Dynamic Evolution of Language Games between two Autonomous Robots

Jean-Christophe Baillie
Lab. of Elec. & Computer Engineering, ENSTA
32, Bd Victor F75015 Paris
jean-christophe.baillie@ensta.fr

Matthieu Nottale
Lab. of Elec. & Computer Engineering, ENSTA
32, Bd Victor F75015 Paris
matthieu.nottale@ensta.fr

Abstract*— The "Talking Robots" experiment, inspired by the "Talking Heads" experiment from Sony, explores possibilities on how to ground symbols into perception. We present here the first results of this experiment and outline a possible extension to social behaviors grounding: the purpose is to have the robots develop not only a lexicon but also the interaction protocol, or** *language game, that they use to create the lexicon. This raises several complex problems that we review here.**

***Index Terms*— Language acquisition, Language games, Symbol grounding, Social behaviors grounding, Autonomous development, Architectures**

I. Introduction

Grounding symbolic representations into perception is a key and difficult issue for artificial intelligence [3], [12], [2], [11]. Including social interactions and, more specifically, language acquisition and development in this context has proven to be a fruitful orientation. One of the recent and successful attempts in this direction is the "Talking Heads" experiment [10], [8], [7]. This experiment involves two cameras interacting in a simplified visual environment made of colored shapes on a white board. This interaction is called a "language game". The agents inside the cameras are developing a *shared grounded lexicon* of words related to visual meanings like "red", "large", "above", etc.

The "Talking Heads" experiment has proven that it was possible to design a computer-based mechanism for language acquisition that shares many properties in common with what can be observed in human language acquisition. It can be seen as a attempt in the direction of a computational model of language acquisition and symbol grounding.

However, the "Talking Heads" had two limitations, which were good simplifications to start with, but that we will consider carefully here:

- The environment was limited and simplified (only colored shapes on a white background, fixed cameras).
- The interaction protocol (language game) was predetermined and hardcoded.

In a first phase, exposed in the first part of this article, we have tried to remove the first constraint, which led us to the "Talking Robots" experiment. With the recent development of relatively cheap and powerful robotic platforms (see Sony, Honda or Fujitsu, among others) research on symbol grounding can move from simulation or simple environments to complex natural environments and embodied systems. The "Talking Robots" experiment is following this trend and try to reproduce the initial Talking Heads experiment with autonomous robots (Aibo ERS7) evolving in an unconstrained visual environment instead of simple cameras. One of the goals is to reinforce the validity of the first results of the Talking Heads in showing that the proposed mechanism for lexicon acquisition stands in the case of a noisy, complex environment.

The purpose of the "Talking Robot" is not only to reproduce the "Talking Heads" but also to serve as a ground to build, in a second phase, a more complex experiment involving not only a predetermined *language game* interaction protocol, but where the robots will dynamically create a protocol and use it to play a language game which will not be known in advance. The image could be the one of two kids starting to play a game, without knowing in advance what game they will play and creating their own rules on the go. The complexity of the environment and interactions is a crucial factor to the success of this task.

This is a very ambitious research project and is related to several difficult and known technical and theoretical problems that we will discuss here.

We will briefly present in this article the structure and technical issues of the existing "Talking Robots" experiment and the underlying scientific questions that we address. We will also present our first results. Finally, we outline a possible extension of the experiment to the more general problem of dynamical creation of language games.

II. Grounding symbols into perception: the "Talking Robots" experiment

Since the converging interaction mechanisms of the original Talking Heads experiment were already existing, the most challenging part of the "Talking Robots" experiment was to overcome the technical issues related to the complexity of the environment[1].

We shortly describe here the problems and our solutions. Putting all these elements together, we successfully obtained converging results, that we present here for the ”Discrimination Game” (comparable to the original experiment) and for several successful ”Guessing Games”.

A. Technical issues

In the Talking Robots, the two agents are ERS7 Aibo robots. They are autonomous and start the interaction from any position in the lab, they use visual cues to locate each other and voice recognition to speak the words used in the experiment. They also use the surrounding environment as a base to extract interesting objects to discuss about:

1) Sound synchronization and speech recognition: We use beep signals with specific tones to automatically determine the leader (the ”speaker”) and the follower (the ”hearer”) in the game, and to later synchronize the robots activities.

