Chapter 13
Emergent Action Language on Real Robots

Luc Steels[1,3], Michael Spranger[2], Remi van Trijp[1], Sebastian Höfer[5], and Manfred Hild[1,4]

This paper is the authors' draft and has now been officially published as:
Luc Steels, Michael Spranger, Remi van Trijp, Sebastian Höfer and Manfred Hild (2012). Emergent Action Language on Real Robots. In Luc Steels and Manfred Hild (Eds.), *Language Grounding in Robot*, 255–276. New York: Springer.

Abstract Almost all languages in the world have a way to formulate commands. Commands specify actions that the body should undertake (such as "stand up"), possibly involving other objects in the scene (such as "pick up the red block"). Action language involves various competences, in particular (i) the ability to perform an action and recognize which action has been performed by others (the so-called mirror problem), and (ii) the ability to identify which objects are to participate in the action (e.g. "the red block" in "pick up the red block") and understand what role objects play, for example whether it is the agent or undergoer of the action, or the patient or target (as in "put the red block on top of the green one"). This chapter describes evolutionary language game experiments exploring how these competences originate, can be carried out and acquired, by real robots, using evolutionary language games and a whole systems approach.

Key words: language acquisition, action language, mirror systems, case grammar

13.1 Introduction

It is obvious that the integration of robots in human activities will require them to understand and execute commands. Action language rests on the same elements as other forms of language: mechanisms for sensory-motor grounding, for perceiving and conceptualizing reality, for parsing and producing sentences, and for maintaining dialog. These mechanisms have been discussed in detail in earlier chapters of

[1]Sony Computer Science Laboratory Paris, e-mail: steels@arti.vub.ac.be
[2]Systems Technologies Laboratories, Sony Corporation, Tokyo, Japan
[3]ICREA Institute for Evolutionary Biology (UPF-CSIC) Barcelona
[4]Neurorobotics Research Laboratory, Humboldt-Universität zu Berlin
[5]Robotics and Biology Laboratory, Technische Universität Berlin

this book. But action language also requires some additional challenging functionalities:

1. Speakers must be able to recognize actions they requested so that they can check whether the appropriate action was performed. Hearers must know which motor control programs to trigger to achieve a requested action. And learners must be able to relate an action carried out by others to their own motor control programs, otherwise they would not be able to learn the meaning of action words by examples shown to them. The latter is known as the *mirror problem* and has been discussed extensively in the recent neuroscience and psychological literature due to the discovery of so-called mirror neurons (Rizzolatti and Craighero, 2004). The observation of mirror neurons tells us that there are circuits in the brain responding both to the perception and the execution of the same action, however it does not provide us yet with a mechanistic theory, i.e., how mirror neurons take on these functions. We will need such a mechanistic theory to solve the mirror problem on physical robots.
2. Speakers and hearers must be able to recognize the affordance of objects. *Affordance* means to understand which objects can be acted upon with what kind of actions (Gibson, 1979). Affordance is based on a combination of object perception, expectations about possible actions, prior knowledge of object affordances, and clues given by language. The latter usually takes the form of constituent orderings, grammatical markers, particles or prepositions and is generally known as a case grammar (Fillmore, 1968). Case grammars specify the roles of objects with respect to an action. For example in the utterance "push the red block to the green one", the red block is the object of pushing because it is the direct object of the sentence and the green one is the target as signalled by the preposition "to".

The remainder of this chapter discusses robotic experiments focusing on how these two issues can be approached from a mechanistic point of view. For the mirror problem we build on earlier research in evolutionary language games as reported in Steels and Spranger (2008a,b, 2012) and for the affordance problem on similar experiments in the emergence of case grammar, as reported in van Trijp (2012). We will see that the solution to both problems lies in setting up the right kind of dynamics so that semiotic networks form that represent the needed associative relations, and this requires in turn a deep integration of all layers of intelligence, so that each component can constrain, interrogate, and inform another one.

13.2 The Mirror Problem

13.2.1 Circumscribing the problem

In order to achieve grounded action language, language users need to know the relations between the visual appearance of an action, the motor control programs

that perform the action, and the proprioceptive streams that are perceived while the action is ongoing. This is more difficult than it may seem because seeing an action being performed does in itself not give an immediate clue about how one can achieve the same action, and occasionally the same visual appearance can be achieved by different motor control programs. Indeed, often the precise details of an action do not matter. For example, picking up a block may happen with two fingers or with a thumb and the remaining fingers.

Language users also need to know, as speakers, which word names an action, and, as hearers, which action is named by a given word. Again this is more difficult than it seems because there may be more than one name for the same action (synonymy), more than one action for the same name (polysemy), an agent may still be learning what action is named by what word, and there may be variation in the population with respect to how to name actions.

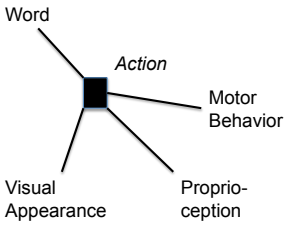

Fig. 13.1 Action language requires knowing the relation between visual appearances of actions, motor control programs that execute them, proprioceptive streams, and words that name the actions.

Figure 1.1 shows the key relationships that language users need to know for successful action language, with a node in the middle (labeled `action`) to bind them together. It is in principle also possible that the different information items are directly linked to each other without such a central node, and this may be an intermediary stage. But a central node is more effective for later processing. Indeed, some kind of central representation of action is what observations of mirror neurons seem to suggest.

We call these networks *semiotic networks*, as they relate words with sensorymotor experiences and motor control programs. Because agents have to learn and develop these networks progressively based on partial evidence, we get typically a much more messy picture as shown in Figure 1.2. The same visual appearance may correspond to many different motor control programs and the same motor control program may give rise to many different appearances depending on the objects being manipulated or the point of view of the observer. The same word may be used for different actions, particularly if agents are not yet sure how to name some action, and the same action may be named by many different words. Particularly in a developing agent or if the language is not yet completely settled we must expect a very complex

network. Moreover it is very unlikely that the networks of two agents are going to be the same because they will each have different histories of interaction with the world. To handle intermediary states, all the links in the semiotic network are assumed to be weighted with a score. The score is adjusted when confidence in a link increases or decreases.

