From: AAAI Technical Report SS-01-05. Compilation copyright © 2001, AAAI (www.aaai.org). All rights reserved.

The role of language in learning grounded representations

Luc Steels
VUB Artificial Intelligence Lab - Brussels
Sony Computer Science Lab - Paris
steels@arti.vub.ac.be

Abstract

There is a broad consensus that the representations used by a cognitive agent must be grounded in external reality through a sensori-motor apparatus. These representations must also be sufficiently similar to those used by other agents in the group to enable coordinated action and communication, and they must be acquired autonomously by each agent. This paper first tries to clarify the terminology and issues involved. Then it argues that language plays a crucial role in the learning of grounded representations because it is a source of feedback and constrains the degrees of freedom of the representations used in the group. The idea of a language game is introduced as framework for concretising the structural coupling between concept formation and symbol acquisition and some experiments are briefly discussed.

1 Defining the problem

Given the terminological confusion in the cognitive sciences it is worthwhile to define more precisely the issues we try to address.

1. In computer science, a *representation* is a physical state of a machine (computer memory for example) which acts as a "stand in" for something else. The physical state becomes thus a means to store information and physical processes operating over the state can implement whatever representational transformations we wish to enact. Thus a representation of a number in a computer is a configuration of digital states. Calculation takes place by changing these states. Objects, concepts, actions, etc. can likewise be represented in an artificial agent by postulating internal states for each of these. Cognitive processes like decision-making, language parsing, object recognition, etc. can then be conceived as physical operations over these internal states.

A representation (both the physical medium chosen and the convention used to map information onto this medium) is arbitrary with respect to what one wants to represent. The only requirement is that the mapping is systematic and that processes operating over a representation are consistent with respect to the mapping that has been adopted. Thus we can not only use binary representations of numbers but also hexadecimal representations, and make use of marks on a laser disk as well as electromagnetic states of an electronic circuit as the state medium. Note that once we are at the level of physical states and physical processes, no additional homunculus is involved to "interpret" representations.

In neuroscience parlance, the equivalent of the computer science notion of a representation is the notion of a neural correlate. This is a biological state (for example the activation of a neuron or set of neurons) which stands for something else, like a control signal for an arm, the recognition of a concept, experience of the color red, etc. Neural processes operating over these physical states, usually thought to take the form of the selective propagation of signals through a network, are the neural correspondence of the physical operations carried out over computational states.

So even if computational implementations are very different from biological implementations, the notion of representation is similar in AI and (cognitive) neuroscience. Using representations and operations over representations to explain cognitive functions seem now so natural and obvious that it is difficult to follow philosophers who claim that cognition does not involve representations. Perhaps they simply have not understood what representations are or are using the notion of representation in another way.

2. We say that a representation is *grounded* when there is an autonomous process that transforms sensations (i.e. data flowing from sensors or motors into internal states) into internal representations and transforms internal representations into motor activations. Through these grounding processes, the agent can coordinate his activities with the world and other agents. The representation need not be an exact, full, veridical representation of the world, and it can be analog or categorial, but it needs to be sufficiently detailed and faithful to support the agent's interaction with the world and others.

Grounding is trivially achieved for devices like a calculator. The user pushes buttons which directly activate internal representations. It is obviously much more complex for representations about the world. Sensors reflect physical properties of the environment which are not necessarily those that the agent needs to focus on. The information is hidden or incomplete in the sensory-motor data and requires complex processing to get out. Often there is not enough information in the sensory data and so representations have to be hypothesised in a top down fashion and mapped onto the sensory-motor data.

In a lot of (pre 1990) AI work the problem of grounding was (temporarily) abstracted out by supplying the representations directly to the computer and only focusing on the processing aspect. This was a useful strategy for a while but it has been rightfully criticised because not all the representations assumed by early AI programs can be grounded on a physically embodied robot [4], [31]. For example, it is far from obvious that abstract geometric representations about the world, as envisioned by David Marr [20], can be extracted from real world images given the available resources. This has lead to a healthy move towards simpler representations and better exploitation of bodily interaction [23]. We should nevertheless keep in mind that many experiments which involve complex representations - even from the very beginning of AI - have considered the problem of grounding these representations. A typical example is the SRI Shakey robot [22] which was a model for many subsequent robotics efforts. So there is nothing in the notion of representation that makes them inherently not groundable. It is only that the grounding of representations is a very non-trivial and difficult technical problem involving a whole arsenal of statistical and pattern recognition techniques and that not every abstract representation can be grounded.