Voice recognition using standard Hidden Markov Models with the HTK toolkit is performed on a set of 20 syllables to recognize the words created by the speaker as a combination of those 20 syllables. The success rate is approximately 80%. Note that by using spoken language instead of pure frequency sounds, the game is in principle playable by a robot and a human, which is a further extension we plan to run. The following technical choices have also been done with this human/robot constraint in mind.

2) Robot positioning: One requirement for the experiment is to have the two robots standing one next to the other, starting from a random position, and looking in the same direction: we use movement detectors and a waving head motion to locate an Aibo head in the image and pattern recognition based on a FMI-SPOMF transform[6] to identify the orientation of this head. Together with a scale measure, this gives an accurate enough estimation of the position and orientation of the other robot.

3) Stable segmentation: For the robots to have comparable representation of their visual environment, it is necessary to use a stable segmentation algorithm which will give comparable results for salient regions observed by the robots from slightly different points of view: the CSC algorithm[5] has proven to be the most stable after a series of tests. The measure of stability is the entropy of a normalized correlation matrix between two segmentations of two slightly different images, which should ideally be zero. We only keep the most salient regions in the calculation, as the Talking Robots do. The CSC algorithm scores an average of 0.07, which is in practice much better than 0.11 for Recursive Histogram Splitting or 0.19 for Split & Merge.

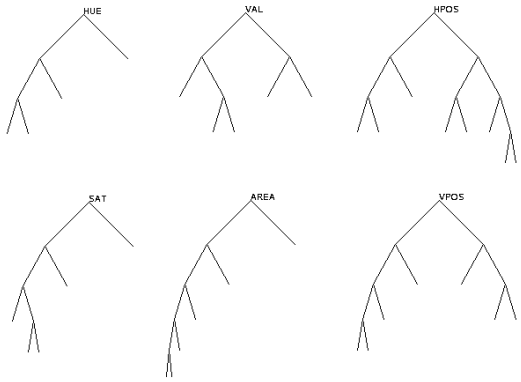
4) Pointing device: In the process of the Guessing Game, we need an accurate pointing device for the robots to designate the objects they are talking about: we use a blinking laser pointer fixed on the robot’s head (controlled by a photo detector on one of the robot’s back LED) and a simple red spot detector which gives very accurate results.

This blinking laser is also used to center the field of view of the two robots at the beginning of the game, enforcing the similarity of the two scenes segmentation.

More technical details can be found in [1].

B. Discrimination Game

The discrimination game[8] is played by one robot only and aims at dynamically create a set of visual categories grounded in the surrounding visual environment of the robot: for example, with objects of different sizes in the image, a category for ”big” and ”small” will be developed but no category corresponding to ”colors” will arise in a black and white environment. Here is an example of the kind of categories grown, as observed in the environment of our lab:



Our games successfully converge with 250 iterations to a set of approximatively 80 to 90 categories in the unconstrained environment of the lab, and a success rate above 80% (see fig 1, 2).

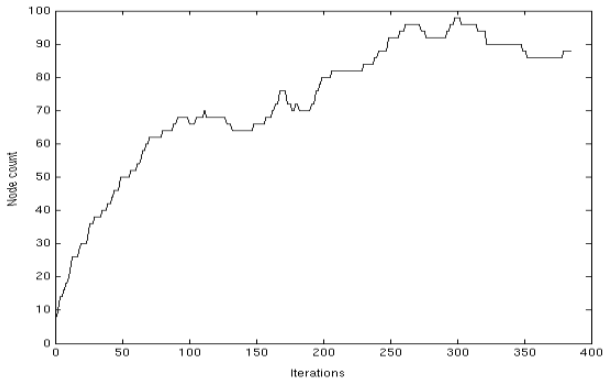


Fig. 1. Stabilization of the number of categories

Categories are mainly developed for position and area. If we turn the vision of the robot to black and white, the saturation discrimination tree (SAT) shrinks in a few iterations, as expected.