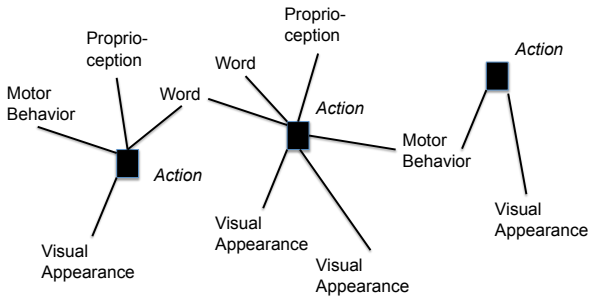

Fig. 13.2 In real world settings, semiotic networks are more complex because each node in the network can be related in more than one way to the same type of node.

The key question is how agents can build up the required relationships and how they can get networks that approach as much as possible those shown in Figure 1.1. We argue that this can happen only if information from many different sources is maximally used. The network can only be built up in a piecemeal fashion, and become progressively coordinated and aligned. This requires that we exploit the whole systems approach to the full, including the potential for self-organization in the multi-agent system, and that we have sequences of interactions between the different members of a population.

More concretely, the following processes can help to establish semiotic networks for action language:

1. Robots can learn about the relationship between the visual signature and proprioception of an action on the one hand and their own bodily movements by inspecting their own body directly, or by looking at a mirror, and monitoring the proprioceptive and motor signal streams (see Figure 1.3). Such experiments, carried out on the Sony humanoid robots (Fujita et al, 2003), were already described in Steels and Spranger (2008a), and an earlier chapter by Höfer et al (2012) has shown how semantic information can be extracted from proprioceptive and visual data streams in a bottom-up unsupervised manner. Both visual appearances and proprioceptive experiences can be captured in terms of prototypes in a suitable conceptual space, so that nearest neighbor computation can be used to classify them.
2. Robots can also use internal simulation in order to learn about the expected relation between their own movements and motor command signals, proprioceptive streams, and possibly rough visual appearance, for example by using a stick figure representation (see Figure 1.4 from Steels and Spranger, 2008b). Such an

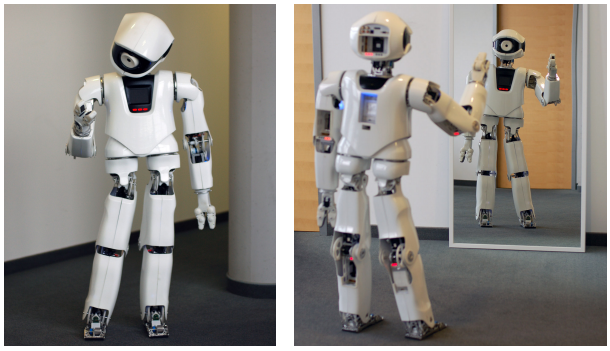

Fig. 13.3 A robot can look at its own body while performing movements or look at itself in the mirror while motorbabbling, in order to acquire relations between visual appearances of actions, proprioceptive experiences, motor commands, and actions.

approach was also used in earlier experiments on Sony humanoid robots, specifically in order to guess the meaning of unknown words (Steels and Spranger, 2008b), and an earlier chapter by Schilling (2012) already discussed in great detail how such internal simulation models can be implemented using recurrent neural networks.

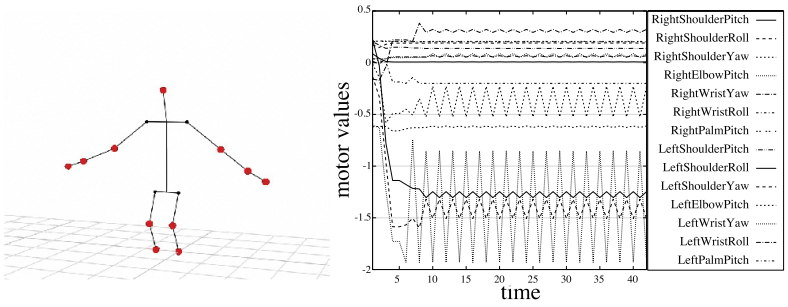

Fig. 13.4 Simulating the effect of certain movements so that expected proprioceptive streams and visual prototypes can be derived is another source for building up semiotic networks. A stickfigure can be used internally to get a rough idea of what a simulated action will look like visually.

3. Another source of information, which requires the integration of several more components, is verbal interaction (Steels and Spranger, 2008a). When an action is named by a speaker, this signals to the hearer that this action is to be differentiated from other actions. If the speaker does not recognize this action, he can then get clues about its visual appearance after the speaker shows the action, and he can get in turn feedback from the speaker on whether his own execution of the action is conform to that expected by the speaker, which means that he can learn

about the appropriate motor control commands to achieve that action. In this way there can be a top-down influence of language on what counts as an action and on the inventory of actions that agents in the group are sharing. The remainder of this section focuses on language game experiments that exploit language in this way.

13.2.2 The Action Game

We assume that agents individually acquire different actions by exploring their body and the world. For example, they may engage in motor babbling or engage in interactions with objects and progressively acquire actions that could be used as building blocks to achieve something in the world. Experiments in this direction are described by Oudeyer et al (2007), a.o. Although speaker and hearer will thus develop an inventory of actions, there is not necessarily any coordination across different agents because each agent will have had a different history of interactions with the world. This is where action games become relevant.



Fig. 13.5 An example of interaction during the language game. Left: The speaker (right robot) asks the hearer (left robot) to execute an action. Middle: The hearer interprets and performs the action. Right: The speaker signals failure and then performs the correct action.

