zone, "responded" with a small amplitude BFM. B, while still roaming around the room, drew a few more vertical lines from the top of the pad to the bottom and then an indistinct scribble. Finally, after a bit more roaming, B moved to the reset zone.

  The interaction highlights the fact that B's habit of performing irrelevant behavior during offline interactions has become stable and (possibly) more persistent.

- *Round 9.* In CR, W drew a series of indistinct scribbles, followed by a vertical line on the right side of the pad. Meanwhile, B moved from HR to FR. Finally, W moved to HR and the round was lost.

  There was an offline interaction. B kept roaming around the room while W was in a reset zone. After a bit of time, W drew an indistinct scribble in the center of the screen, probably to invite the partner to go to a reset zone. B responded with an indistinct scribble in the middle of the screen and went to a reset zone.

  The interaction highlights the fact that W has become aware of B's irrelevant offline behaviors and attempts to stop them.

- *Round 10.* In CR, W drew a prolonged vertical line on the right side of the pad, without moving. B moved from FR to TR, then back to FR. The round was lost and there was an offline interaction. B, while roaming around the room, produced a series of indistinct scribble. During this time, W remained in a reset zone.

After 10 rounds of the game, there is no evidence of effective communications between the players. Moreover, the players seem incapable of establishing functional frames for joint attention. This situation will remain unchanged for about 70 more rounds.

A comparison between Pairs A and B highlights a fundamental difference in the capability to perceive and create the conditions that support fruitful interactions.

Players in Pair A never attempted to hold an offline disjointed view interaction. Most likely, it did not even occur to them that there was anything relevant to do in such context. At the same time, they quickly developed the habit of holding offline same view interactions, which were the key to their success. Players in Pair B, on the contrary, quickly developed the habit to of holding offline disjointed view interactions (on B's initiative) and ignored opportunities for offline same view interaction. Considering the dynamics of the game that we illustrated early on in this section, the difference indicates that Pair A was much more inclined than Pair B to establish a functional sociocognitive frame for their interactions, from the very beginning of the game. A key factor for this was the fact that players in Pair A were constantly monitoring each other's behavior, and responded adaptively to it. On the contrary, players in Pair B often completely ignored each other's behavior and, consequently, little if any of their behavior became adapted to that of the partner. As illustrated in the rounds presented above, this difference had three dramatic consequences for the development of a communication system.

The first consequence concerns the use of the body as a resource for communication. Players in Pair A grounded the meaning of their signs primarily by exploiting the opportunities offered by embodied communication. That is, the body of the agents, albeit minimal in complexity, became a powerful tool that was used creatively to express new meanings (e.g. Round 5), as well as to repair unsuccessful interactions (e.g. Round 9). Players in Pair B never exploited these resources. For them, the body of the agents represented merely a challenge, since it forced them to hold different views of the game world.

The second consequence concerns the development of a sense of shared purpose and effectiveness in the game, which enabled players to diagnose problems. After a small number of interactions, both players in Pair A began to be able to clearly perceive their successes and their failures at communicating. This provided a sense of shared purpose to players, which, in turn, enabled them to repair the problems they encountered in the game, constantly improving their capacity to communicate. On the contrary, players in Pair B did not seem to develop the capacity to clearly perceive their problems in communicating. After a number of pointless interactions, they seem to become used to idea that their interactions could not be steered in any useful way.

The third consequence is straightforward After 10 rounds, Players in Pair A communicated all that was needed to win the game; Players in Pair B did not.

11.5 Conclusions

What can we learn from these experiments and particularly from comparing them? In what follows we illustrate three tentative conclusions.

A first conclusion is that many of the things that are implemented by design in the robots, such as routines for establishing frames of joint attention, turn out to be challenging for the human players to set up. That is, the human experiments focus as much on the prerequisites towards emergent communication (sharing cooperative goals, establishing joint attention, using communication, controlling the communication medium)

as on the emergence of the communication system itself. This is good news; human experiments have the potential to provide useful knowledge for robotic implementations of ever more realistic prerequisites for communication. For example, the data of the human players in Pair A presented above suggest that an important prerequisite that needs to be implemented in robots is the capacity to use the body in an expressive manner. Pair A's players had minimally embodied agents at their disposal. Yet, by using subtle space–time cues, Pair A's players harnessed powerful communicative devices out of the minimal bodies of their agents. Implementing this capability in robot experiments may provide valuable insight into the design of natural communication systems. At the same time, robot experiments may provide important controls for human experiments. Humans that participated in the experiments presented in Section 11.4 knew how established communication systems work in the social world to which players belonged. Due to the experimental limitations, this knowledge could not be used directly to communicate, but players could use it to guess how to set up a functional communication system. Although pairs' relatively high level of failure in the game suggests that such guesses did not provide an easy solution to the problem of developing functional communication systems from their very foundations, it would be desirable to ascertain more precisely their role in the human experiments. Robot experiments offer an opportunity to do so, since robots can be programmed so as to not posses any knowledge about how pre-established communication systems work. For example, the robots in the experiment presented in Section 11.3 had no pre-established knowledge about the usefulness of perspective reversals for communication. They discovered its usefulness when faced with the challenge to set up a communication system from its very foundations.