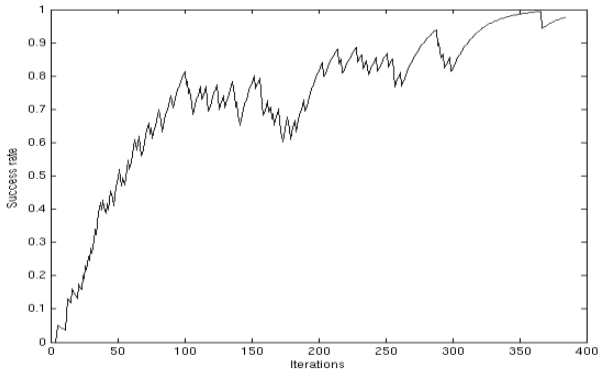
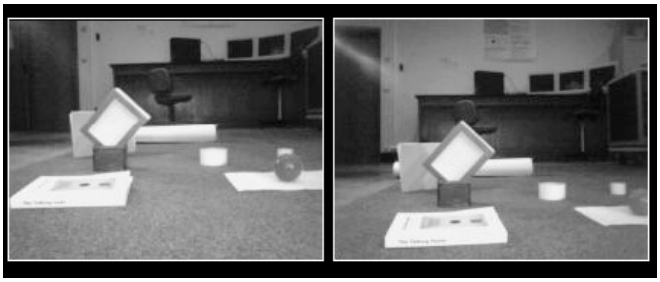


Fig. 2. Convergence of the success rate

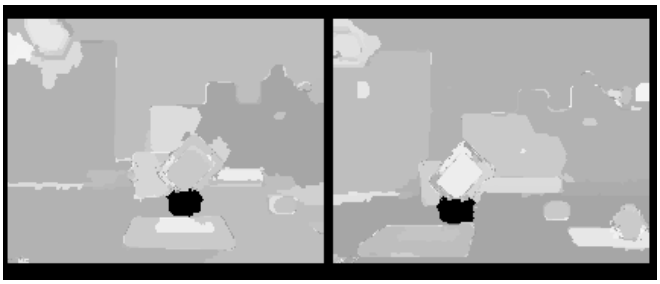
C. Guessing Game

The ”Guessing Game” has been run on a limited set of experiments. Although we have still no statistical results yet, we have observed in detail several games to compare them to the original Talking Heads.

The kind of visual scenes the robots deal with is the natural environment of the lab with several colored objects on the floor. Here is an example of the two viewpoints from two robots playing a game:



In one session, the speaker creates the word ”kakomima” for a topic associated to the meaning ”WIDTH [0.5 , 0.75]”. The hearer correctly hears the word but fails to interpret it. The laser pointing mechanism used by the speaker to tell where was the topic is activated and the topic is located by the hearer, as shown in the following extracted segmentation from both robots (the topic is in black):



But in this example, the most relevant category found by the hearer for this topic is ”HUE [0.25 , 0.375]”. The game fails and the hearer is creating a wrong association.

This situation is happening most of the time and we have to improve the constancy of segmentation and context to increase the probability of having the speaker and the hearer both categorizing a shared topic in the same way. Apart from this, the original mechanisms of the Talking Heads have been successfully re-implemented and run.

III. Grounding social behaviors into perception: evolving language games

Trying to generalize the Talking Robots to a more general experience where the interaction is dynamically created is a challenging problem. What is mainly interesting, and which makes the project difficult, is to design it in a way which exhibits some general principle that is scientifically relevant to the understanding of the coupling of social interactions and the development of grounded semantics.

The desired result of this ”extended” experiment is to have the robots develop a set of shared social behaviors corresponding to news language games. Several questions arise:

- What internal representation should be used for a behavior?
- What is the general architecture and how do categories, words, behaviors and perception interact?
- How can the behaviors be compared one to the other?
- What is the main driving force of the system while creating, deleting and using behaviors?

The last two points particularly have been the focus of recent new researches on motivation and self development for autonomous robots[9], [4].

A. Internal representation of behaviors

We can take the existing Guessing Game as a typical example of the kind of interaction the robot should build. The Guessing Game can be described in terms of internal states, conditional branching, event catching related to the perception layer and actions in a general sense (physical actions, speaking words or changing one’s internal state). One suitable structure to describe such a behavior is a state machine with transitions triggered by conditions and synchronization wait points (where a condition is blocking until is becomes true).