An Action Game works as follows (see Figure 1.5). Two agents are randomly chosen from the population and downloaded in two robot bodies. The robotic agents then go through the following script:

1. The speaker chooses randomly an action from his inventory of actions.
2. The speaker retrieves the name for this action in his vocabulary and transmits that to the hearer. If there is no name yet, the speaker may invent a new name.
3. The hearer retrieves the intended action by looking up the name in his own vocabulary and evokes the associated motor behavior.
4. The speaker observes the action carried out by the hearer and checks whether it fits with the prototypical visual body-image of the action he had originally chosen. Note that this requires that the speaker has an association between visual body-image and action.
5. If this is NOT the case, the speaker signals failure. The speaker activates his own motor behavior associated with the action in order to repair the communication,

so that there is an opportunity for the hearer to learn the speaker's name for the action.
6. Otherwise the speaker signals success.

As discussed in the previous chapter (Gerasymova and Spranger, 2012), it is crucial that speaker and hearer *align* after each successful interaction. This means that the scores of all the links implicated in a decision are increased in the case of a successful game and decreased when there is an unsuccessful game. A lateral inhibition dynamics can be used to speed up convergence, which means that alternative links (for example other names for the same action or other actions for the same name) are decreased in the case of a successful game.

Step 5 is obviously the most difficult step, and the hearer can potentially try out different motor control programs until the speaker agrees that this is the one he originally had in mind. The semiotic network can be exploited to speed up acquisition in a multitude of ways:

- It is possible that the hearer already knew the action (i.e. had a motor control program and visual prototype for it) but did not know that the action was named that way. In this case, the visual prototype should match with the action performed by the speaker and the appropriate connections can be added in the network.
- It is possible that the hearer had another action associated with the name, in which case the game is negative evidence that this was an appropriate link.
- When the hearer does not know the action, he may engage in simulations, starting from the motor control programs for actions that match closely to the visual appearance observed. By varying parameters, the space around these motor control programs can be explored until a good match with the speaker's correct action is obtained. And this action can then be tried.

The action game can also be played between a human and a robot, as shown in Figure 1.6. In this case the human asks the robot to do a certain action and shows it when the robot fails to perform the action. Conversely, the robot may ask the human to do an action and this way get confirmation about links in the network that have low degrees of confidence. It is well known from studies of language acquisition that such active forms of learning are much more effective than purely passive observational forms of learning.

13.2.3 Experimental results

We have done a number of experiments where this action game is played by a population of agents. Here we look at one experiment in the emergence of a lexicon of action names from scratch, discussed in more detail in Steels and Spranger (2008a). The experiment involves 10 agents and an action repertoire of 5 actions. The states of two randomly chosen agents are downloaded into the body of the robots before they play a game and the changed states are uploaded after the game. One agent

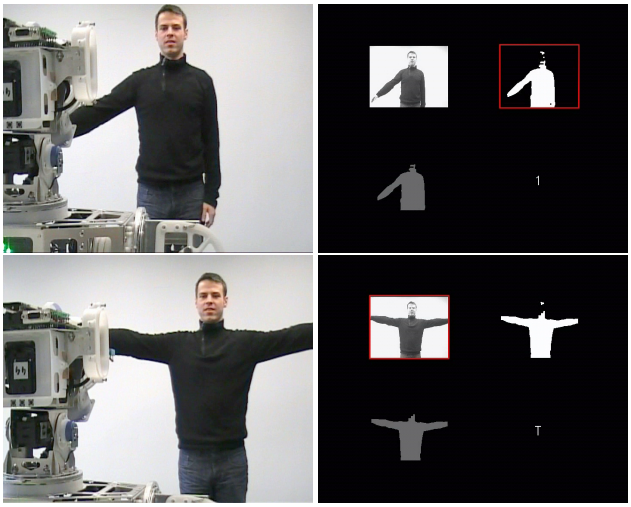

Fig. 13.6 Action game played between human and MYON robot. Here the human experimenter plays the role of the robot. Because the body shape is similar, the same feature extraction and pattern recognition mechanisms can be used as for games between two robots.

takes on the role of speaker and the other one that of hearer. Figure 1.7 gives an example of part of the semiotic network built by one of the agents. We observe indeed that networks typically are at the start partially unconnected or with many competing connections between nodes due to the uncertainty at the initial stage, both because agents are not yet fully certain of the connections between visual prototypes, motor behaviors and actions and because the choice which word to use for naming a particular action has not yet completely settled in the population. However we observe that gradually network complexity clears up to leave only networks with 'correct' connections and only a single word becomes dominant for each action to be named. Consequently there is routine success in language games.

These results are more clearly shown in Figure 1.8. It is a snapshot of the semiotic dynamics for the first 500 games. The top figure plots on the y-axis the communicative success, lexicon-size, and invention and adoption frequency, with the number of games shown on the x-axis. We see that communicative success rises steadily, to reach basically 100 % success. The number of words initially overshoots because words are being invented by agents not knowing that other words have already invented by others, but alignment gradually causes convergence. The invention rate peaks early on until all agents have at least one word for an action and the adoption rate gradually decreases until all agents know the words that are being used. At the bottom, we show the average score in the population for different words competing for naming one single action. We see clearly that a winner-take-all situation arises after about 300 games, with one word dominating for naming this particular action.