A second conclusion is that humans clearly operate within a framework of repair and consolidation strategies to set up a communication system, confirming that this is a useful framework to analyze the dynamics of emergent communication. Moreover, the human data indicated an important aspect of human repair and consolidation frameworks. Repair strategies used by one player had to be coordinated with those used by the other, a metalevel of coordination that is currently absent in robot experiments. Implementing such metalevel of coordination in robots will be an instructive challenge.

A third conclusion is that perspective reversal is a key ingredient of communication, both for robot and human players. In the case of the robot players, handling perspective reversal means to geometrically transform your own visual experience so as to reconstitute what it could have been for the other and only when this is systematically integrated in the language system do we see successful communication (cf. Figure 11.2). In the case of the human players, perspective reversal means above all to make a reasonable guess about how the other player will interpret your movements given what s/he can know about your own position and how s/he may interpret your sign. Most of the failures in human pairs occurred when one of the partners was unable or unwilling to adopt the perspective of the other.

Finally, we suggest one general conclusion about the theoretical framework adopted in this chapter. In principle, there are different ways in which human individuals or

artificial agents may arrive at a coordinated communication system: genetic evolution, intergenerational cultural evolution, or intragenerational (collective) problem solving. Here we explored the latter. We argued that if individuals/agents come to the task of communicating with a battery of problem-solving strategies for setting up a framework for joint attention (Tomasello and Farrar 1986) and joint action (Sebanz *et al.* 2006), for diagnosing communication failures and for repairing them by expanding or adjusting their communication conventions, a communication system will gradually arise and remain adaptive as individuals/agents encounter more or different challenges. The success of robotic agents to autonomously bootstrap a communication system (discussed in Section 11.3) and the empirical data from human experiments (discussed in Section 11.4) demonstrate that this approach is not only viable, but also empirically testable.

Acknowledgements

The preparation of this chapter was promoted and supported by the Center for Interdisciplinary Research of the University of Bielefeld. The help of Christian Kroos, Theo Rhodes and Michael Richardson in performing the experiments that provided the data for this chapter is gratefully acknowledged by Bruno Galantucci. Bruno Galantucci's project was supported by an NIH grant (DC-03782) to Haskins laboratories. This research of Luc Steels and coworkers was supported by the Sony computer Science Laboratory under a EU FET grant ECAgents (IST-1940).

References

Axelrod R (2005). Agent-based modeling as a bridge between disciplines. In KL Judd and L Tesfatsion, eds. *Handbook of Computational Economics,* Vol. 2: *Agent-Based Computational Economics,* Handbooks in Economics Series. North-Holland.

Baronchelli A, Dall'Asta L, Barrat A, and Loreto V (2007). The role of topology on the dynamics of the Naming Game. *European Physics Journal Special Topics,* **13**, 233–5.

Baronchelli A, Felici M, Caglioti E, Loreto V, and Steels L (2005). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics,* (P06014). http://arxiv.org/pdf/physics/0509075

Bickerton D (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences,* **7**, 17–388.

Boyd R and Richerson PJ (1985). *Culture and the Evolutionary Process.* Chicago: University of Chicago Press.

Brennan SE and Clark HH (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology-Learning Memory and Cognition,* **22**, 1482–93.

Briscoe T (2000). Grammatical acquisition: inductive bias and coevolution of language and the language acquisition device. *Language,* **76**, 245–96.

Briscoe T, ed (2002). *Linguistic Evolution Through Language Acquisition: Formal and Computational Models.* Cambridge: Cambridge University Press.

Cangelosi A and Parisi D (1998). The emergence of a 'language' in an evolving population of neural networks. *Connection Science,* **10**, 83–97.

Cangelosi A and Parisi D, eds (2002). *Simulating the Evolution of Language.* London: Springer-Verlag.

Clark HH (1996). *Using Language.* Cambridge: Cambridge University Press.

Clark HH and Wilkes-Gibbs D (1986). Referring as a collaborative process. *Cognition,* **22**, 1–39.

Croft W (2000). *Explaining Language Change: an evolutionary approach.* London: Longman Publishing Group.

Galantucci B (2004). Toward an experimental method for studying the emergence of human communication systems. *Dissertation Abstract International*, **65**, 2673B. (UMI No. 3134786).

Galantucci B (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, **29**, 737–67.

Galantucci B, Fowler CA, and Richardson MJ (2003). Experimental investigations of the emergence of communication procedures. In R Sheena and J Effken, eds. *Studies in Perception and Action*, VII, pp. 120–4. Mahwah, NJ: Lawrence Erlbaum Associates.

Galantucci B, Kroos C, and Rhodes T (2006). Rapidity of fading and the emergence of duality of patterning. In A Cangelosi, ADM Smith, and K Smith (Eds.), The Evolution of Language-Proceedings of the 6th International Conference on the Evolution of Language, pp 413–15. London: World Scientific

Garrod S and Anderson A (1987). Saying what you mean in dialog—a study in conceptual and semantic coordination. *Cognition*, **27**, 181–218.

