Connection Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713411269

The robot in the mirror

Luc Steels [ab]; Michael Spranger [b]
[a] Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium [b] Sony Computer Science Laboratory, Paris, France

Online Publication Date: 01 December 2008

To cite this Article Steels, Luc and Spranger, Michael(2008)'The robot in the mirror',Connection Science,20:4,337 — 358
To link to this Article: DOI: 10.1080/09540090802413186
URL: http://dx.doi.org/10.1080/09540090802413186

Connection Science
Vol. 20, No. 4, December 2008, 337–358

The robot in the mirror

Luc Steels[a,b]* and Michael Spranger[b]

[a]*Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium;* [b]*Sony Computer Science Laboratory, Paris, France*

(*Received 13 February 2008*)

Humans maintain a body image of themselves, which plays a central role in controlling bodily movement, planning action, recognising and naming actions performed by others, and requesting or executing commands. This paper explores through experiments with autonomous humanoid robots how such a body image could form. Robots play a situated embodied language game called the Action Game in which they ask each other to perform bodily actions. They start without any prior inventory of names, *without* categories for visually recognising body movements of others, and *without* knowing the relation between visual images of motor behaviours carried out by others and their own motor behaviours. Through diagnostic and repair strategies carried out within the context of action games, they progressively self-organise an effective lexicon as well as bi-directional mappings between the visual and the motor domain. The agents thus establish and continuously adapt networks linking perception, body representation, action and language.

Keywords: body image; self organisation; lexicon acquisition; embodied language games

1. Introduction

The term body image refers to a representation or collection of representations that embodied agents must maintain in order to move about in the world, plan and execute action, perceive and interpret the behaviours of others, build and search an episodic memory, and understand or produce language about action, for example commands. The subject of body image has received great attention lately, particularly in the neurological literature, because of puzzling phenomena such as phantom limbs, mirror box experiments, unusual pain, out of body experiences, etc. (Ramachandran and Hirstein 1998; Rosenfield 1988; Blanke and Castillo 2007). These disorders and experiments make it quite clear that human body image is not a simple, fixed innately given internal representation, but a dense network of representations that forms in development and continues to change throughout life.

The question addressed here is how the representations of body image develop and become co-ordinated. There is no simple answer to this question, because different representations use different sources and they are changing on different time-scales. For example, the child's body

*Corresponding author. Email: steels@arti.vub.ac.be

ISSN 0954-0091 print/ISSN 1360-0494 online

DOI: 10.1080/09540090802413186
http://www.informaworld.com

model must be changing as he/she grows up and arms and legs are becoming longer or further removed from the ground. The actual motor-body-image is changing every time some body part is moved or being moved by an external force. In the neuroscience literature, a lot of discussion of the relation between body image and motor behaviour is framed within the context of mirror neurons (Rizzolatti, Fadiga, Gallese and Fogassi 1996). Their discovery has shown that there are neuronal cells that become active both when a particular action is seen (body image of other) and when the action is performed (motor behaviour of self). The relation of these findings to understand body language and its origins has been recognised and explored (Rizzolatti and Arbib 1998).

In earlier robotics work, body related representations were usually fixed and relations between them could therefore be programmed by hand, as happened, for example, in the pioneering robot Shakey (Nilsson 1984). However, today the cognitive robotics community recognises that body representations are highly complex, that the relations between body image, motor control and language are dynamic, and that the conceptual and linguistic repertoires for talking about the body are language-specific. So it is no longer considered to be possible to pre-program body representations or body language. A number of studies show how body-related representations get bootstrapped and remain adapted as the robot interacts with the world and others. Particularly the relation between visual representations and recognition of bodily action and the bodily action itself has been intensely studied through research on imitation, which usually presupposes that agents are able to recognise and enact movement primitives and then combine them to learn new movements by demonstration (Billard 2002; Matarič 2002; Demiris and Johnson 2003).

This paper discusses experiments with the Sony QRIO humanoid robot (Gutmann, Fukuchi and Fujita 2005) in order to understand better the nature, function and operation of body image. Our goal is to come up with precise working models that have a dual function: (i) they suggest predictive models for future psychological, neurological and neuronal observation, experimentation, and theorising, and (ii) they suggest avenues for engineering robots that can acquire and adapt embodied communication systems. We have built further on existing results in the self-organised formation of grounded language and meaning, validated in computational and robotic experiments (Steels 2003). Earlier work has already shown how a lexicon can form through language games using a lateral inhibition dynamics (Steels 1995), how perceptually grounded categories, such as colour categories, can form and be co-ordinated among individuals (Steels and Belpaeme 2005; Vogt and Divina 2007), and how a spatial vocabulary can form in a population of embodied agents based on perceptually grounded spatial models of the world and perspective reversal (Steels and Loetzsch 2008). There has also been prior work on how robots can learn a vocabulary for talking about actions being taught by a human experimenter (Steels and Kaplan 2001; Roy and Pentland 2002; Dominey 2003). All these experiments are based on the idea that language, including the conceptualisations of the world expressed by language, can be viewed as a complex adaptive system that is constantly shaped and reshaped by its users in order to achieve successful communication. Here we apply the same principles to examine the formation and role of body image, including its role in enabling language about the body. Similar to research by Triesch, Jasso and Deak (2007), we assume that there is nothing special about mirror neurons but that neurons become mirror neurons because they take on a particular role in networks.

Concretely, we discuss two experiments. The first one (called the mirror experiment) is viewed as preparatory. Robots learn the bi-directional mapping between visual-body-image and motor-body-image by standing before the mirror, executing actions, and observing the visual-body-images that they generate. Once a group has each learned this mapping, they play language games settling on names for these actions. The language games that we study are action games, in which the speaker asks the hearer to do an action and the game is a success if the hearer performs the requested action. If the game fails, the speaker repairs the communication by performing the action himself. In a second experiment (called the body language experiment), robots do not learn the bi-directional mapping between visual-body-image and motor-body-image through a mirror but through the language game itself. They start without knowing the relation between visual-body-image and motor behaviours and without having a pre-defined lexicon. We will show that not only is the bi-directional mapping between visual-body-image and motor-body-image emerging but also the lexicon for naming bodily actions and their visual appearance. Both co-evolve and bootstrap each other.

The first two sections of the paper define some relevant terminology and the problems that need to be solved to achieve grounded language about the body and actions of the body. Then the mirror experiment and its results are derived. Next, the body language experiment is introduced and results analysed. We conclude with directions for future work.

2. Terminology

A body consists of physical components (arms, fingers, legs, torso) that are connected and hence have three-dimensional angular positions with respect to each other. We need to make a distinction between the motor aspects, which concern the actual control and sensing of bodily movement, the visual aspects, which concern the way that bodies are perceived, and the cognitive aspects, which concern the way that the body is conceptualised and being talked about. We argue that there is not a single body representation but different ones from each of these points of view. Moreover, agents not only build and maintain representations of themselves, but also, to some level of detail and accuracy, of other agents.

More concretely, we follow the standard robotics literature and distinguish the following aspects on the motor side:

- *Body model.* This is a three-dimensional representation of the body that contains information about which components the body has, how they are attached to each other, where the motors are located, where proprioceptive sensors are located, and the size and extent of each component. In the current experiment, this body model is fixed and pre-programmed.
- *Motor control program.* This is a sequence of desired positions of the various motors and hence of the bodily components that they articulate. They are usually expressed in terms of angles.
- *Motor command stream.* This is a sequence of signals to the motors to start moving in a certain direction with a particular acceleration in order to achieve a certain position (see Figure 1).
- *Proprioception.* These are various sensory streams coming from the motors or coming from other sensors attached to the body (gyroscope, acceleration sensor, temperature sensing, etc.) that help to determine how far a motor control program effectively achieves its purpose.
- *Motor behaviour.* This is an ensemble of motor control programs, integrating motor commands, proprioception and feedback loops in order to achieve a particular behaviour such as walking, picking up an object, waving, or raising the left arm. The behaviours may integrate additional non-motor sensory sources and they may exploit the physical properties of the body to achieve their effect (Steels 1994; Pfeifer, Bongard and Grand 2007). For example, obstacle avoidance could integrate infrared reflectance of objects and rely on the effect of bumping against an obstacle to generate another (random) body position. The behaviours needed in the present experiment typically require about 20 motor commands for reaching a body posture and various control and feedback loops between them.