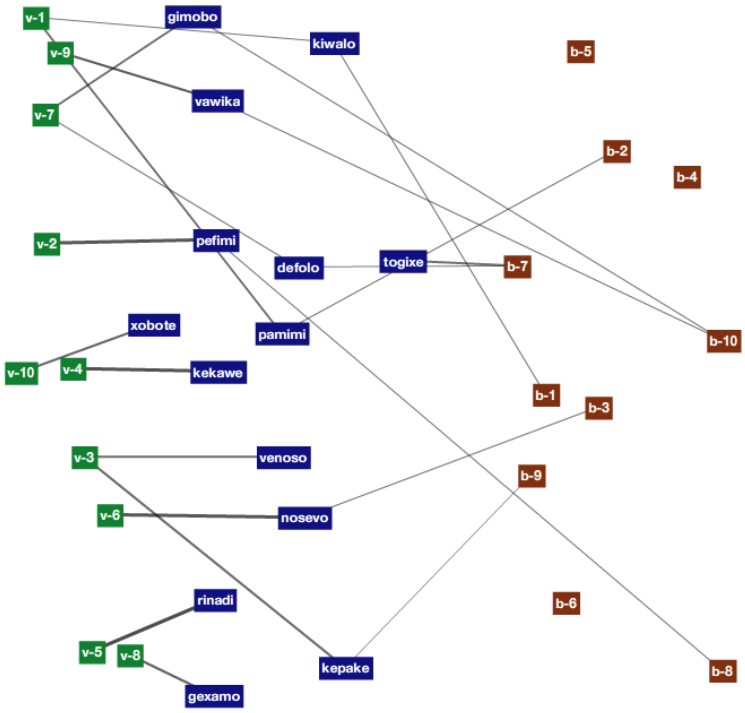

Fig. 13.7 A snapshot of the semiotic network built by one agent as the experiment unfolds. Relations are shown between invented words for actions (in the middle), visual prototypes (v_1, v_2, etc.) (on the left) and motor behaviors (b_1, b_2, etc.) (on the right).

The experiment therefore convincingly shows how language can help agents acquire which actions are shared in the population and to coordinate the relation between visual appearances and motor control programs in their own semiotic networks.

13.3 The Affordance Problem

13.3.1 Circumscribing the problem

When actions involve other objects, embodied agents are confronted with an additional problem, which is knowing about the possible participant roles in an action (usually called the event structure) and the affordance of objects that might fill these roles, i.e. understanding what can be done with an object and how this can be done

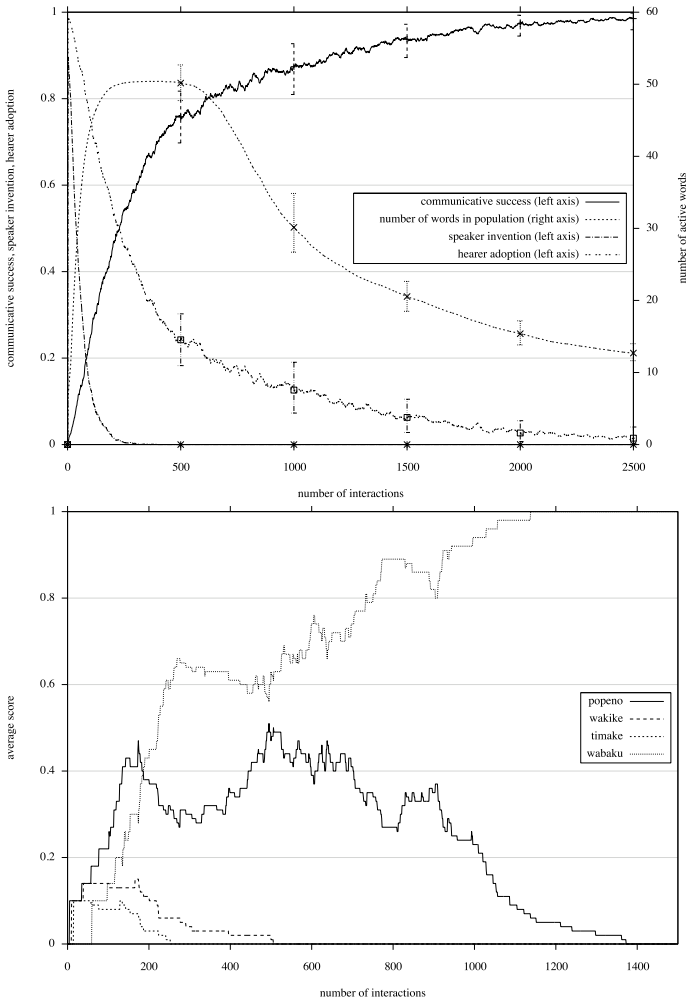

Fig. 13.8 Graphs zooming in on the first 2500 language games in a multi-agent experiment with 10 agents naming 10 actions. The top figure shows the global behavior and the bottom figure the average score for all the words competing for naming the same action. The winner-take-all dynamics that coordinates the lexicon among the agents is clearly visible and agents reach total success in their action games.

(Gibson, 1979). This is crucial for language, particularly for dealing with situations when a phrase could not be recognized well or when it contained unknown words. Expectations from the state of the world, the ongoing interaction, and expectations about what object might be involved in an action can then help to nevertheless figure out what the unknown speech elements could be about, and possibly predict missing words or guess meaning of unknown words. For example, if the following utterance is perceived: "Use ..⟨*unrecognized*⟩.. to nail down this plank", the agent can derive from knowing about the named action that some kind of hammer is needed or something that can play the role of hammer. If the utterance is "Use the hammer on the table to nail down this ... ⟨*unrecognized*⟩ ..." then something that could be nailed to something else, such as a plank, could be the referent of the missing nominal phrase.

Gibson, who originated the concept of affordance, argued that affordance can be perceived in a direct, immediate way without sensory processing, but nobody has so far proposed a mechanistic theory how this magical process is supposed to work. We hypothesize instead that learning about affordance is based on integrating knowledge from many sources over a long time span in order to build up the semiotic networks that can then be exploited to integrate the different manifestations of an action (visual, proprioception, motor control and language).

The first step to operationalize and test this hypothesis is to enrich the semiotic network with relations that cover not only the action itself (as in Figure 1.1) but also the participants in the action (see Figure 1.9). Each of these participants is characterized by certain properties that allow an object to fill this role and it may have also a way in which it is marked in language.

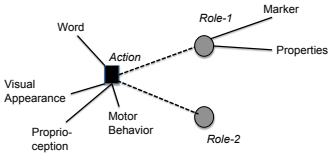

Fig. 13.9 When actions involve objects, the original semiotic network needs to be extended with information about the roles that are expected, the properties of these roles and how they are marked through language.