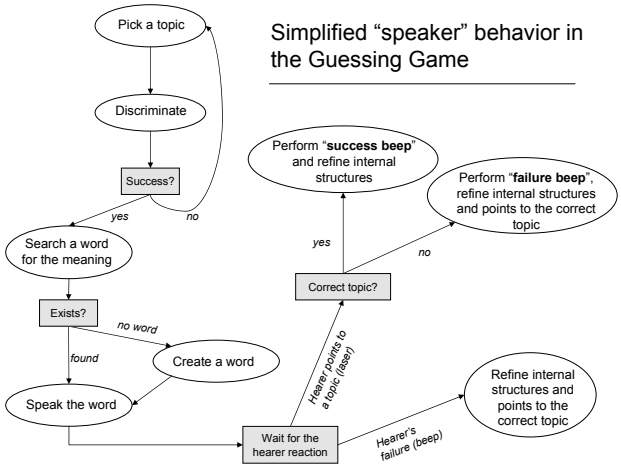
The grammar to build behaviors is then a combination of a graph building capability, a set of admissible conditions expressed over perceptions and internal structures (existing words, existing categories, scores of words/meaning associations), synchronization conditions and typical admissible action descriptions[1].

[1]Shared local variables to store temporary structures are also necessary, but we will not detail this point here

One key element of the behaviors we aim at is that they terminate at some point and that there are exactly two outcomes possible: we will call these outcomes ”success” and ”failure”, even if the intuitive meaning of success and failure is not clear considering the given behavior.

Another important property is that the behavior must somehow modify the internal structures of the robot, according to the interaction of the robot with the outside world, so that the nature of the interaction has a chance to evolve. This implies that at least one of the actions is modifying internal structures and one of the tests is relative to the perception channels. This might not be enough to constrain the class of behaviors to non trivial ones, but it is already a first requirement.

As an illustration, here is the behavior corresponding the speaker side of the guessing game:



B. Behavior evaluation

The most obvious evaluation for a given behavior is to measure how much ”success” outcome it produces when it is run. However, this is not a sufficient measure for two reasons:

1) A trivial behavior which would systematically yield to ”success” and do nothing else would get a very high evaluation. Obviously, this is not an interesting behavior by any practical definition. A more interesting notion is the measure of the increase in ”success” outcome, a first order derivative of the first measure. In that case, an ”interesting” behavior is a behavior whose success rate increases over time, implicitly leading to learning and increase of skills[2].
2) A behavior like the ”speaker” part of the Guessing Game can not be considered independently of what the other robot is doing. It might be ”interesting” (increase of success rate) with a partner running a ”hearer” behavior, and totally uninteresting otherwise. In other words, a behavior evaluation is only meaningful on a couple of behaviors, which we will call a ”game”.

The last point implies that the robot stores a representation of the other robot's behavior as part of his own behavioral state. Technically speaking this means that the interactions, or ”games”, will be stored as a couple of a behaviors with a ”leader” and a ”follower”.

This simple *theory of mind* leads to a second kind of evaluation measure for a game, which is the adequation between the supposed behavior model of the partner and its actual behavior. Failure to this regard might be detected in several ways:

- An expected synchronization never occurs and a timeout is activated.
- Some observed actions of the partner are not coherent with its supposed current state. This supposes of course that there is a reliable way of observing the actions, which is in itself a difficult unresolved problem.
- Synchronization signals (like success or failure) are detected while they should not happen yet or they should not happen at all.

The final evaluation of a game is then the combination of its *score*, which is the increase of success rate over a given period of time, and of its *synchronicity*, which is a measure of how appropriately it describes the currently happening interaction, including the partner's behavior. The score of a game should be considered only if the synchronicity is high enough, otherwise no judgment on the game can be assessed.

It is also interesting to note that if a game implements a form of grounded communication in a sufficiently changing environment, it will have a good score since it will adapt the agent representations to the changes of the environment and lead to an increase of success rate. In other words, games that are grounded communication-based games and not determined in a closed environment should have a high score, which is an interesting side effect.

C. Driving force

Let us suppose that the robots have already a repertoire of games, with corresponding scores in memory. We propose that there are three courses of action that a robot can take at start: 1. Do nothing, 2. Observe the other robot's actions, 3. Initiate a game.

In the second case, the robot will update the synchronicity of the games stored in memory by comparing what the partner is actually doing to its own available behavior descriptions. Once again, this is a technically very difficult task, but in principle it would yield to an estimation of the most ”synchronized” behavior descriptions among the repertoire of known games and for the observed actions. If there is no

[2]Trivial behaviors could also arise in that case, with for example agreed increase of success between the agents. That is why the constraints stated in III-A are important components of the admissible behavior definition.

synchronized behavior available, the robot will stimulate the creation of new behavior descriptions and start again.