The MOTOR-BODY-IMAGE combines information derived from the motor system with the body model and is changing continuously as the body moves (Figure 2, middle image). A distinction needs to be made between the EXPECTED MOTOR-BODY-IMAGE, which combines information of the motor control program and the body model, and the SENSED MOTOR-BODY-IMAGE, which combines information from proprioception and the body model. Discrepancies arise all the

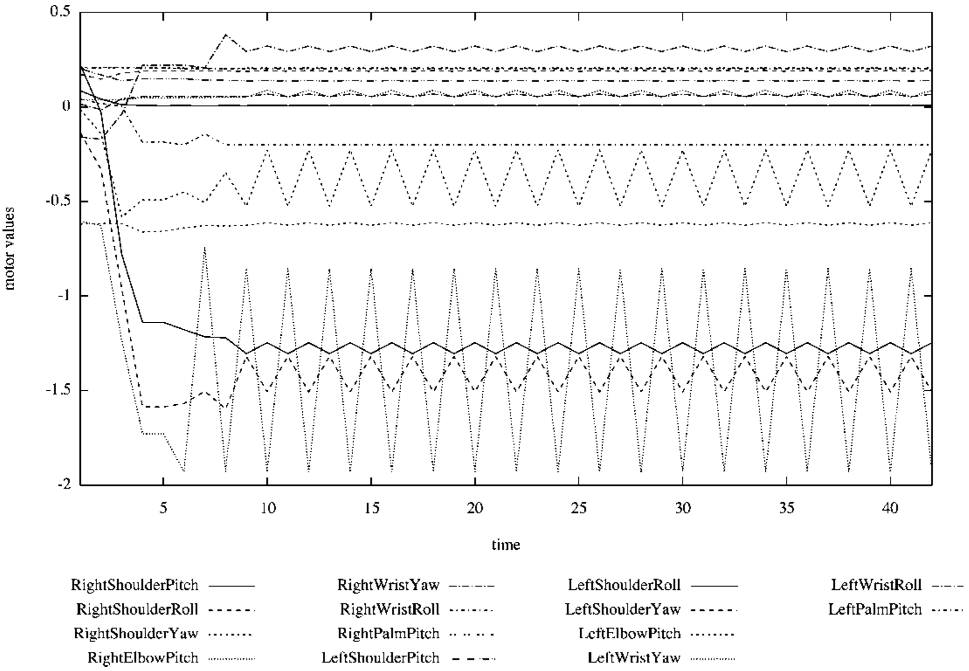

Figure 1. Example of motor control data streams for a waving behaviour. The motor activity of raising the arm and then the oscillatory movements of arm joints are clearly visible.

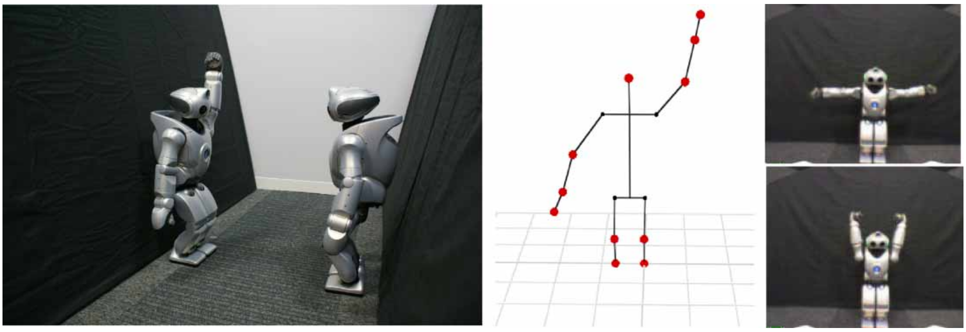

Figure 2. The experimental set-up, where two humanoid robots face each other and play an Action Game, asking each other to achieve a certain posture, such as raise the left arm, or stretch out both arms. The middle image shows the expected motor-body-image, constructed by integrating the motor command stream and the motor body model. The right image shows two example postures as seen through the camera of another robot.

time between the expected and sensed motor-body-image, which in normal circumstances are corrected as part of motor behaviours. For example, the motor control program may issue a command to the motors controlling the lower arm to reach a certain angle, but if the arm is blocked by an obstacle, then proprioceptive feedback alerts the motor control behaviour that the motor control program cannot achieve its goal and hence that another way must be used to reach the desired target. The SIMULATED MOTOR-BODY-IMAGE uses a detailed physical model of the body and the physical behaviours of the robot to reconstruct the expected and sensed motor-body-images given a particular state of the world and information about which behaviours are active.

We follow the standard vision literature that makes the following distinctions in the visual domain:

- *Actual visual image*. This representation is based on the raw visual perception to be treated by a battery of standard image processing algorithms in order to do foreground/background distinction, segmentation, feature extraction, model reconstruction and tracking. We define the SENSED VISUAL-BODY-IMAGE as the outcome of visual analysis that pertains to the perception of a robot body in the environment (self or other).
- *Expected visual image*. It is well known that pure bottom-up vision is very limited and therefore needs to be complemented with the projection of top-down expectations derived by extrapolating from the past state using Kalman filtering, for example. In this present experiment, the predicted visual image is constantly compared with the actual visual image and the prediction model continuously updated to improve it. We define the EXPECTED VISUAL-BODY-IMAGE as being the outcome of this predictive process.
- *Event structure*. In the case of the body, we are dealing with a set of objects that are moving in co-ordinated ways. It is therefore possible to apply image analysis techniques for detecting events and their hierarchical structure and thus develop a VISUAL BODY-EVENT-STRUCTURE (as in Siskind 2001 or Steels and Baillie 2003). This analysis must be based on an inventory of primitive events and how they are combined into more complex events.
- *Simulated visual image*. If a simulated motor-body-image exists, then it is possible to compute a SIMULATED VISUAL-BODY-IMAGE and a SIMULATED BODY-EVENT-STRUCTURE, based on an accurate mapping from simulated behaviour to its visual appearance. Even if one starts with accurate initial conditions, a simulation would deviate very quickly from reality, partly because sensing is never accurate enough and it is impossible to take all physical aspects into account.

For all these representations, we can make a further distinction between body image (visual or motor) of SELF and OTHER, whether sensed, expected, or simulated, even though one robot can obviously never have a sensed motor-body-image of another robot and the sensed visual-body-image a robot has of himself is always very partial. Each motor- and visual-body-image is located within the coordinate system of a particular robot and if a robot perceives another robot, it is always from the perspective of his own coordinate system.

Clearly all these body-related representations play a role in body language. Body language is the agent's inventory of words and grammatical constructions for describing the body and its actions. It is mapped on to the motor-body-image and motor behaviours and on to the visual-body-image and the perceived body-event-structures through the intermediary of the CONCEPTUAL BODY-IMAGE and the CONCEPTUAL ACTION-STRUCTURE. In order to play an action game, the speaker requesting the action must understand the relation between the visual-body-image that he wants to see achieved and the motor-body-image. He must know which action can achieve it, which behaviours need to be done for this action and how this action is expressed in language. The listener must parse the sentence to recover the intended action and the body parts it involves and he must map them to his own behaviour. To recognise whether the hearer performed the requested action, the speaker must be able to predict the expected visual-body-image of the other and compare it with the actual observed body-image of the other.

The different conceptual and linguistic inventories are only partially shared between agents and are surely language-dependent. For example, many languages (such as Serbo-Croatian) do not lexicalise the distinction between arm and hand, foot and leg, finger and thumb, or finger and toe. Actions are categorised more subtly and differently in different cultures and languages (Talmy 2000). Natural languages are therefore a key to understanding how a particular language community conceptualises the body and actions carried out by the body, and there is not a simple one–one relation between the cognitive conceptualisation of the body and the way it is grounded, just as there is no such simple relation between colour categories and the physics and visual processing of colour (Steels and Belpaeme 2005).

The great advantage of using real robots is that all these different representations can be operationalised and made entirely concrete and grounded in real sensing and vision, even in real time. The motor-body-image (sensed and expected) is sensed and processed in real time as the robot moves about. The visual-body-image (sensed and expected) is derived through image processing starting from the raw visual image captured by the cameras. The simulated motor- and visual-body-images are the most difficult to compute because they require very detailed and realistic physical models, are computationally very intensive, and require a very realistic model of the world. It is not necessary for actual robot behaviour and it will not be considered in the present paper (even though it plays a central role in many theories of body image, see, for example (Jeannerod (2001) and Feldman (2006)).

In several of our earlier experiments (see, e.g. Steels and Loetzsch 2008) robots were already equipped with mechanisms for perspective reversal and this will also play a crucial role in the present experiment. Perspective reversal means that one robot tracks the position of another robot in an ongoing interaction using his own self-centred coordinate system and then, based on his own visual perception, computes what the world looks like from the viewpoint of another robot. This reconstruction is necessarily error-prone, because a robot can never totally accurately perceive the other robot's position, and incomplete, because the other robot might see objects or faces of objects that are invisible to the first robot. Nevertheless, perspective reversal is a crucial ability for language, both as speaker and hearer. For example, if one robot says 'point right', this is ambiguous. It could mean to the right of the speaker or to the right of the hearer. If he says 'raise your right hand' this is not ambiguous because it is the arm to the right from the perspective of the hearer. Here we apply perspective reversal to the motor-body-image. The observer must project the motor-body-image of the other on to that of his own, so that the image generated by the other can be compared with the one he would produce himself with the same movement. If one robot stands in front of another, this means, for example, that movement of the arm to the left of the body projects to movement of the right arm. Robots perform this kind of geometric transformation of the image before feature detection. Watching oneself in the mirror is very special, because the movement of a left or right body part appears as a movement of the same body part, so that no perspective reversal is to be carried out.

This paper focuses on one key question: How can visual-body-image and motor-body-image become co-ordinated and what role does language play in this process? In order to get any concrete experimentation done at all, we provide robots with a body model and focus on reaching body postures (such as 'raise your left arm') as opposed to ongoing dynamic body behaviours (such as 'wave your right arm'). The expected and sensed self motor-body-images are derived in real time by the robot when an action is being executed, by combining the motor control streams and proprioceptive sensing streams with the body model. Robots are equipped with a sophisticated vision system that derives visual-body-images using standard state-of-the-art techniques from computer vision. There is foreground/background separation, segmentation, feature extraction, and object recognition (see Figure 3 later; see Spranger (2008) for a similar system). Top-down expectations are also computed using standard computer vision techniques, so that robots can use an expected visual-body-image for improving the sensed visual-body-image. There is no simulated visual- or motor-body-image in the present experiments (without making any claims about their necessity or psychological plausibility). As we focus only on static body postures, there is no event-recognition system either, even though we use it in some of our other experiments. Crucially, what the robots do *not* know and have to learn in the experiments is the relation between visual-body-image and motor-body-image. They are able to maintain and use a lexicon in the form of associations between words and visual images of body posture or actions, but they do *not* know *a priori* which words are used for which actions or for which body images.