As with the earlier example of simple actions, realistic semiotic networks will look a lot messier, partly because it is well known that in human languages the marking of participant roles is not only language-dependent (and therefore it has to be learned), but also speaker-dependent, and shifting during language learning and

language evolution. Moreover markers of participant roles tend to be polysemous, meaning that the same marker is typically used for a variety of roles. For example, the Latin dative expresses roles such as the purpose of the action (dativus finalis), the beneficiary of the action (dativus commodi), who is concerned with the action (dativus ethicus), etc. On top of that, the same role can be expressed with a variety of markers depending on further context. For example, in Hungarian a verb normally using an accusative for the object undergoing the action switches to genitive when the sentence is negated.

Language typologists use semantic maps to represent these complex relationships (see Figure 1.10 from Haspelmath, 2003). The map shows the different possible roles in an action and how certain markers cover specific regions on these maps. This way the expression of participant roles in different languages can be compared. Semantic maps can also be used to display progress in language acquisition or how markings are changing over time in (artificial) language evolution.

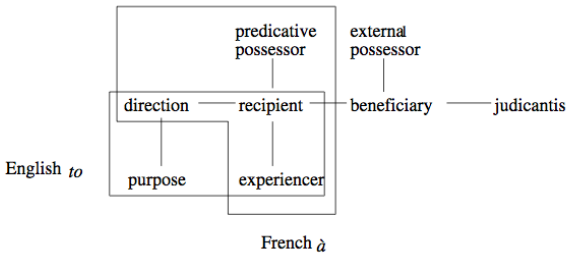

Fig. 13.10 A semantic map contains the different participant roles of actions together with regions covered by certain markers. In this case, there is a set of roles covered by the preposition "to" (in English) or "*à*" (in French). Different languages cover different areas of such maps and maps develop and shift during development and evolution.

The following processes help to establish semiotic networks for action language:

1. What kind of objects can play a role in a particular action and what their respective roles are can be acquired by 'playing' with the object, i.e. by manipulating the object - or several objects - trying to see what kind of actions they afford and how exactly the objects need to be manipulated to reach certain effects. For example, different push actions can be tried with objects of a different nature whereby it becomes clear that some objects (like blocks) can be pushed whereas heavy or fixed objects cannot be pushed. There may also be locations on an object where the pushing should take place. The notion of 'pushable' is of course very difficult to pin down and such categorizations presumably emerge only after many interactions and continue to be shaped by new experiences as the agent builds up more experience.
2. The affordances of objects can also be explored in simulation, although this requires that not only the body itself is simulated but also parts of the environment

that are interacting with the body, in other words that simulations are similar to those discussed earlier for evolving behaviors using evolutionary algorithms (Pasemann et al, 2012). There are two obstacles however to achieve this. The first question is how the world models necessary for such rich simulations can be acquired autonomously by the agent and the second one concerns complexity. Simulating interaction with objects at a sufficiently detailed level to derive useful predictions is in general a very computer-intensive process. The Berkeley 'Neural Language Theory' group has carried out concrete experiments in this direction using Petri-nets and they have explored the connection to language including the marking of participant roles (Feldman and Narayanan, 2004).

3. The third source for building up semiotic networks comes from language. As explained already, human languages conceptualize the world for language, meaning that they segment and categorize objects and their relations, introduce perspectives, highlight certain aspects as foreground, and so on, and they express these conceptualizations explicitly either in a lexical way or through grammar. Conceptualizations include also ways to categorize actions and how actions are taking place (for example in terms of temporal structure) and ways to categorize the roles of participants in terms of semantic abstractions, such as target or goal.

Different languages have different ways of conceptualizing actions and expressing them. For example, Atsugewi (an American Indian language) has a variety of prefixes and suffixes with the verb stem that indicate properties either of the verb or of the different participant roles of the actions denoted by the verb. Thus in the case of motion verbs, there are suffixes that signal the kind of path taken by the moving object, like "moving into a volumetric enclosure", or the kind of object that is moving, like "involving sticky material", and there are prefixes that circumscribe the type of event, for example what caused the motion, or what kind of instrument was involved to cause the motion. All of these provide hints for building up the rich semiotic network that constrains what kind of event might be expressed by a particular utterance (Talmy, 1988). But it also works the other way round. If confronted with an unknown action or unknown words for describing an action, these markers provide hints about the nature of the action that might be involved, what kind of objects can play a role in them and what the roles of these objects are. Concepts such as 'volumetric enclosure' or 'moving into' provide hints of the affordances of objects and how they relate to actions.

13.3.2 The Action Game Extended

We have been carrying out experiments, including the MYON robots (Hild et al, 2012, this volume), in order to see how markers for participant roles can be acquired or invented. In the experimental setup (see Figure 1.11), the robots are seated at a table with a set of physical objects on top of it. Each of these objects can be manipulated in various ways, that is each object affords a set of actions. For example, there is a fairly flat block which can only be pushed and a block with a peg that can

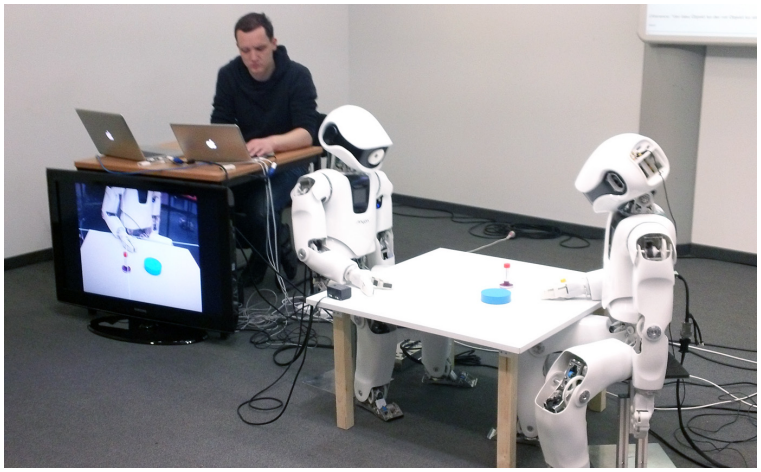

Fig. 13.11 Action language game played between two MYON robots. The human experimenter (in this case Michael Spranger) can monitor internal states as the experiment unfolds. The vision input of the left robot is shown on the monitor.

be picked up or pushed. The robots play games similar to the ones described earlier for researching the mirror problem. The speaker asks the hearer to perform a certain action and the game is a success if the hearer executes that action.