3. Representations are *symbolised* when there are external tokens (speech sounds, gestures, scratches on a piece of paper, configurations on a display) that are associated with the representation and used for external communication with another agent. The relationship is entirely conventional. Sender and receiver must agree, but there are in principle endless possibilities. The process of relating a representation to its symbolisation and vice-versa must be carried out autonomously by each agent. We say that a symbol is grounded iff its representation is grounded.

The relations considered so far are summarised in the semiotic square depicted in figure 1. By sensation, I mean the perceptual or motor data streams that directly connect the agent to the world. By representation I mean a conceptual representation useful for decision making, language or other cognitive tasks. The semiotic square is reminiscent of the semiotic triangle familiar from the semiotic literature [9] which relates world, concept, and symbol. The relation between a symbol and the world is the *reference relation.* The relation between a symbol and a concept is the *meaning relation.* In the philosophy of language literature, the reference relation and the meaning relation are studied as such, independently of how this relation is established by a cognitive agent. This kind of research in formal semantics is of interest when one wants to investigate how a symbol system can in principle be related to the world, but is a very different topic from the one considered here.

The problem of symbolisation is trivially solved by a calculator which transforms the internal representation of a number into an external representation on a display and which displays on the buttons the conventional representations of numbers so that users know which button to push. It is extraordinarily difficult in the case of natural language, because the conventions are not universal, they are open-ended, and involve a non-trivial multi-level mapping from representations to symbols.

```
sensation --------  world
    |                 |
    |                 |
    |                 |
representation ---- symbol
```

Figure 1: The semiotic square summarises the relation between world, sensation, representation, and symbol.

Just for clarification, it is perhaps important to sharpen in this context the notion of symbolic processing, as it has been used in AI. Symbolic processing means that a set of symbols, possibly without any relation to the world, such as postulated in a logical calculus, is mapped onto internal representations and the internal representations are processed conform to given symbol manipulation rules, for example the rules of natural deduction of the predicate calculus. The outcome of processing is then translated back into symbols. A logic programming systems such as PROLOG or a functional programming language like LISP provide this facility. They can handle millions of symbols and their compilers optimise for fast symbolic processing. This technology is of enormous value for building non-trivial cognitive agents but does not address in itself grounding nor learning.

4. *Learning grounded representations* means that the agent allocates states for certain representations but also - and more importantly - that the agent learns to use the representation appropriately, more specifically (1) the ability to relate the representation to the world through the sensori-motor apparatus, and (2) the use of the representation for some purpose, such as making a decision about the next action to take.

Learning a symbolisation means to acquire the relation between representations and symbols as required for communication in a specific community. The agent must acquire the ability to activate the intended internal representations given a set of symbols or to select a set of symbols to symbolise a particular representation.

Learning a semiotic system means to acquire both grounded representations and their symbolisations, i.e. the relation between world, sensation, representation, and symbol. This is the problem that the child faces when growing up in a language community and the problem that concerns us further in this paper. We take this problem to be equivalent to the *symbol grounding problem.*

2 Two approaches

The first approach to the symbol grounding problem is to follow a divide-and-conquer strategy. This assumes that there is first of all a process that learns grounded representations. Once the representations are in place, it then postulates a second independent process that associates symbols with the already acquired representations. It has been suggested that this is the way symbol grounding happens in humans [12] and various experiments have been done following this approach [8].

However a second approach is possible, as first suggested in Steels [27], [29], in which there is a strong *structural coupling* between the two. This means that learning representations and learning their symbolisation go hand in hand and influence each other. A representation that has been learned can be the subject of symbolisation but communication through symbols provides important feedback to representational learning.

In my opinion the structural coupling approach is the only viable way to explain the massive build up of representations and symbols that humans use and it can be profitably used in artificial systems. This approach seems paradoxical at first because instead of solving two difficult problems one by one, we try to solve both of them at the same time, which intuitively seems to be even more difficult. Here are the reasons why I nevertheless believe that a structural coupling approach is better.