Garrod S and Doherty G (1994). Conversation, coordination and convention—An empirical investigation of how groups establish linguistic conventions. *Cognition*, **53**, 181–215.

Garrod S and Pickering MJ (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, **8**, 8–11.

Goldin-Meadow S (2003). *The Resilience of Language: what gesture creation in deaf children can tell us about how all children learn language.* New York: Psychology Press.

Goldin-Meadow S, McNeill D, and Singleton J (1996). Silence is liberating: Removing the handcuffs on grammatical expression in the manual modality. *Psychological Review*, **103**, 34–55.

Healey PGT, Swoboda N, Umata I, and Katagiri Y (2002). Graphical representation in graphical dialogue. *International Journal of Human-Computer Studies*, **57**, 375–95.

Healey PGT, Swoboda N, Umata I, and King J (2007). Graphical language games: Interactional constraints on representational form. *Cognitive Science*, **31**, 285–309.

Hockett CF (1960). The origin of speech. *Scientific American*, **203**, 89–96.

Hudson Kam CL and Newport EL (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, **1**, 151–95.

Hutchins E (1995). *Cognition in the Wild.* Cambridge, MA, US: MIT Press.

Kegl J (1994). The Nicaraguan sign language project: An overview. *Signpost*, **7**, 24–31.

Kimbara I (2006). On gestural mimicry. *Gesture*, **6**, 39–61.

Kirby S (2000). Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In C Knight, M Studdert-Kennedy, and JR Hurford, eds. *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pp. 303–23. Cambridge University Press.

Kirby S and Hurford J (2002). The emergence of linguistic structure: an overview of the iterated learning model. In A Cangelosi and D Parisi, eds. *Simulating the Evolution of Language*, pp. 121–48. London: Springer Verlag.

Krauss RM and Weinheimer S (1964). Changes in reference phrases as a function of frequency of usage in social interaction—a preliminary study. *Psychonomic Science*, **1**, 113–14.

Lakoff G and Johnson M (1999). *Metaphors We Live By.* New York: Basic Books.

Larson R, Borroff M, and Yamakido H, eds (2007). *The Evolution of Language.* Cambridge, UK: Cambridge University Press.

Lefkowitz N (1991). *Talking Backwards Looking Forwards: The French Language Game Verlan.* Tübingen: Gunter Narr Verlag.

Minett J and Wang W (2005). *Language Acquisition, Change and Emergence: Essays in evolutionary linguistics.* Hong Kong: City University of Hong Kong Press.

Mufwene S (2001). *The Ecology of Language Evolution*. Cambridge: Cambridge University Press.

Pardo JS (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, **119**, 2382–93.

Pickering MJ and Branigan HP (1999). Syntactic priming in language production. *Trends in Cognitive Sciences*, **3**, 136–41.

Pickering MJ and Garrod S (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, **27**, 169–226.

Pinker S and Jackendoff R (2005). The faculty of language: what's special about it? *Cognition*, **95**, 201–36.

Polich L (2005). *The Emergence of the Deaf Community in Nicaragua*. Gallaudet University Press.

Sebanz N, Bekkering H, and Knoblich G (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, **10**, 70–6.

Steels L (1997). The synthetic modeling of language origins. *Evolution of Communication*, **1**, 1–34.

Steels L (2000). Language as a complex adaptive system. In M Schoenauer, K Deb, G Rudolph, X Yao, E Lutton, JJ Merelo, and H-P Schwefel, eds. *Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*, PPSN VI, Lecture Notes in Computer Science, pp. 17–26. Berlin: Springer- Verlag.

Steels L (2003). Evolving grounded communication for robots. *Trends in Cognitive Sciences*, **7**, 308–12.

Steels L and Belpaeme T (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, **28**, 469–89.

Steels L, Kaplan F, McIntyre A, and Van Looveren J (2002). Crucial factors in the origins of word-meaning. In A Wray, ed. *The Transition to Language*, pp. 252–71. Oxford, UK: Oxford University Press.

Steels L and Loetzsch M (2007). Spatial language in dialogue. In KR Coventry, T Tenbrink, and J A Bateman, eds. *Perspective Alignment in Spatial Language*. Oxford: Oxford University Press.

Steels L and Wellens P (2006). How grammar emerges to dampen combinatorial search in parsing. In P. Vogt *et al.*, eds. *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, pp. 76–88. Springer.

Tomasello M (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Tomasello M (2005). Beyond formalities: The case of language acquisition. *Linguistic Review*, **22**, 183–97.

Tomasello M and Farrar MJ (1986). Joint attention and early language. *Child Development*, **57**, 1454–63.

Traugott E and Heine B (1991). *Approaches to Grammaticalization*, Vol I and II. Amsterdam: John Benjamins Publishing Cy.

Wagner K, Reggia JA, Uriagereka J, and Wilkinson GS (2003). Progress in the simulation of emergent communication and language. *Adaptive Behavior*, **11**, 37–69.