However, there are some important differences to the earlier experiments. The first difference is related to the environment and the action agents can perform. The environment is manipulated in such a way to clearly carve out the communicative pressure for expressing participant roles. For instance, in the setup depicted in Figure 1.11, both the blue and the red object afford a *push* action. In English, one might say 'push the red object *to* the blue object", in order to signal which object is the target. This allows us to study the mechanisms necessary for evolving a language that has no markers to one that does so.

The second difference is related to the processing capabilities of agents. Here, agents have to use compositional semantics and syntax in order to discriminate actions. We use the Incremental Recruitment Language (IRL, see Spranger et al, 2012, this volume) for processing of compositional semantics and Fluid Construction Grammar (FCG, see Steels et al, 2012, this volume) for compositional syntax. The two systems together enable agents to process phrases which involve actions as well as participants and the relation between them.

Lastly, we equip agents with acquisition and invention operators that orchestrate language change. This sets up the right type of communicative pressure and capabilities to study the acquisition and evolution of complex action language in a grounded scenario.



Fig. 13.12 Semantic program autonomously constructed by an agent, in order, to talk about a possible action. This semantic structure denotes a *push* action which has *the red object* as undergoer and *the blue object* as target. Notice the unique selector maps onto the determiner "the".

Let us exemplarily go through an interaction that agents were engaged in using the setup depicted in Figure 1.11. Both agents are up-front equipped with a pidgin-style English without participant markers.

1. The speaker observes the scene and picks a target action he wants the hearer to perform. In this case, he wants the hearer to push the red object to the blue object.
2. The speaker constructs a meaning (see Figure 1.12) and from that he constructs the utterance "push the red object the blue object".

3. Before uttering the phrase, he re-enters (Steels, 2003) the phrase and detects that this phrase is actually ambiguous and could be interpreted as meaning *push the red object to the blue object* or *push the blue object to the red object*. The agent fixes the problem by *inventing* the new marker "-ko" and attaches it to the *undergoer*. The speaker says to the hearer "push the red object-ko the blue object"
4. The hearer parses the utterance. Of course, he does not know what the meaning of the marker is. But given the flexible interpretation mechanisms of FCG and IRL, he constrains the possible actions down to either pushing the red block to the blue one or the other way around. He then chooses one of the actions at random. Let us suppose, he chooses the correct one and he pushes the red object to the blue object.
5. The speaker checks whether the hearer performed the correct action. Here, this is the case and the speaker signals to the hearer that he was correct.
6. This provides an opportunity to the hearer who can now be sure that his guess was correct and learn that the marker *-ko* is marking the undergoer of the push action.

We use these sorts of interactions to research how a case grammar, i.e. a grammar that marks participant roles in language, can emerge. The next section elaborates on one such experiment already discussed in more detail by van Trijp (2012).

13.3.3 Experimental Results

The experiment involves a population of 10 agents and a repertoire of 21 actions. There are 6 monovalent actions (such as 'move' and 'appear'), 11 bivalent actions (such as 'grasp' and 'approach') and 4 trivalent actions (such as 'give' and 'take'), which yields 40 event-specific participant roles. Each language game, two agents are randomly drawn from the population to act as either the speaker or the hearer.

The case grammars that evolve in the experiments show striking similarities with case systems found in natural languages. Figure 1.13 shows the two most frequent case markers from one of the emergent languages in the experiments. Apart from a couple of idiosyncratic case markers, this particular language has evolved an agentive vs. non-agentive system, which follows the universal trend in language to distinguish between agents and patients. As can be seen in the figure, the left marker *-tux* typically covers participant roles that involve an actor (such as getting closer, moving away, appearing and going outside). The right marker *-pav* typically covers participant roles that undergo an action (such as objects that are being grasped or approached), involve a state (such as being visible or standing still) or a location (such as the source or destination of a movement).

Where do these case markers come from? When describing the actions they perceive, the agents only have a lexical language at their disposal and hence always have to infer from their situation model which object played which role in the action. In order to avoid this effort, the agents start to categorize the participant roles

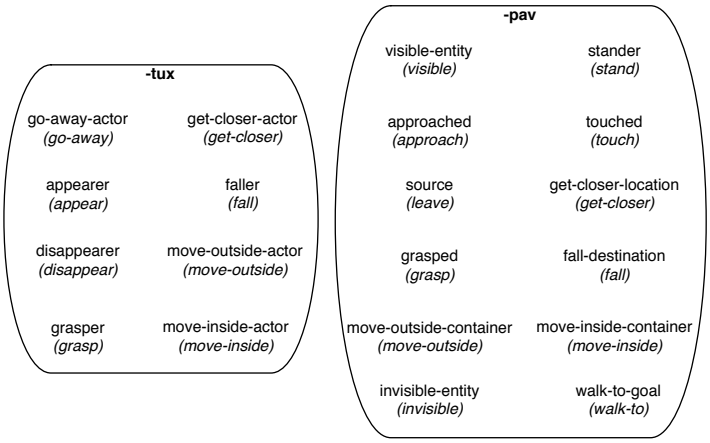

Fig. 13.13 This diagram shows the two most frequent case markers of one of the emergent languages in the experiments. The markers capture an agent versus patient distinction, which is a universal trend in natural languages as well. The markers are boldfaced; the participant roles they cover are grouped together with the events they belong to in italics between brackets.

of an action in terms of more abstract semantic roles and invent case markers for expressing those roles in language, using a meta-level architecture as described by Beuls et al (2012).