In the third case, the robot decides to initiate a game as a ”leader”, picking up the behavior with the best score or using information from the second phase to enter a game initiated by the other robot, as a ”follower”. Once again, if there is no game available in memory with a high enough score, the robot will try to create new games and try them out.

The robot can stop the current game and restart to the initial state, choosing between the three courses of action. This should happen in two situations: the increase of success rate is low for a long period of time (the game becomes ”boring”), the synchronicity of the game is too low (the partner is not following the planned behavior).

These simples rules based on score and synchronicity need to be implemented in a real experiment to be tested, but they are reasonable first steps in the design of a mechanism where turn-taking and complex rewarding interactions could build up between two autonomous robots, without supervision or predetermined structures. The driving forces of the whole system is to maximize the synchronicity and the score of its internal game representations. The synchronicity measures how ”shared” the games are and the score measures how ”adaptive” they are. As we already mentioned, grounded games should have a priori high scores in a changing environment.

D. Naming the games

In a similar way like the Talking Robots, it is important that the robots keep a associative memory between forms (words) and game descriptions. When the leader initiates a game it will speak a word to describe it before the interaction starts. The follower can hear that word and, based on its own associative memory, try to guess what the interaction will be. This is important for two reasons:

- The follower could not guess a game by looking at the first actions of the leader only. Too many games will start in the same way and the result would be a low synchronicity in general and a poor convergence.
- We want to ground words describing actions into social interaction.

There should also be a feedback mechanisms by which a failed interaction can lead to a better convergence of the robot’s internal games. The follower must not only compare the leader observed behavior to his own stored behaviors but try to establish on his own a simple model of this behavior. This model will be used in association with the word spoken at start to create a new form/game association, hopefully capturing some part of the leader’s unknown game. This is probably the most technically difficult part of the experiment.

Interesting extensions can be imagined where the two robots enter a negotiation phase to know what game should be played (based on the expected learning power of the game for each of the participants) and who should be the leader and the follower. For the moment, our robots will behave in an ”egoist” way to this regard and try to impose their view.

E. Game creation

An important issue that we have not addressed yet is the mechanism by which games are created. The visual categories in the Talking Heads/Robots are not created at random but by refining existing categories. The analogy is not totaly relevant here since the starting ”blank” categories in the Talking Robots reflect the pre-existing perceptive channels (WIDTH, AREA, HUE, SATURATION,...), whereas there is no such things for behaviors. But the principle of incremental refinement of existing structures, coupled with the creation of new simple ”empty” structures to increase diversity, is necessary in a selectionnist approach[10].

The refinement can be made by several mechanisms:

- Further develop a state node of a behavior by subdividing it into two nodes.
- Add a branching condition or synchronization before or after an existing node.
- Add transitions between existing nodes.

In the process, all nodes are not equal and several measures could be used to prefer one node to the other, like the frequency by which it is visited when the behavior is run.

The creation of an ”empty” behavior is limited to a single test node with two outcomes: success or failure. One other option to create new behaviors is to ”clone” existing high score games, in a genetically inspired way.

Symmetrically to this ”growing” capability, there should also be a ”pruning” mechanism to reduce the complexity of behaviors. Like in the case of growing, nodes can be deleted or merged and transitions can be removed, based for instance on the frequency of their usage.

Altogether, the pruning and growing of existing behaviors must take place in proportion to the score of the corresponding game, to further develop successful games with a lateral inhibition mechanism, and reduce the influence of less successful games.

F. General architecture

The software architecture of the experiment describes what are the main structures in the system and how they are connected and chained. It must be sufficiently general to be adapted to a large set of ”games” and it should help to make clear what are the most important stages in the information processing.

Figure 3 shows a simplified description of the architecture, which correspond to the Talking Robots and possible extensions. The Talking Robot use only the left side of this schema, ignoring ”Perception of Interaction” and the behavior repertoire.