There are many ways to learn grounded representations, but broadly speaking mechanisms fall into two classes: unsupervised or supervised learning. In the case of unsupervised learning, clustering techniques (possibly implemented as neural networks such as the Kohonen network) extract from a series of data invariances that are then equated with interesting representations. However we note two things: (1) not all representations of interest to a cognitive agent are reflected as invariants in sensori-motor data, and (2) often there is more than one possible way to cluster the data depending on the dimensions that are considered.

The latter generates the problem that if different agents each independently develop representations about the world, there is no guarantee that they arrive at mutually compatible representations. Today the experimenter carefully designs the features that are input to the learning system, carefully selects appropriate example sets, and then tweaks parameters until an appropriate clustering comes out. This is not quite the autonomous learning that we would hope for. To do otherwise however has turned out to be very difficult indeed, particularly if the sensory-input is really taken in its raw form, i.e. bitmaps captured by a camera, motor states, the audio signal directly coming from a microphone, etc.

In the case of supervised learning, the agent is given a series of cases as well as feedback whether the representations being developed are appropriate with respect to some task. Thus if the task is classification, the agent would be given examples and counterexamples, if the task is action in the world, the agent gets a feedback signal whether the action was successful (as in reenforcement learning algorithms). Because the task can incorporate some form of coordination with other agents, it is in principle possible to steer the acquisition of representations in such a way that they are compatible with those used by others, by incorporating in the feedback some element that is related to representation sharing. But the critical question here is: Where does the feedback come from? In real world circumstances, feedback is never direct and obvious, specifically not concerning internal representations. Feedback comes only through the *use* of a representation, generating the well known credit assignment problem. If the designer has to carefully determine feedback and prepare the example sets, then we are missing something fundamental.

There are many users of representations. For example for planning actions, particularly at a microlevel (like for grasping an object), the agent needs adequate categorisations of reality dedicated to that task. So action execution can be a possible provider of feedback and is undoubtly a force guiding concept formation. Language is another big user of representations because before anything can be said the world must be conceptualised in the way that has been lexicalised and grammaticalised in the language (and this can differ substantially from one language to another [24]. But language is not only useful because it provides representational feedback, it also helps a community of agents to settle on similar representations. This is why we have emphasised this in our work.

The next section describes the processes involved in some more detail.

3 Learning symbolisation

There exists a large literature on learning grounded representations which includes the algorithmic machine learning [21] and neural network literature as well as statistical pattern recognition techniques [3]. In contrast, it is only in the past few years that work has intensified studying how the relation between representations and symbols might be acquired. First in computer simulations (see examples in [14]) and then in experiments on robotic agents (see e.g. [33], Billard and Dautenhahn, [32]). So far, representations and symbols have been taken to be atomic, although some researchers have been considering structured representations and symbols ([15], [1], [30]). Generally speaking, this research on symbol learning converges on the same sort of solution [28]: Agents need an associative memory storing relations between representations and symbols. One representation can be associated with many symbols, and one symbol with many possible representations. Each association has a score which denotes how well the association reflects the consensus in the group, as far as the agent can tell. So the agent's lexicon consists of triples <r,s,m> with r the representation, s the symbol and m the score. The agents implement the following behaviors (called the Naming Game):

1. Speaker behavior. Suppose that the speaker needs to find a symbol for communicating representation R. The speaker collects all associations <r,s,m> in his lexicon where R = r, and picks out the one with highest score

m. s is then the symbol to be communicated.

2. Hearer behavior. Suppose that the hearer receives the symbol S. He then collects all associations <R',s,m> where s = S in his memory. The association with the highest score m is chosen and R' is hypothesised to be the meaning of s.

Assume that speaker and hearer then get feedback on whether R = R' (more on feedback later). If R = R' the game is successful and both agents increase the score of the association they used and decrease the score of competing associations. These are those associations of the speaker with the same representation but a different symbol, or those associations for the hearer with the same symbol but a different representation. If the game is not successful, both agents decrease the score of the association they used.

It can be shown that an agent acquires a set of conventions in a group given these behaviors (see [28]). Moreover if agents take turn being speaker and hearer and a speaker is allowed to invent a new symbol occasionally when he does not have an association yet to symbolise a particular representation, a set of conventions can establish itself from scratch in the population (figure 2). It would bring us too far from the topic of this paper to explain why this mechanism works, but in short the rules of the game embody a self-organising positive feedback loop. Associations that are successful become even more so because their score goes up, so that they become used even more frequently and hence propagate in the rest of the population. The dynamics is similar to that of ant societies self-organising a path or to increasing returns as studied in non-equilibrium economics.