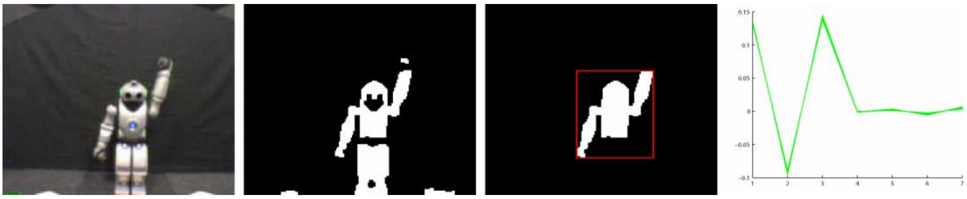

Figure 3. Aspects of visual processing. From left to right we see the source image, the foreground/background distinction, the result of object segmentation (focusing on the upper torso) and the feature signature of this posture (explained in the text).

We now report two experiments. The first experiment is a preparatory experiment in which a robot co-ordinates visual-body-image and motor-body-image by standing before a mirror, executing actions and observing the visual images of his own body that these actions generate. They play a language game called the Action Game, in which robots ask each other to perform actions in order to achieve a certain posture. The second stage consists of an experiment in which two robots interact with each other and perform a similar kind of co-ordination, but now using language instead of the mirror. As in our other language game experiments, there can be many more agents than robot bodies. We download at the start of a particular interaction the cognitive state of a particular agent in the on-board processors of the robot, and upload the altered state after an interaction. Moreover, to speed up experimentation we have also recorded and archived the full visual and motor streams for a large number of interactions so that we can do off-line experiments with recorded data.

3. The mirror experiment

In the first experiment, each robot stands before a mirror in order to acquire the relations between his own motor-body-image and (a mirror image) of his own visual-body-image (see Figure 4).

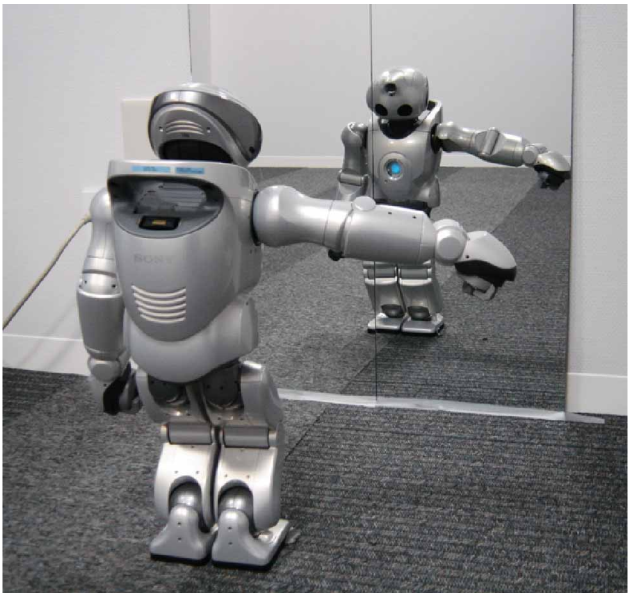

Figure 4. A humanoid robot stands before a mirror and performs various motor behaviours, thus observing what visual-body-images these behaviours generate.

Once the robots have acquired these relations, they engage in language games, asking each other to take on a certain posture. As all robots have exactly the same body, a robot can use a visual-body-image of himself in order to categorise the body image of another robot, after perspective reversal. It turns out that given these assumptions, well understood techniques from language dynamics can solve the task easily.

The internal cognitive system of the robots can be fruitfully analysed in terms of a network that is linking sensory images with motor behaviour (see Figure 5). All the links in these networks have weights that change over time based on the outcome of a game. The network has the following nodes:

- *Sensory experience*. The vision system not only segments the robot body against the background (Figure 3) but also extracts a series of features that characterise the shape of the robot body. As these features have to be shift and scale invariant, we have used normalised central moments as defined in the Appendix. Values for each of these features are combined into a feature vector that constitutes the sensory experience of a particular perceived body image at a particular moment in time. When the robots are watching another robot, they perform perspective reversal on the visual image, which means that they perform a geometric transformation. As the robots are always standing in front of each other in this experiment, this means a 180-degree transformation.
- *Prototypes*. The specific visual feature vectors must then be classified in terms of prototypical views on body postures, as defined more formally in the Appendix. A prototype is a feature vector with the same dimensions as the sensory experience and with a typical point as well as minimum and maximum accepted deviations from this point. The best matching prototype is found by distance computation and then the prototype is adjusted to assimilate better the new experience. When no prototype is matching, a new one is created, with the sensory experience being the first seed. An example of a prototype is shown in Figure 3, right-most image. The seven features are displayed on the x-axis and their values (scaled) on the y-axis. The graph shows a band connecting minimum and maximum boundaries of each feature values for this prototype (which is very thin in this particular case).
- *Postures*. Each prototype is linked to a node for a posture on the one hand and to a motor behaviour on the other hand. These posture nodes function like mirror neurons, in the sense that they become active both when the robot sees a body image of another robot (possibly

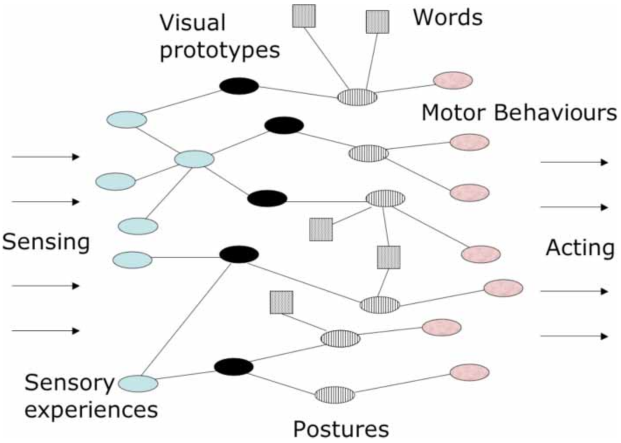

Figure 5. Network linking sensory experiences, visual prototypes of postures, nodes for postures acting as mirror neurons and nodes triggering the motor behaviour that achieves the posture. Nodes for words (shown as squares) are associated with the posture nodes.

himself) that matches best with a prototype of the posture *and* when the robot wants to perform the motor behaviour that generates this posture.

- *Motor behaviour*. Each posture node is linked to a motor behaviour that can achieve that particular posture when executed. The relation between a posture and a motor behaviour is acquired in this experiment by prior kinesthetic teaching. For each posture, the robot's body is brought manually into that posture by a human experimenter. The robot records the critical angles with which to construct the motor control program, the expected proprioception, and possibly additional controls and feedback loops needed to achieve the posture.

The links between sensory experiences and their prototypes are established using standard prototype-based machine learning techniques and the links between postures and motor behaviours are based on kinesthetic teaching. But how are the links acquired between the visual prototypes of a posture and the node for this posture? This is where the mirror experiment plays a crucial role because it generates the conditions under which associative learning can take place: a robot stands before the mirror, selects a posture and activates the corresponding motor behaviour. This motor behaviour generates a sensory image and hence a sensory experience categorised with a particular prototype. Through standard Hebbian learning (which enforces the connection between nodes that are simultaneously active, see Bishop 1995) the link between the prototype and the posture is established and progressively enforced.

Once this network is established, the development of a shared lexicon is straightforward based on the following well-known Naming Game dynamics (Steels 1995, 2005). All agents maintain an associative memory with weighted links between names and postures. In requesting a name, the speaker chooses the name connected to the posture with the highest weight and, in interpreting a name, the listener chooses the posture connected to this name with the highest weight. When no name exists yet for a posture, the speaker can invent a new name by generating a random combination of syllables and storing it in his network. When a listener encounters an unknown name, he guesses the meaning based on the repair actions of the speaker and then expands his own associative memory. As we are dealing with a distributed population of autonomous agents, it is entirely possible, in fact unavoidable, that one agent invents a new name even though there is already a name circulating in the population, but it is known that a lateral inhibition dynamics will easily co-ordinate the lexicon. Lateral inhibition means that after a successful game the used associations are increased and competing associations (alternative names for the posture in the case of the speaker, alternative postures for the name in the case of the hearer) are decreased.

A more precise definition of the script for the Action Game used in the mirror experiment follows:

(1) The speaker chooses a posture from his inventory of postures.
(2) The speaker retrieves the name with the highest weight and transmits that to the hearer.
(3) The hearer retrieves the posture with the highest weight in his own lexicon and evokes the motor behaviour associated with it.
(4) The speaker observes the posture adopted by the hearer and checks whether it is classified by the prototypical visual-body-image of the posture he had originally chosen.
(5) If this is *not* the case, the speaker signals failure.
    - The speaker activates his own motor behaviour for achieving this posture in order to repair the communication.
    - The hearer perceives the posture and matches that against his own inventory of prototypical posture images. The hearer then extends his lexicon connecting the nodes for this prototype, the posture and the name of the posture.

(6) Otherwise he signals success and both agents consolidate their inventories using lateral inhibition dynamics, i.e. increasing the weight of the connections that were used and decreasing their competitors.

This script is executed by randomly chosen agents in pair-wise interactions. Agents take turns being speaker and hearer. Typical results of a sequence of consecutive games are shown in Figure 6. The graph plots the global behaviour of the population *after* each individual has co-ordinated motor behaviour and visual-body-image through the mirror. One hundred per cent success is reached easily after about 3000 games. Already after 1500 games there is more than 90% success. The graph shows the typical overshoot of the lexicon in the early stage as new words are invented in a distributed fashion followed by a phase of alignment as the agents converge on an optimal lexicon.