Figure 1.14 provides more insights into the semiotic dynamics of the evolved case systems. The chart at the bottom shows the competition of case markers over time as observed by a single agent. The X-axis shows the number of language games played in the entire population, the Y-axis shows 'type frequency' or the number of participant roles that can be covered by a marker. As can be seen, there is a lot of variation and strong competition between different markers in the beginning of an experiment. As agents start to change the weights in their semiotic networks depending on the communicative outcome of each language game, however, three markers become dominant in the language.

The top chart shows three different measures (cognitive effort, coherence and systematicity), which are all recorded in successful language games. The X-axis shows the number of language games played in the whole population (with two agents playing at each game), the Y-axis shows the values of the three measures between 0 and 1. Cognitive effort indicates the semantic ambiguity of utterances and thus how much effort an agent needs for interpreting them (see Steels, 2012). As can be seen, cognitive effort is reduced to a minimum after 6000 language games, which indicates that the case grammar allows the agents to interpret their language with less ambiguities. The coherence measure tracks whether the population converges on the same links in their semiotic network. In the beginning of the experiments, coherence decreases as agents start expanding the nodes in their semiotic network to

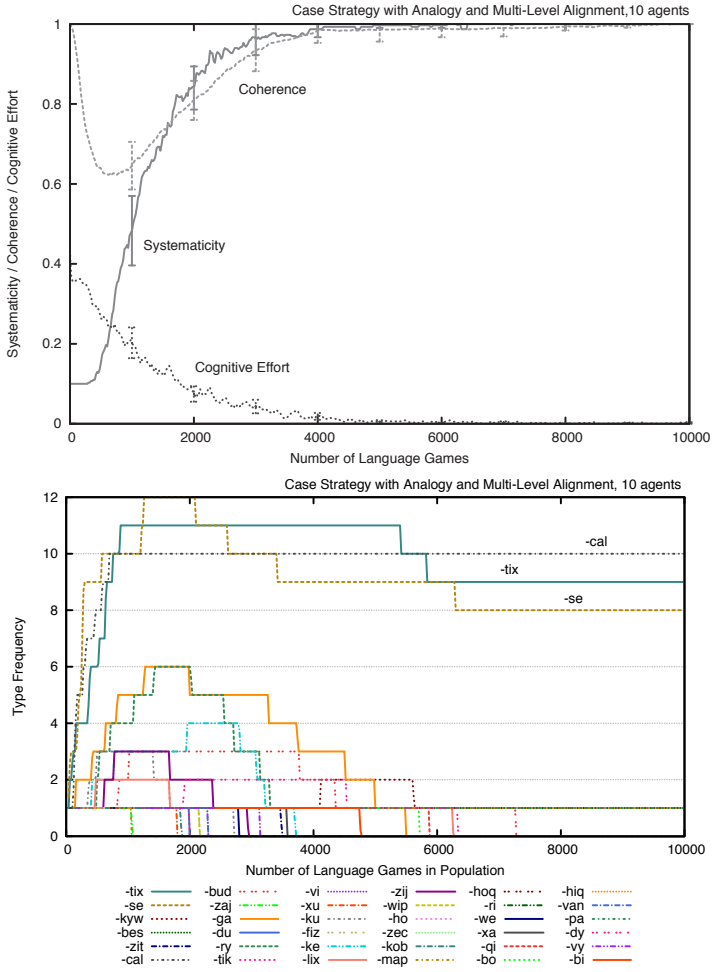

Fig. 13.14 Top: The agents develop a case grammar in order to reduce their cognitive effort in interpretation. As can be seen, effort indeed becomes minimal after 6000 games. The coherence measure shows that all agents converge on the same preferences in how they mark event structure, whereas the systematicity measure shows that the agents succeed in developing a systematic, non-arbitrary case language. The bottom graph shows the type frequency of different case markers over time, as assumed by a single agent. In the beginning there is fierce competition, but gradually three case markers become dominant.

involve not only actions, but also the participants of those actions, how they can be categorized in terms of semantic roles, and so on. After 1000 games, however, coherence starts increasing again in the population until the maximum value is reached after 6000 games. Finally, the systematicity measure shows whether the agents develop an arbitrary mapping between participant roles and case marker, or whether the mapping becomes systematic across constructions. Here again, the agents succeed in developing a fully systematic case system.

Crucially, there are many ways in which agents can build semiotic networks that integrate their visual perception of actions, affordances of objects and the linguistic conventions for expressing actions. As a result, each emergent language in the experiment evolves a different case system. Figure 1.15 compares the two most frequent case markers of two languages that evolved independently of each other using the same experimental set-up. The left marker *-pav* is the same one as shown in Figure 1.13, and covers roles that would be classified by linguists as 'Patient', 'State', 'Source' and 'Destination'. The marker *-pav* thus functions as a non-agentive case in its language. The right marker *-cat*, on the other hand, is a more agentive-like category, including participant roles such as a 'mover' and an 'approacher'. However, it overlaps with *-pav* in that it covers states, such as standing or being visible. Finally, *-cat* also involves 'touch' (i.e. visual overlap between two participants).