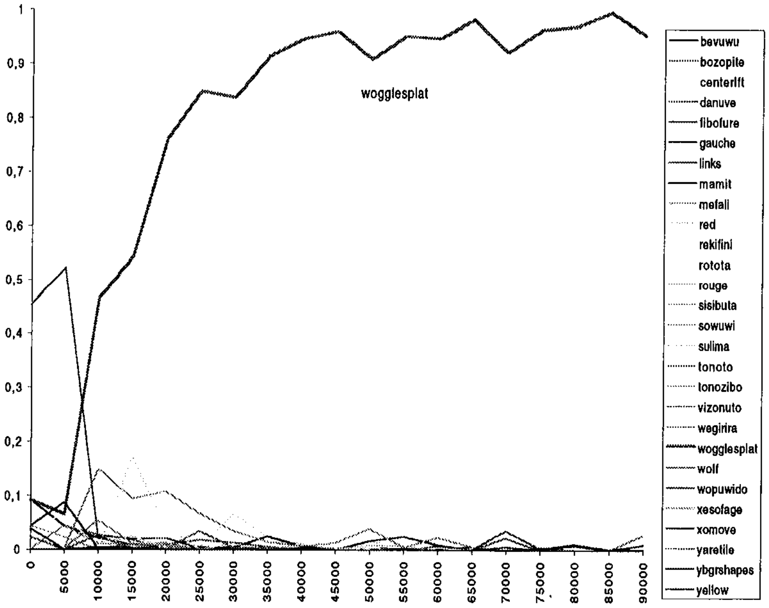

Figure 2: A meaning-form diagram which graphs for a specific meaning all the possible forms and their score. A winner-take-all situation is clearly observed. X-axis shows language games and y-axis the score of forms. There is a steadily growing population reaching 1500 agents towards the end

4 Learning semiotic systems

Let us now focus on the coupling of this symbol acquisition process with the learning of a grounded representation. I will take object-recognition as example task with instance-based learning ([21], ch 4) as the method to learn the representations of the objects. Let us assume that for every object known to the agent, there are a set of views stored as vectors in the n-dimensional feature space (figure 3). An object is recognised by a nearest neighbor match. When a new image has been captured and segmented, the closest stored view is retrieved and the object is found with which this view is associated. Feedback either confirms that the image indeed contains or does not contain the object. In both cases a new view can be stored to refine or correct future object-recognition. As in all supervised learning algorithms, the question is where the feedback is going to come from. This is where language becomes relevant.

When a child is learning to recognise the object ball, it is initially not clear at all what counts as a ball. A non-supervised clustering algorithm that would accidently hit the right internal representation of a ball is almost excluded. In practice, parents point to various examples of a ball (or various situations which generate different views of the same ball) and then say the word ball. The child

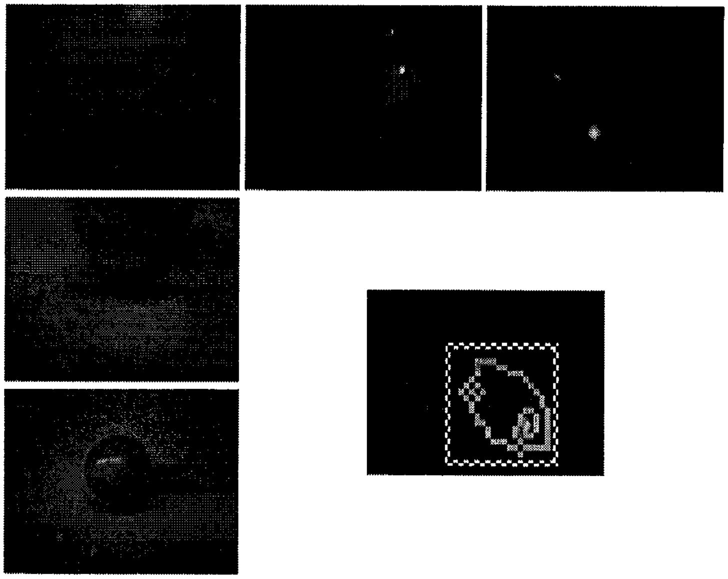

Figure 3: Different views of a ball stored as instances in the object memory.