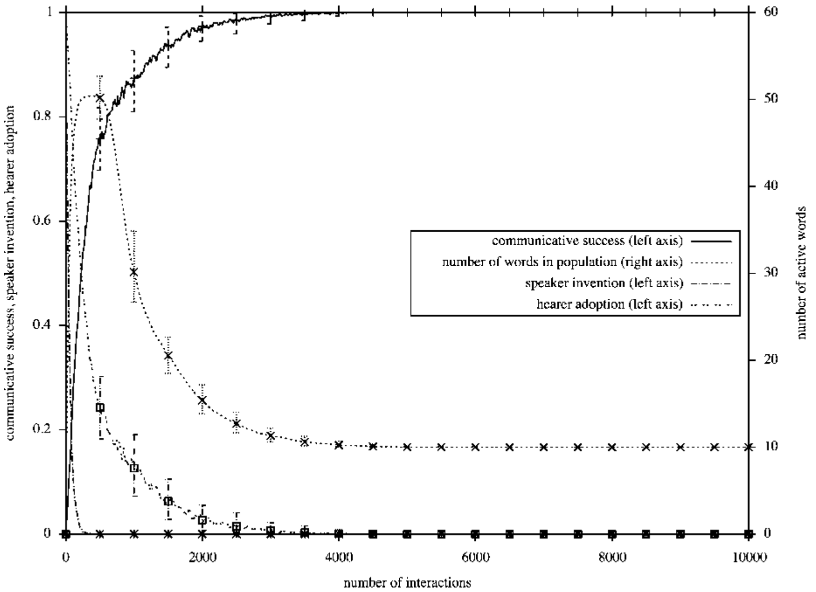

Figure 6. Results of the Action Game played by a population of 10 agents for 10 postures. Robots have first co-ordinated their own visual-body-images and own motor behaviours by standing before a mirror and observing their visual appearance after executing a particular motor behaviour. The number of language games (in this case 5000 games) is shown on the x-axis, each game involving two agents. The running average of communicative success and average lexicon size as well as invention and adoption frequency are shown on the y-axis.

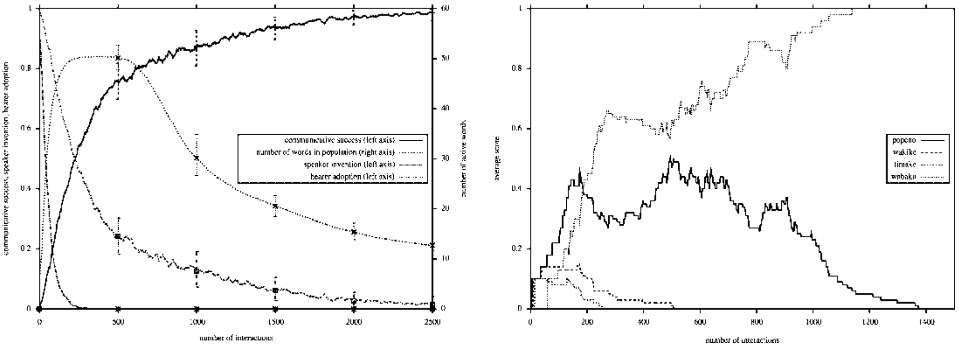

Figure 7. This figure zooms in on the first 2500 games. To the left we see the global behaviour showing a close-up of the same results as in the previous figure and to the right the average score for all the words competing for naming the same posture. The winner-take-all dynamics that co-ordinates the lexicon among the agents is clearly visible.

Figure 7 shows a snapshot of the semiotic dynamics for the first 2500 games. On the left y-axis we plot again the communicative success, lexicon-size, and invention and adoption frequency. On the right, the average score in the population for different words competing for the same meaning is shown. We see clearly that a winner-take-all situation arises after about 300 games, with one word dominating for naming this particular posture.

4. Language can mediate body image co-ordination

We have seen that once adequate links exist between the prototype for a visual-body-image of a posture and the motor behaviour that generates this posture (here captured in the posture nodes that act as mirror neurons), it is straightforward for a group of agents to self-organise a lexicon of names that can be used for describing a particular posture or as commands for asking another robot to achieve it, and thus for playing the Action Game. Now we turn to the question of whether robots could also co-ordinate visual-body-image and motor behaviours through language *without* using a mirror first. It turns out that the answer is positive, showing the potential of language to co-ordinate perception and behaviour among autonomous agents without any form of telepathy, central control or prior design. We do not claim that language is the only way in which autonomous agents can co-ordinate motor behaviour and visual body image, in fact any kind of co-operative game will do, but it is definitely one route. Moreover, we do not claim that we present here the most efficient solution, rather we want to show the simplest solution in terms of cognitive complexity of the mechanisms used by the agents. If agents become ‘smarter’, for example by being able to guess what action another agent intends him to perform because of the shared context and ongoing co-operative behaviour, then they have stronger heuristics to explore the search space.

The agents play basically the same language game as before. They take turns asking each other to achieve a certain posture, such as ‘raise left arm’ or ‘put up both arms’, but the speaker no longer repairs miscommunication by performing the motor behaviour himself that he originally requested. The crucial difference is now that the networks built up by the agents do not contain nodes for postures, hence no ‘mirror neuron’ nodes exist linking posture with motor behaviour on the one hand or with visual prototypes of the posture generated by this motor behaviour on the other. Instead we get networks as shown in Figure 8. Nodes for words are now linked with visual prototypes of body images on the one hand or with motor behaviours on the other. As there is no

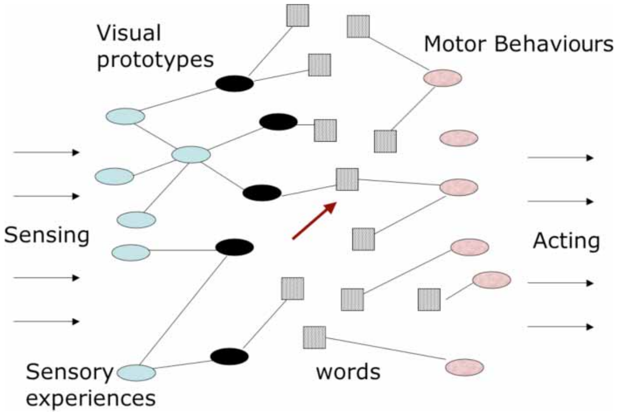

Figure 8. Networks built by agents in the second experiment without the mirror. There are no nodes for postures linking visual prototypes with motor behaviours. Nodes for words (indicated by squares) are either linked to body image or to motor behaviour or both.

direct link between visual prototypes and motor behaviours, there is not necessarily agreement either. Thus, there might be a word W linked to the visual prototype of a posture P, which is at the same time linked to a motor behaviour M that generates a different posture. Visuo-motor co-ordination arises only when the same word names both the visual prototype of a posture and the motor behaviour that generates it. This is the case only for the name pointed at by an arrow in Figure 8. Nevertheless, agents can already have success in communication even without this kind of visuo-motor co-ordination, as explained shortly.

The script for the Action Game in this experiment is as follows:

(1) The speaker chooses randomly a visual prototype from his inventory of prototypes.
(2) The speaker retrieves the name with the highest weight associated with this prototype and transmits that to the hearer. If there is no name, then the speaker chooses randomly a motor behaviour and then the best name for this motor behaviour (if there is none he invents a new name for the visual prototype). With this step the hearer possibly introduces a new hypothesis about a connection between a visual prototype and a motor behaviour, even though this could be entirely wrong.
(3) The hearer retrieves the motor behaviour with the highest weight associated with this name in his own lexicon and evokes the motor control program associated with it. When the hearer has never heard this word before, he makes a random choice within his own inventory of motor behaviours and creates a new link, thus also introducing a new hypothesis.
(4) The speaker categorises the visual image generated by the hearer's movements using his inventory of prototypical visual-body-images. The speaker then checks whether this visual prototype is associated to the name of the prototype he originally chose in step 2.
(5) If this is *not* the case, the speaker signals failure.
(6) Otherwise, speaker and hearer have co-ordinated lexical associations and they can both reinforce this relation, i.e. they increase the score of the link between the name and the prototype (for the speaker) and the name and the motor behaviour (for the hearer).

Note that there is much less feedback given compared with the earlier game because in the mirror experiment the speaker would carry out the 'correct' motor behaviour for the chosen posture to repair the communication, whereas now the speaker just signals failure. Note also that agents change the connections between words and visual prototypes only when they are speaker and they change the connections between words and motor behaviours only when they are hearer.

The results of this second experiment for a population of five agents and five postures are shown in Figure 9. The graphs show the global behaviour of the population focusing on the first 5000 language games. One hundred per cent success is reached after about 5000 games and stays stable (unless new postures get introduced and then the networks of all agents expand to cope with this). Already after 3000 games there is more than 90% communicative success. The graph shows the typical overshoot of the lexicon in the early stage as new words are invented and a phase of alignment as the agents converge on an optimal lexicon, which is now five words to name each of the postures. The frequency of invention and adoption is also shown and dies down as the lexicon stabilises. So this is already a very important result. Clearly agents are able to self-organise a lexicon of commands even if no prior lexicon exists and even if they are neither pre-programmed nor have been able to learn mappings from motor behaviours to visual postures before lexical self-organisation starts.