If linguists would be confronted with the utterances produced by the agents of each language, they would compare the two languages by drawing a semantic map

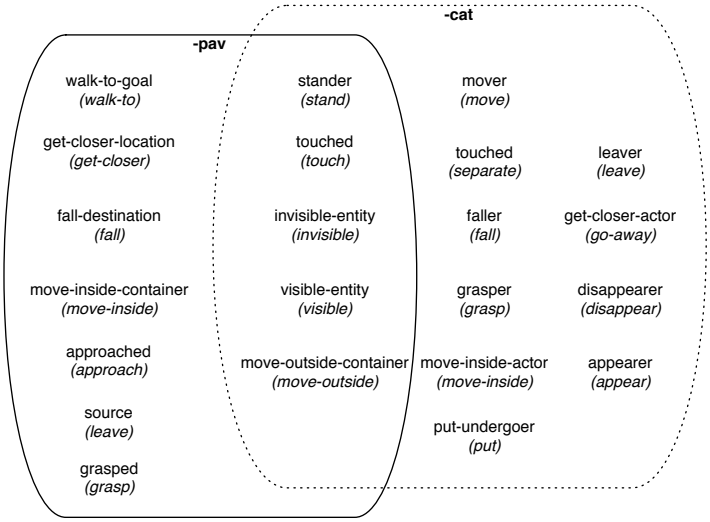

Fig. 13.15 This diagram compares the two most frequent case markers in two independently evolved languages. As can be seen, they carve up the space of participant roles in different, but overlapping ways.

similar to the one shown in Figure 1.16. Such semantic maps show that even though each language organizes the event structure of actions in its own way, a coherent 'conceptual space' can be drawn that shows recurrent patterns in the emergent case systems. Universal trends in language can thus be attributed to the distributed processes whereby language users shape and reshape their language.

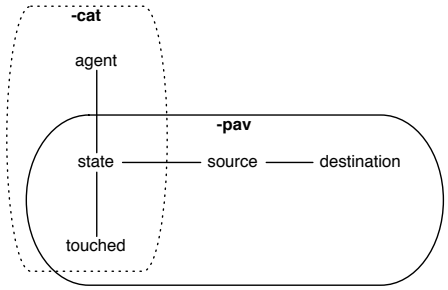

Fig. 13.16 Just like natural languages, the emergent case languages can be compared to each other using semantic maps.

13.4 Conclusions

Describing and requesting actions is one of the primary functions of language. But human languages are very different from programming languages or command-based interfaces to consumer devices. Actions and the roles of participants in actions are conceptualized in a language-dependent way and these are expressed using a variety of lexical and grammatical resources. We have argued that action language relies on rich semiotic networks that relate visual appearances, proprioception, motor control, and language. These networks are built up progressively through a diverse set of interactions with the world, internal simulations, and verbal interactions. Some evolutionary language game experiments were discussed that try to operationalize these ideas on autonomous humanoid robots. Although important results could already be achieved, it is clear that grounded language is a highly complex skill which requires still further research and many more case studies.

Acknowledgements

This research was conducted at the Sony Computer Science Laboratory in Paris and the Humboldt University neurorobotics group in Berlin. It was partly financed by the European FP7 project ALEAR.

References

Beuls K, van Trijp R, Wellens P (2012) Diagnostics and Repairs in Fluid Construction Grammar. In: Steels L, Hild M (eds) Language Grounding in Robots, Springer, New York

Feldman J, Narayanan S (2004) Embodied meaning in a neural theory of language. Brain and Language 89:385–392

Fillmore C (1968) The case for case. In: Bach E, Harms R (eds) Universals in Linguistic Theory, Holt, Rinehart and Winston, New York

Fujita M, Kuroki Y, Ishida T (2003) A small humanoid robot sdr-4x for entertainment applications. In: Advanced Intelligent Mechatronics, 2003. AIM 2003. Proceedings. 2003 IEEE/ASME International Conference on, vol 2, pp 938–943

Gerasymova K, Spranger M (2012) Learning Temporal Language: a Case Study for Russian Aspect. In: Steels L, Hild M (eds) Language Grounding in Robots, Springer, New York

Gibson J (1979) The Ecological Approach to Visual Perception. Houghton Mifflin, Boston

Haspelmath M (2003) The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In: Tomasello M (ed) The New Psychology of Language, vol 2, Lawrence Erlbaum, Mahwah, New Jersey, pp 211–242

Hild M, Siedel T, Benckendorff C, Thiele C, Spranger M (2012) Myon, a New Humanoid. In: Steels L, Hild M (eds) Language Grounding in Robots, Springer, New York

Höfer S, Spranger M, Hild M (2012) Posture Recognition Based on Slow Feature Analysis. In: Steels L, Hild M (eds) Language Grounding in Robots, Springer, New York

Oudeyer PY, Kaplan F, Hafner V (2007) Intrinsic motivation systems for autonomous mental development. IEEE Transactions on Evolutionary Computation 11(2):265–286

Pasemann F, Rempis C, von Twickel A (2012) Evolving Humanoid Behaviours for Language Games. In: Steels L, Hild M (eds) Language Grounding in Robots, Springer, New York

Rizzolatti G, Craighero L (2004) The mirror-neuron system. Annual Review Neuroscience 27:169–192

Schilling M (2012) Grounded internal body models. In: Steels L, Hild M (eds) Language Grounding in Robots, John Benjamins Pub, Amsterdam

Spranger M, Pauw S, Loetzsch M, Steels L (2012) Open-ended Procedural Semantics. In: Steels L, Hild M (eds) Language Grounding in Robots, Springer, New York

Steels L (2003) Language re-entrance and the 'inner voice'. Journal of Consciousness Studies 10(4-5):173–185

Steels L (2012) Grounding Language through Evolutionary Language Games. In: Steels L, Hild M (eds) Language Grounding in Robots, Springer, New York

Steels L, Spranger M (2008a) Can body language shape body image? In: Bullock S, Noble J, Watson R, Bedau MA (eds) Artificial Life XI, The MIT Press, Cambridge Ma

Steels L, Spranger M (2008b) The robot in the mirror. Connection Science 20(4):337–358

Steels L, Spranger M (2012) Emergent mirror systems for body language. In: Steels L (ed) Experiments in Cultural Language Evolution, John Benjamins, Amsterdam

Steels L, De Beule J, Wellens P (2012) Fluid Construction Grammar on Real Robots. In: Steels L, Hild M (eds) Language Grounding in Robots, Springer, New York

Talmy L (1988) Force dynamics in language and cognition. Cognitive Science 12:49–100

van Trijp R (2012) The evolution of case systems for marking event structure. In: Steels L (ed) Experiments in Cultural Language Evolution, John Benjamins, Amsterdam