may ask for a ball, and get it. She may be asked to give the ball and get feedback whether she gave the right object, She may hear another child say "Where is my ball" and get a response, and so on. Each verbal interaction where a symbol is used in a specific situation, is an opportunity not only to learn the symbol-representation relation but also the representation-sensation-world relation. Note that there is never explicit feedback on the representation itself, only on the total: representation + symbolisation as used in a specific verbal interaction. We call such a verbal interaction a language game. One language game relevant for object-recognition is the "Look Game".

The Look Game

1. *Shared attention.* By pointing, eye gazing, moving an object or other means the speaker draws the visual attention of the hearer to the topic. The speaker emits a word aiding to share attention, like "look", and observes whether the hearer gazes towards the topic. Based on this activity, both agents are assumed to each have an image that includes the object (although their views on the object will of course be very different).

2. *Speaker behavior.* The speaker then recognises the object with a nearest-neighbor match using his own object-memory, yielding a representation of this object R. The speaker then collects all associations <r,s,m> in his lexicon where R = r, and picks out the one with highest score m. s is then the symbol to be communicated.

3. *Hearer behavior.* The hearer receives the symbol S and collects all associations <R',s,m> where s = S in his memory. He also performs object-recognition using his own object-memory yielding the representation R".

4. *Feedback.*

4.1. If the hearer does not have an association in memory for S, this means that S designates a new object. So the hearer creates a new object O and stores the image segment as a view of O. He then stores a new association between O and S in this lexicon.

4.2. If the hearer has an association in memory for S with R' = R", then the score of this association is increased and the score of competing associations (i.e. other associations involving the same symbol S) are decreased. The image segment can be stored as a new view of R".

4.3. If the hearer does not have an association in memory for S with R' = R", then this means that S designates a new object. The hearer performs the same action as in 4.1.

When this game is being played the agents can be shown to acquire not only a memory to recognise the different relevant objects in their environment (relevant from the viewpoint of verbal interaction with others) but also to acquire the

Many similar language games can be invented. Each of them combine the use and learning of grounded representations as well as the use and learning of a symbolisation for these representations. We have put this methodology in practice in a number of experiments, using various robotic bodies, ranging from Lego-based small mobile robots [33], to steerable cameras in a "talking heads" experiment [29] and four-legged dog-like AIBO robots. In each experiment a certain game was defined and consecutive games played to test language learning and representation learning. In the more recent experiments on the AIBO robot, we have started to combine many different games and also addressed the problem of interaction with humans. We are also extending this methodology for humanoid robots.

5 Conclusions

This paper advocates a tight structural coupling between processes for learning grounded representations and learning symbolisations of them. Both constrain each others' degrees of freedom and enable the learner to get feedback about the adequacy of a representation. Our experiments in simulation and on real robots have sufficiently demonstrated that applications can be constructed in a straightforward way using these principles. At the moment we are specifically targeting research on language games for humanoid robots.

This work raises many additional interesting issues. For example, there has been a longstanding debate between nativists who claim that language learning amounts to learning labels for existing categories and relativists such as Whorf who claim that each language implies a different categorisation of reality. The structural coupling of concept formation and language acquisition advocated in this paper explains how a relativistic view is not only possible but unavoidable. If language enables and influences the learning of representations then it is easy to see how representations can become language specific. Of course representations are still strongly constrained by the world and tasks carried out in the world as well - they are not completely conventional or arbitrary. But they need not be innate to explain how they can become shared.

Acknowledgement

This paper was triggered by discussions with Steve Harnad as part of the committee concerning the Ph.D. thesis by Paul Vogt and by discussions with Erik Myin on the Wittgensteinian view of language. It is based on a large amount of technical work conducted in collaboration with Edwin De Jong, Frederic Kaplan, Angus Mcintyre, Joris Van Looveren, Jelle Zuidema and Paul Vogt. Frederic Kaplan was responsible for the AIBO implementation of language games.

References

[1] Batali, J. (1998) Computational Simulations of the Emergence of Grammar. In: Hurford, J. et.al. (1998).