Communicative success does not necessarily mean that agents relate all visual prototypes 'correctly' to the corresponding motor behaviours. As explained shortly, communicative success is possible without full co-ordination. Nevertheless, Figure 10 shows that the 'right' correspondences progressively emerge. The figure plots the aggregated strength (y-axis) of the relation between visual prototypes (x-axis) and motor behaviours (z-axis) in this population over time. Aggregated

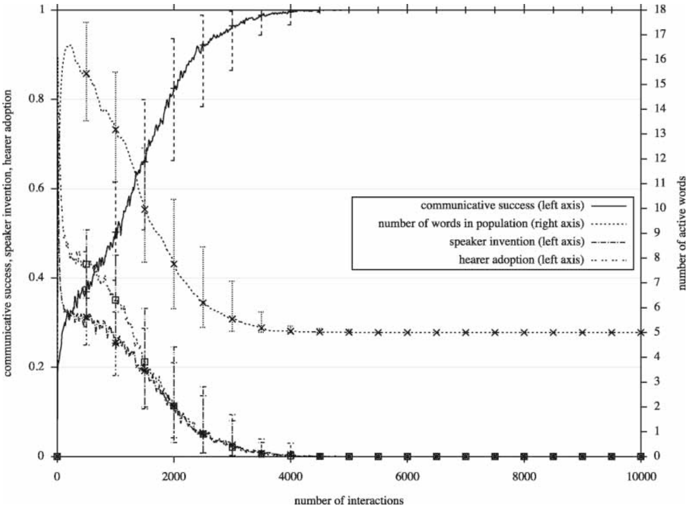

Figure 9. Results of the Action Game played by a population of five agents for five postures. Robots have *not* co-ordinated their visual-body-images and motor behaviours by using a mirror prior to playing the action game. The x-axis plots the number of language games. The running average of communicative success and average lexicon size as well as invention and adoption frequency are shown on the y-axis.

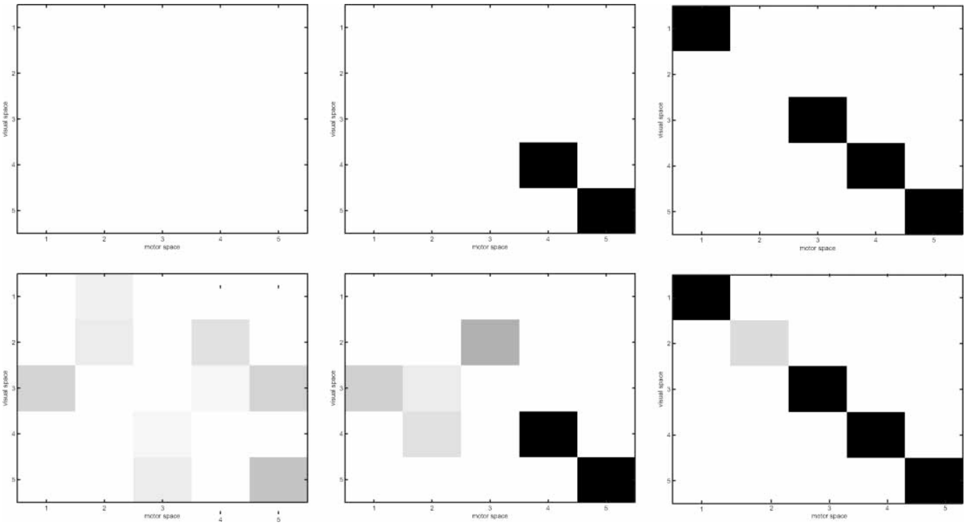

Figure 10. Relation between visual prototypes and motor behaviours for a population of five agents negotiating names for five postures. The diagrams show the aggregated strength of the relations by product (top row) and sum (bottom row) over time; the left column after 500 interactions, middle column after 1500 interactions and right column after 2500 interactions. The diagram shows that agents progressively achieve the correct mappings and agree on them. The strength of a connection is depicted with black meaning strongest possible connection and white no connection.

results for 100 experiments are shown with two diagrams. As explained in more detail in the Appendix, the first one (top row) shows the evolution of agreement among the agents (computed by taking the product of the strength over all different agents and all experiments) and the second one (bottom row) shows whether agents make any link at all (computed by taking the average strength of all relations between visual prototypes and motor behaviours for all agents for a series of experiments). If there is a single column between a particular visual prototype p_i and

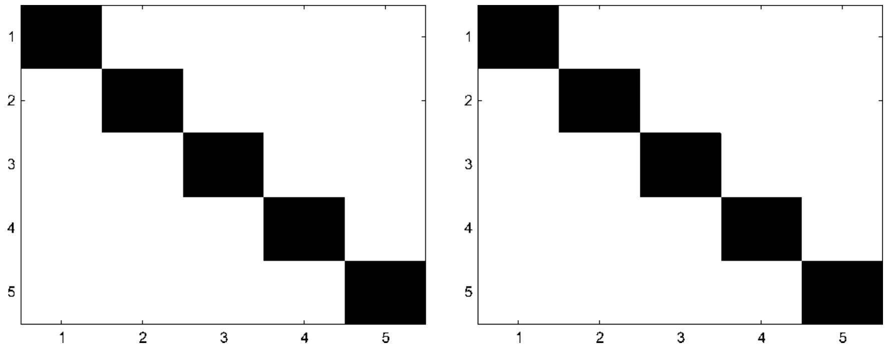

Figure 11. Relation between visual prototypes and motor behaviours for a population of five agents for five postures after they have reached a steady state. The diagrams combine the results of 100 different experiments, showing the aggregated strength (y-axis) of the relation between visual prototypes (x-axis) and motor behaviours (z-axis) as mediated by language. The left diagram reflects how many links are present (computed by taking the sum) and the right one how far agents agree (computed by taking the product).

its corresponding motor behaviour m_i with strength 1.0, then this means that all agents have this particular mapping (bottom row) and agree on it (top row).

Figure 11 shows the final steady state reached after 3000 games. We see that total coherence is reached with respect to visuo-motor co-ordination by all agents in all experiments. This result is remarkable because there appears to be a chicken and egg situation. If all agents have a correct association, then a lexicon can easily self-organise as shown in the mirror experiment. Alternatively, if there is a shared lexicon (for example pre-programmed) then we can imagine the agents learning a new association between visual prototype and motor behaviour easily. But how is it possible that agents co-ordinate their lexicons *and* acquire coherent associations between visual-body-image of other and motor behaviour of self, without any form of telepathic exchange or prior design, and without any nodes functioning as mirror nodes linking the two?

Let us examine first a situation arising in two-agent games at an intermediary stage (see Figure 12). Assume agent a_1 has associated the visual prototype p_2 with the word w_3 and motor

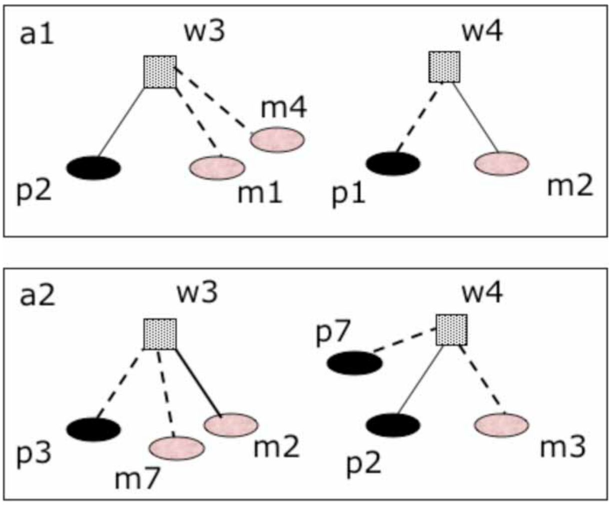

Figure 12. Successful games may arise between two agents even without visuo-motor co-ordination. This figure shows an intermediate state of two agents where different prototypes or motor behaviours are associated with the words. The word w_3 can be successfully used by a_1 to evoke m_2 by a_2 and the word w_3 can be successfully used by a_2 to evoke m_2 by a_1.

behaviour m_2 with the word w_4. Another agent a_2 has associated w_4 with p_2 and w_3 with m_2. The agents have in addition other associations with a lower strength. There is no visuo-motor co-ordination within each agent, assuming p_2 is the correspondent prototype for m_2, but nevertheless agents can successfully communicate. When a_1 wants to see p_2 he communicates w_3, and a_2 will perform m_2, which indeed reproduces p_2 for a_1. The same with turns taken is true for w_4, p_2 and m_2. So we see that success in the language game does *not* necessarily mean that each agent uses the same word both for recognising and for producing a particular posture. They can have perfect communicative success without visuo-motor co-ordination. The situation changes completely, however, when a third agent a_3 enters the population. He will be confronted with two competing words for achieving p_2: w_3 from a_3 and w_4 from a_2. Only one of them can survive the lateral inhibition dynamics, pushing also the other agents to choose one or the other word for p_2. So three or more agents can only communicate successfully when prototypes and motor behaviours are co-ordinated within each agent. Given this observation, we predict that in a population with only two agents no visuo-motor coherence can arise, despite successful communication. Indeed, this is borne out in experiments. Figure 13 shows the results of action games played between only two agents. They quickly reach communicative success but visual prototypes and motor behaviours are not properly co-ordinated within each agent. When we inspect the agent networks we see that sometimes only one or two motor behaviours are correctly linked to their corresponding visual prototype and sometimes there is even no visuo-motor co-ordination at all (as shown in Figure 14).