The robot perception is divided into a part devoted to object identification and categorization (used by the Talking

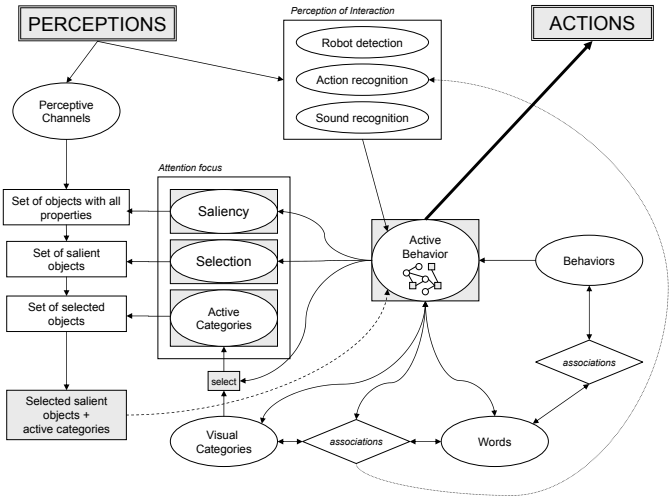


Fig. 3. General architecture of the extended experiment

Robots) and a part devoted to the perception of interactions. The first part extracts objets via a segmentation algorithm and filter them through several perceptive channels. The ”saliency” stage keeps only the most salient objects, the ”selection” stage allows to arbitrarily pick one or several objects and the ”active categories”, selected from the repertoire of known visual categories, filter only the objects corresponding to those ”active categories”. If all available categories are declared ”active”, this stage is performing a classification, instead of a filtering. The final result is usable by the behavior which controls the saliency level, the pickup strategy and selects active categories.

Associations between a repertoire of words and the known visual categories, as well as between words and known behaviors are also provided. They are accessible and modifiable by the behavior.

The second part of the perception layer is devoted to the extraction of more high level information regarding the robots: where the partner is, what is the action it is performing, what are the sound spoken. The action recognition can possibly benefit from the knowledge on existing words and meanings in the robot, to focus on relevant events. This is symbolized by the long arrow from bottom to top.

Finally, the behavior produces ”actions” which are sounds, movements, synchronization tasks or updates of internal structures like the ”word” repertoire or ”visual categories” repertoire.

This is a simple architecture to start with. It is adapted to the implementation of the Guessing Game and, in principle, to other interesting interactions.

IV. Conclusion

The Talking Robot experiment has started to successfully reproduce in a complex environment the results obtained in the simplified environment of the Talking Heads, reinforcing the validity of the first experiment. Building on our experience on this embedded version, we plan to start a new set of experiments involving not only a predetermined *language game* interaction protocol, but where the robots will dynamically create a protocol and use it to play a language game which will not be known in advance. The difficulties of this project have been described here, with first attempts towards solutions. Among the most critical requirement is the definition of a converging dynamics and how it interfaces with the components of the system (perception, categories, meanings) together with an acceptable visual perception of actions performed by other robots.

References

[1] Baillie. Grounding symbols in perception with two interacting autonomous robots. *Proceedings of the 4th International Workshop on Epigenetic Robotics*, 117, 2004.

[2] Brooks. Elephants don’t play chess. *Robotics and Autonomous Systems*, 6(1&2):3–15, June 1990.

[3] Harnad. The symbol grounding problem. *Physica D 42*, pages 335–346, 1990.

[4] Oudeyer and Kaplan. Intelligent adaptive curiosity: a source of self-development. *Proceedings of the 4th International Workshop on Epigenetic Robotics*, 117:127–130, 2004.

[5] Rehrmann and Priese. Fast and robust segmentation of natural color scenes. In *ACCV (1)*, pages 598–606, 1998.

[6] Qin sheng Chen, Michel Defrise, and F. Deconinck. Symmetric phase-only matched filtering of fourier-mellin transforms for image registration and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(12), 1994.

[7] Steels. The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103(1-2):133–156, 1998.

[8] Steels. Language games for autonomous robots. *IEEE Intelligent Systems*, pages 16–22, sept/oct 2001.

[9] Steels. *The Autotelic Principle*, volume 3139. 2004.

[10] Steels and Kaplan. Bootstrapping grounded word semantics. In T. Briscoe, editor, *Linguistic evolution through language acquisition: formal and computational models*, chapter 3, pages 53–73. Cambridge University Press, Cambridge, 2002.

[11] Luc Steels and Jean-Christophe Baillie. Shared grounding of event descriptions by autonomous robots. *Robotics and Autonomous Systems*, 43(2-3):163–173, 2003.

[12] Ziemke. Rethinking grounding. *In Austrian Society for Cognitive Science, Proceedings of New Trends in Cognitive Science - Does Representation need Reality, Vienna.*, 1997.