This analysis shows that the semiotic dynamics co-ordinating perception, action and language across agents can be seen as a relaxation process. There are two processes going on. The first one is the co-ordination of the lexicon with the damping of synonymy and homonymy as in any naming game. The second one is the co-ordination inside each agent of visual prototypes of postures and

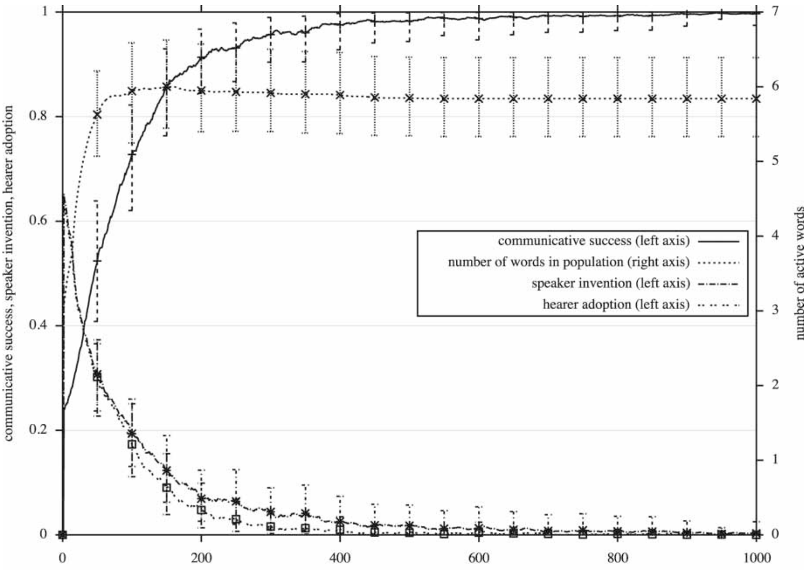

Figure 13. Results of the Action Game played by a population of two agents for five postures (1000 interactions, 100 experiments averaged). Robots have *not* co-ordinated their visual-body-images and motor behaviours by using a mirror, but only through language. The x-axis plots the number of language games. The running average of communicative success and average lexicon size as well as invention and adoption frequency are shown on the y-axis. In contrast to an experiment with more than two agents (see Figure 9), the populations do not necessarily agree on the optimal set of words, while still reaching 100% communicative success.

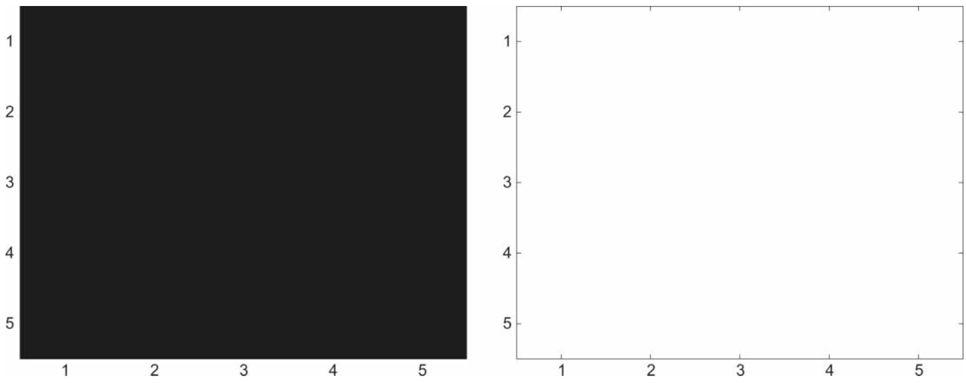

Figure 14. Aggregated results from 100 experiments with a population of two agents for five postures, similar to Figure 11. The left diagram shows the average strength between visual prototypes (x-axis) and motor behaviours (y-axis) and the right diagram shows the product. The left diagram is completely black because both agents have links for all possible mappings (averaged overall experiments), and the right diagram is completely white because they do not agree on such mappings (overall experiments).

the motor behaviours that generate them. The latter reaches a steady state when there is complete visuo-motor co-ordination when there are more than three agents. Both processes are driven by optimisation for communicative success.

5. Further enhancements

The results of this second experiment are remarkable, but a key issue is of course whether the solution will scale up. Agents are searching within a space of possible combinations between words, visual prototypes and motor behaviours. This space grows depending on the number of words (which is itself a function of the size of the population) and the number of possible actions the robots can perform. Figure 15 measures the time until the last failed interaction between any two agents in the population. Two curves are shown; the top one relates to the script used in the previous experiment. We see that when the population size increases; the time to success scales rather badly. A similar exercise is scaling the number of actions, which gives a similar result. Figure 16 shows that the time towards global success scales even worse when scaling the number of actions.

It is tempting to seek more sophisticated learning methods or provide additional sources of information for the robots, such as expectations about which actions might be performed. These things would undoubtly help; but here we are interested to see whether the interaction scripts for the Action Game itself could have any influence, which would be the case if both agents increase the amount of information that they can draw from the language game. This happens to be the case with the following script:

(1) The speaker randomly chooses an image-schema from the known schemata as the topic of the conversation.
(2) He looks up a word associated with the schema. If there is none he invents one.
(3) He looks up the motor behaviour associated with the word. If there is no associated motor behaviour he picks one randomly.
(4) The speaker utters the word and performs the motor behaviour picked.
(5) The hearer parses the uttered word and looks up the image-schema and motor behaviour associated with the word.

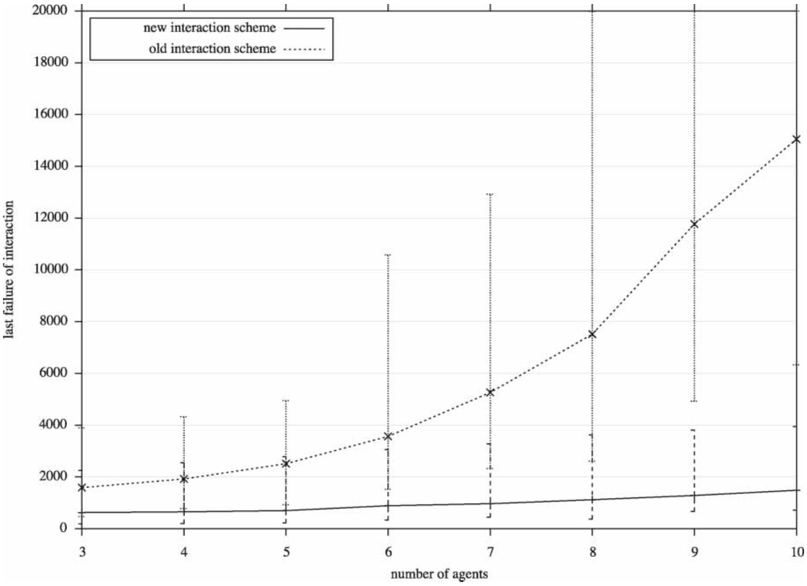

Figure 15. Two interaction schemes are compared for how they scale with respect to the number of agents. The old interaction scheme used in the previous experiment scales much worse than the new interaction scheme where agents exchange more information within a single language. All experiments in this graph are about aligning and co-ordinating five actions. For every combination of agents and actions, 100 experiments are run and the result is averaged. The error bars show the min, max values.

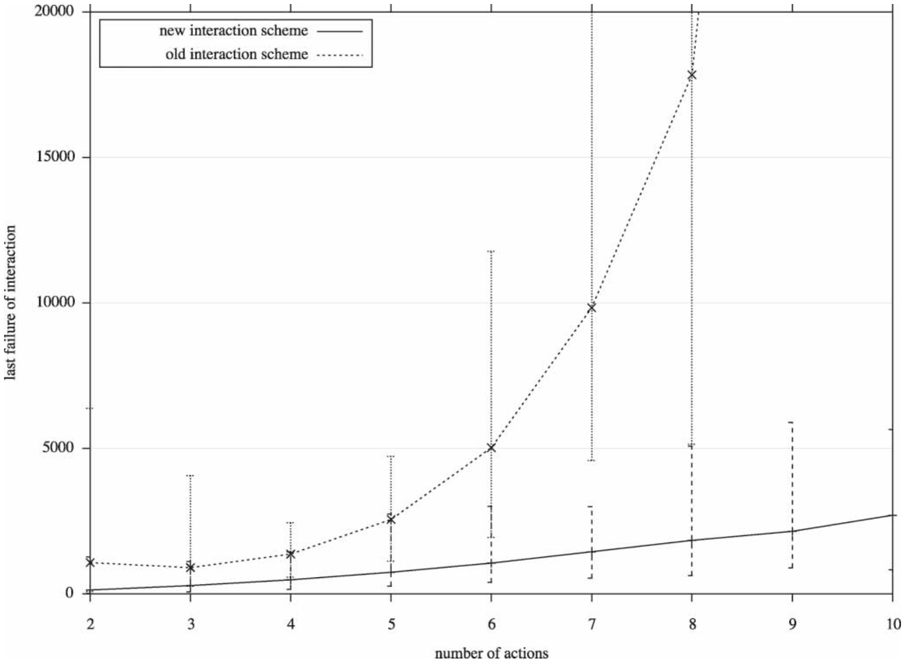

Figure 16. Two interaction schemes are compared with respect to the number of actions. We see an that the old interaction scales rather badly whereas the new interaction scheme performs well given a growing number of actions. Already with eight actions, 20,000 interactions are sometimes not enough in the old paradigm to reach a steady state. All experiments in this graph involve a population of five agents. The number of actions is depicted on the x-axis. For every combination of agents and actions, 100 experiments are run and for every experiment the last failed interaction is counted and averaged. The error bars show min, max values.

(6) If there are no image-schemata associated with the word he classifies the image-schema resulting from the motor behaviour performed by the speaker and associates it with the word.
(7) The hearer performs the motor behaviour associated with the word. If there is none he picks one of the known motor behaviours randomly.
(8) Both agents determine the success of the interaction. The speaker determines whether the hearer performed the intended behaviour. The hearer determines whether the image-schema associated with the uttered word corresponds to the motor behaviour performed by the speaker.

In a sense a game has four possible outcomes. The speaker and hearer both agree, the speaker agrees or the hearer agrees and both disagree. Given this outcome both update their lexicons. As shown in Figures 15 and 16, we see that with this new interaction scheme, the system can cope with scaled number of agents and number of actions. This experiment demonstrates that not only cognitive mechanisms but also interaction patterns can have a dramatic influence on performance.

In the second experiment, agents do not have ‘mirror-neuron’ nodes that become active either when a visual image is seen or when the corresponding motor action is performed. Nevertheless, they have co-ordinated visuo-motor mappings using the same word to refer both to the visual image of an action performed by the other and to the motor control program for the action. The next question is whether the introduction of such nodes would lead to better performance. This does not turn out to be the case, as Figure 17 shows. In this experiment, agents follow a script that is closer to the one used in the mirror experiment. Words are associated with posture nodes, as in Figure 5, and the same update rules are used to increase the strength of connections

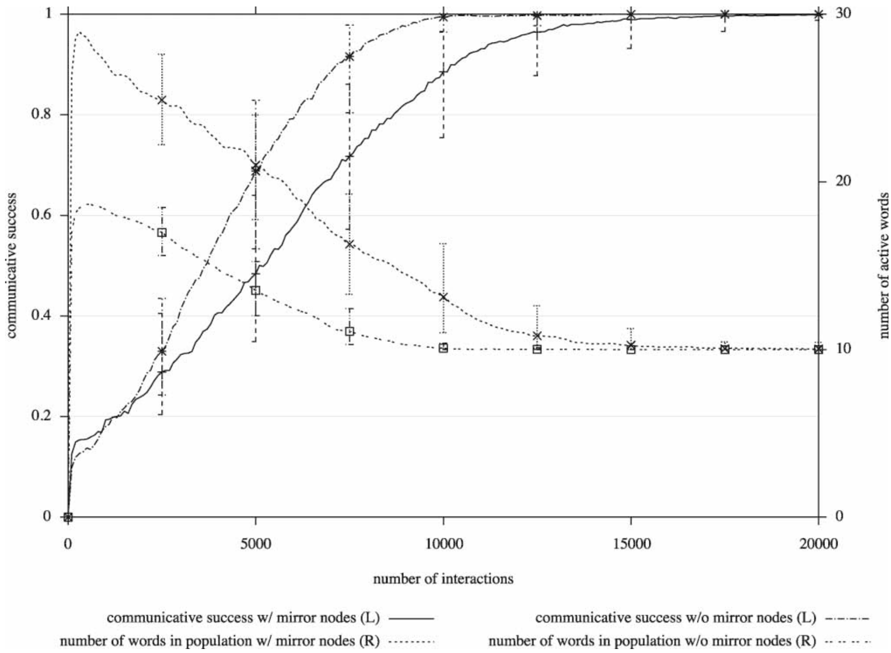

Figure 17. Experiment in which agents perform visuo-motor co-ordination without prior exposure to mirrors and with and without explicit ‘mirror’ nodes for postures. We see that the performance of the system using explicit ‘mirror’ nodes is worse than the solution where agents just add links between words and visual postures and between words and motor behaviours. This graph shows the communicative success and the lexicon size for experiments with populations of five agents, negotiating and co-ordinating 10 actions. The experiment is repeated 100 times with and without mirror nodes, the results are averaged.

between posture nodes and actions or visual images, in the case of a successful game, or to add new connections and to decrease strength of others, in the case of failure. We see, however, that performance is worse than in the experiment where they do not use posture nodes. The reason is that posture nodes do not really help because they increase the search space of possible networks. Note that the Hebbian learning of the posture nodes (used in the mirror experiment) cannot be used here because the visual image and the corresponding motor behaviour are never active at the same time as the motor behaviour is performed by the other robot. So the possible role of mirror nodes in visuo-motor co-ordination through language or other forms of co-operative action remains an intriguing question for further research (Steels and Spranger 2008).

6. Conclusions

We have shown in this paper that physically embodied agents can co-ordinate their visual experiences and motor behaviours either through observing themselves in the mirror or through language. The agents were set up to play a language game in which they ask each other to take on certain postures and give feedback whether the posture was actually reached. The agents start without a prior inventory of names, without a prior inventory of visual prototypes for body postures, and without knowing the relation between the visual prototypes of postures and the motor behaviours that re-enact them. We have seen that it suffices to set up the right kind of semiotic dynamics, establishing networks that link perception, categorisation, action and naming, in order to see networks arise that are not only sufficiently aligned across agents to achieve successful communication, but also co-ordinate visual perception and motor behaviour as a side effect. In the body language experiment, the co-ordination is not explicitly represented with posture nodes as in the mirror experiment, but arises implicitly through the language game, even with the simple consolidation strategies used here.

There are many other avenues that can be explored further. One obvious further research topic is to tackle descriptions of dynamical gestures as opposed to postures. However, the most intriguing issue concerns the role of mirror nodes. It is easy to see how such nodes may form in the case where a mirror is employed, but, surprisingly, when establishing visuo-motor co-ordination through language they were not needed. However, when additional sources of knowledge are used to improve the efficiency for finding the right mappings, such as simulation of actions to imagine internally what they might look like, mirror nodes become essential (Steels and Spranger 2008).

Acknowledgements

This research was carried out at the Sony Computer Science Laboratory in Paris with partial support from the ECAgents project, funded by the EU Future and Emerging Technologies programme (IST-FET) as IST-1940, and the ALEAR project, funded by the EU Cognitive Systems programme. We are extremely grateful to Masahiro Fujita, Hideki Shimomura and their team at the Sony Intelligent Systems Research Lab in Tokyo for giving the opportunity and support to perform the QRIO experiments, and to Martin Loetzsch for his help in setting up and carrying out the experiments. The information provided here is the sole responsibility of the authors and does not reflect the EU Commission's opinion. The Commission is not responsible for any use that may be made of data appearing in this publication.

References

Billard, A. (2002), 'Imitation: A Means to Enhance Learning of a Synthetic Proto-language in an Autonomous Robot,' in *Imitation in Animals and Artifacts*, eds. K. Dautenhahn and C. L. Nehaniv, Cambridge, MA: MIT Press, pp. 281–311.

Bishop, C. (1995), *Neural Networks for Pattern Recognition*, New York: Oxford University Press.

Blanke, O., and Castillo, V. (2007), 'Clinical Neuroimaging in Epileptic Patients with Autoscopic Hallucinations and Out-of-body Experiences Case Report and Review of the Literature,' *Epileptologie*, 24, 90–96.

Demiris, Y., and Johnson, M. (2003), 'Distributed, Predictive Perception of Actions: A Biologically Inspired Robotics Architecture for Imitation and Learning,' *Connection Science*, 15, 231–243.

Dominey, P. (2003), 'Learning Grammatical Constructions in a Miniature Language from Narrated Video Events,' *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, Boston, MA.

Feldman, J. (2006), *From Molecule to Metaphor*: *A Neural Theory of Language*, Cambridge, MA: MIT Press.

Gutmann, J., Fukuchi, M., and Fujita, M. (2005), 'Real-time Path Planning for Humanoid Robot Navigation,' in *International Joint Conference on Artificial Intelligence*, pp. 1232–1238.

Hu, M. (1962), 'Visual Pattern Recognition by Moment Invariants,' *IEEE Transactions on Information Theory*, 8, 179–187.

Jeannerod, M. (2001), 'Neural Simulation of Action: A Unifying Mechanism for Motor Cognition,' *Neuroimage*, 14, 103–109.

Matarič, M. (2002), 'Sensory-motor Primitives as a Basis for Imitation; Linking Perception to Action and Biology to Robotics,' in *Imitation in Animals and Artifacts*, eds. K. Dautenhahn and C.L. Nehaniv, Cambridge, MA: MIT Press, pp. 392–422.

Mukundan, R., and Ramakrishnan, K. (1998), *Moment Functions in Image Analysis*: *Theory and Applications*. London, UK: World Scientific.

Nilsson, N. (1984), 'Shakey the Robot,' Technical Note 323, AI Center SRI International, Menlo Park, CA, USA.

Pfeifer, R., Bongard, J., and Grand, S. (2007), *How the Body Shapes the Way We Think*: *A New View of Intelligence*, Cambridge, MA: MIT Press.

Ramachandran, V., and Hirstein, W. (1998), 'The Perception of Phantom Limbs,' *Brain*, 121, 1603–1630.

Rizzolatti, G., and Arbib, M. (1998), 'Language within our Grasp,' *Trends in Neurosciences*, 21, 188–194.

Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996), 'Premotor Cortex and the Recognition of Motor Actions,' *Cognitive Brain Research*, 3, 131–141.

Rosenfield, I. (1988), *The Invention of Memory*: *A New View of the Brain*, New York: Basic Books.

Roy, D., and Pentland, A. (2002), 'Learning Words from Sights and Sounds: A Computational Model,' *Cognitive Science*, 26, 113–146.

Siskind, J. (2001), 'Grounding the Lexical Semantics of Verbs in Visual Perception Using Force Dynamics and Event Logic,' *Journal of Artificial Intelligence Research*, 15, 31–90.

Spranger, M. (2008), 'World Models for Grounded Language Games,' Diploma thesis, Humboldt-Universität zu Berlin.

Steels, L. (1994), 'A Case Study in the Behavior-oriented Design of Autonomous Agents,' *From Animals to Animats*, 3, 445–452.

Steels, L. (1995), 'A Self-organizing Spatial Vocabulary,' *Artificial Life*, 2, 319–332.

Steels, L. (2003), 'Evolving Grounded Communication for Robots,' *Trends in Cognitive Sciences*, 7, 308–312.

Steels, L. (2005), 'The Emergence and Evolution of Linguistic Structure: From Lexical to Grammatical Communication Systems,' *Connection Science*, 17, 213–230.

Steels, L., and Baillie, J. (2003), 'Shared Grounding of Event Descriptions by Autonomous Robots,' *Robotics and Autonomous Systems*, 43, 163–173.

Steels, L., and Belpaeme, T. (2005), 'Coordinating Perceptually Grounded Categories through Language: A Case Study for Colour,' *Behavioral and Brain Sciences*, 28, 469–489.

Steels, L., and Kaplan, F. (2001), 'AIBOs First Words. The Social Learning of Language and Meaning,' *Evolution of Communication*, 4, 3–32.

Steels, L., and Loetzsch, M. (2008), 'Perspective Alignment in Spatial Language,' in *Spatial Language and Dialogue*, eds. K.R. Coventry, T. Tenbrink and J.A. Bateman, Oxford: Oxford University Press (in press).

Steels, L., and Spranger, M. (2008), 'Can Body Language Shape Body Image?,' in *Proceedings of the Eleventh International Conference on Artificial Life*, pp. 577–584.

Talmy, L. (2000), *Toward a Cognitive Semantics*, Cambridge, MA: MIT.

Triesch, J., Jasso, H., and Deak, G. (2007) 'Emergence of Mirror Neurons in a Model of Gaze Following,' *Adaptive Behavior*, 15, 149–165.

Vogt, P., and Divina, F. (2007), 'Social Symbol Grounding and Language Evolution,' *Interaction Studies*, 8, 31–52.

Wood, J. (1996), 'Invariant Pattern Recognition: A Review,' *Pattern Recognition*, 29, 1–17.

Appendix A. Moments

We use normalised central moments to describe the shape of the upper torso of the robot in the binary (black and white) image provided by the underlying vision modules (see third image from the left in Figure 3). These moments are global, translation and scale-invariant shape descriptors. As they take the whole body (global) into account they do not handle occlusion well. Additionally, these moments lack the rotation invariance property of other shape descriptors (like Hu moments; see Hu 1962). However, the lack of rotation invariance is not an issue because the robots are always positioned face-to-face when engaging in language games, as well as their head movements being restricted such that the image plane is always parallel to the ground. Given these assumptions, the following approach becomes effective.

The central moment of order $(p + q)$ can be defined (Hu 1962) for a two-dimensional discrete signal as follows

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y)$$

with $\bar{x}$ and $\bar{y}$ being the centroid following from

$$\bar{x} = \frac{M_{10}}{M_{00}}$$

$$\bar{y} = \frac{M_{01}}{M_{00}}$$

with M_{pq} being the raw moment of order $(p+q)$ defined as

$$M_{pq} = \sum_x \sum_y x^p y^q I(x, y).$$

In these definitions $I(x, y)$ signifies a black and white image, with the shape to be described by the feature being a connected block of pixels x, y with $I(x, y) = 1$. The normalised central moment η_{pq} of order $(p+q)$ is then computed as in Mukundan and Ramakrishnan (1998)

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{1+(p+q)/2}}. \tag{A1}$$

We use the seven normalised central moments from the order of two to the order of three.

Prototypes

Every agent maintains a set of prototypes describing the sensory experiences of observed postures. The sensory experiences are based on input from lower level visual processing modules (foreground/background segmentation, upper torso detection and so forth) by applying Equation (A1) to the binary image describing the upper body torso:

$$s := \begin{pmatrix} \eta_{20} & \eta_{11} & \eta_{02} & \eta_{30} & \eta_{21} & \eta_{12} & \eta_{03} \end{pmatrix}^T.$$

The prototype P for a set of sensory experiences $\{s_1, \ldots, s_N\}$ is simply the average of all features:

$$P := \begin{pmatrix} \eta_{20} & \eta_{11} & \eta_{02} & \eta_{30} & \eta_{21} & \eta_{12} & \eta_{03} \end{pmatrix}^T = \frac{1}{N} \sum_{k=1}^{N} s_k.$$

The sensory experiences are kept in memory and a prototype can therefore be adjusted to assimilate a new sensory experience by recomputing the average again.

Given a set of prototypes $P_1, \ldots, P_M$, a new sensory experience is classified using a Minimum Euclidean Distance Classifier:

$$\text{class}(s) = \arg\min_i \|P_i - s\|$$

which computes the minimum distance to the sensory experience (see Wood 1996).

Visuo-motor co-ordination in the body language experiment

This subsection defines measures for tracking the visuo-motor co-ordination in the body language experiments without explicit posture nodes linking visual prototypes with motor behaviours. Assume a population A of agents that are bootstrapping a communication system for naming N postures (N prototypes and N motor behaviours). For every agent a, it is possible to construct two incidence matrices **V** and **M** containing the weights between visual prototypes and words on the one hand and words and motor behaviours on the other. The matrices are defined as

$$V := \left(v_{ij}\right)_{N \times W}$$

$$M := \left(m_{ij}\right)_{N \times W}$$

with W being the number of words known to the agent. Using these matrices, we can compute the relation from visual prototypes $v_1, \ldots, v_N$ to motor behaviours $m_1, \ldots, m_N$ as mediated by a particular word w:

$$\hat{\mathbf{M}}_{a,w} := \begin{pmatrix} v_{1w} & v_{2w} & \ldots & v_{Nw} \end{pmatrix}^T \cdot \begin{pmatrix} m_{1w} & m_{2w} & \ldots & m_{Nw} \end{pmatrix}.$$

The visuo-motor relations for all words known to the agents are computed by summing the matrices built for individual words:

$$\hat{\mathbf{M}}_a := \sum_{w=1\cdots W} \hat{\mathbf{M}}_{a,w}.$$

The algorithm clearly adheres to the rules of matrix multiplication (which is the way this is actually implemented), such that

$$\hat{\mathbf{M}}_a = V \cdot M^T.$$

To quantify how aligned the agents are we compute two matrices, one quantifying which links exist in agents (computed as the sum of the above for all agents) and which links agents agree upon, that is, which links are made by all agents (computed by taking the member-wise matrix product $\circ$):

$$\hat{\mathbf{M}}_{+} := \sum_{a=1\cdots A} \hat{\mathbf{M}}_{\mathrm{a}}$$

$$\hat{\mathbf{M}}_{\circ} := \hat{\mathbf{M}}_{1} \circ \hat{\mathbf{M}}_{2} \circ \cdots \circ \hat{\mathbf{M}}_{A}.$$

These matrices are used as an agent external coherence measure of visuo-motor co-ordination and cannot be computed locally by agents (note that the matrix $\hat{\mathbf{M}}_{\mathrm{agent}}$ is in principle computable locally). They are not directly used by agents and serve only as an indicator of visuo-motor alignment after running an experiment. To that end these matrices can be visualized using a two dimensional height map, which maps a matrix element position to a connection strength. Before visualising these matrices $\hat{\mathbf{M}}_{+}$ and $\hat{\mathbf{M}}_{\circ}$ they are normalised beforehand. $\hat{\mathbf{M}}_{+}$ is normalised by dividing each matrix element through the number of agents, which in most cases is sufficient to ensure members of the matrix are between zero and one, given the assumption that after some time the lateral inhibition dynamics will associate a word at most with one visual prototype and at most with one motor behaviour. The matrices $\hat{\mathbf{M}}_{a}$ are normalised by dividing each visual-prototype, motor-link by the number of contributing words.

Here is an example. Let us suppose agent-1 is taking part in an experiment with two postures (two prototypes and two motor behaviours) and has established the relations between word visual prototypes and motor behaviours as defined by the incidence matrices:

$$\mathbf{V} = \begin{pmatrix} 0.5 & 1 & 0 \\ 0 & 0.2 & 1 \end{pmatrix}.$$

$$\mathbf{M} = \begin{pmatrix} 0 & 1 & 0 \\ 0.3 & 0 & 1 \end{pmatrix}.$$

Given these relations it is possible to compute the incidence matrices describing the relation between visual prototypes and motor behaviours for every word known to the agent:

$$\hat{\mathbf{M}}_{1,1} = \begin{pmatrix} 0.5 & 0 \end{pmatrix}^{T} \cdot \begin{pmatrix} 0 & 0.3 \end{pmatrix} = \begin{pmatrix} 0 & 0.15 \\ 0 & 0 \end{pmatrix}$$

$$\hat{\mathbf{M}}_{1,2} = \begin{pmatrix} 1 & 0 \\ 0.2 & 0 \end{pmatrix}$$

$$\hat{\mathbf{M}}_{1,3} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

These matrices are combined to reflect the overall scores agent-1 assigns to particular visuo-motor relations mediated by language:

$$\hat{\mathbf{M}}_{1} = \begin{pmatrix} 1 & 0.15 \\ 0.2 & 1 \end{pmatrix},$$

which in this particular case shows that the agent strongly connects v_1 and m_1 and v_2 and m_2. Given a similar representation for agent-2

$$\hat{\mathbf{M}}_{2} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

we can compute inter-agent visuo-motor co-ordination coherence, which yields

$$\hat{\mathbf{M}}_{+} = \begin{pmatrix} 0.5 & 0.075 \\ 0.1 & 1 \end{pmatrix}$$

$$\hat{\mathbf{M}}_{\circ} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

The two agents of this particular population establish all possible connections but they only agree on a mapping of v_2 to m_2.