[2] Bekey, G. and A. Knoll (2000) Proceedings of the First IEEE Workshop on Humanoids. Cambridge Ma.

[3] Bischop, C.M. (1995) Neural Networks for Pattern Recognition. Oxford Univ Press, Oxford.

[4] Brooks, R. (1999) Cambrian Intelligence : The Early History of the New AI. The MIT Press, Cambridge Ma.

[5] Byrne, A. and D.R. Hilbert (1997) Readings on Color. (Volume I: The Philosophy of Color. Volume 2: The Science of Color) The MIT Press, Cambridge Ma.

[6] Clancey, W. (1997) Situated Cognition : On Human Knowledge and Computer Representations (Learning in Doing - Social, Cognitive and Computational Perspectives). Cambridge Univ. Press, Cambridge UK

[7] Daelemans W. (1999) Memory-based language processing: introduction. - In: Journal of experimental and theoretical artificial intelligence, 11:3(1999), p. 287-292

[8] De Jong, E.D. (1999). Analyzing the Evolution of Communication from a Dynamical Systems Perspective. In Proceedings of the European Conference on Artificial Life ECAL'99, 689-693. Springer-Verlag LNCS, Berlin.

[9] Eco, Umberto (1968) La struttura assente. Milano 1968.

[10] Edelman, S. (1999) Representation and recognition in Vision. The MIT Press, Cambridge.

[11] Hardcastle, W. and N. Hewlett (1999) Coarticulation. Theory, Data and Techniques. Cambridge University Press, Cambridge.

[12] Harnad, S. (1990) The symbol grounding problem. Physica D 42: 335-346.

[13] Horswill, I. (1993) Polly: A Vision-Based Artificial Agent. In: Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI-93), Washington DC, MIT Press.

[14] Hurford, J, C. Knight and M. Studdert-Kennedy (eds.) (1999) *Approaches to the Evolution of Human Language.* Cambridge Univ. Press. Cambridge.

[15] Kirby, S. (1999) Function, Selection and Innateness: the Emergence of Language Universals. Oxford University Press, Oxford.

[16] Ladefoged, P. (2000) Vowels and Consonants: The sounds of the world's languages. Blackwell Pub. Oxford.

[17] Labov, W. (1994) Principles of Linguistic Change: Internal Factors. Blackwell Pub. Oxford.

[18] Lindblom, B., P. MacNeilage, and M. Studdert-Kennedy (1984) Self-organizing processes and the explanation of phonological universals. Linguistics, 21 (1).

[19] Marco, R. P. Sebastiani, and P. R. Cohen (2000) Bayesian Analysis of Sensory Inputs of a Mobile Robot. Proceedings of the Fifth International Workshop on Case Studies in Bayesian Statistics. Lecture Notes in Statistics, Springer, New York, NY, 2000.

[20] Marr, D. (1982) Vision. Freeman, San Francisco.

[21] Mitchell, T. (1997) Machine Learning. McGraw-Hill, New York.

[22] Nilson, N.J. (1984). Shakey the robot. SRI A.I. Center Technical Note 323.

[23] Pfeifer, R. and C. Scheier (1999) Understanding Intelligence. The MIT Press, Cambridge Ma.

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

[25] Pylyshyn, Z.W. (ed.) (1987) The Robot's Dilemma. Norwood NJ: Ablex Publishing Co.

[26] Pylyshyn, Z. (2000) Situating vision in the world. Trends in Cognitive Science, 4(5), May 2000, pp 197-207.

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

[28] Steels, L. (1997) The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1-35.

[29] Steels, L. (1999) How Language Bootstraps Cognition. In: Wachsmutt, I. and B. Jung (eds.) KogWis99. Proceedings der 4. Fachtagung der Gesellschaft fuer Kognitionswissenschaft. Infix, Sankt Augustin. p.1-3.

[30] Steels, L.: The Emergence of Grammar in Communicating Autonomous Robotic Agents. In: Horn, W. (ed.) Proceedings of ECAI 2000. IOS Publishing, Amsterdam. (2000)

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

[32] Steels, L. and Kaplan, F. (1998) Situated Grounded Word Semantics. In *Proceedings of IJCAI-99*, Stockholm. Morgan Kauffman Publishing, Los Angeles. p. 862-867.

[33] Steels, L. and P. Vogt (1997) Grounding Adaptive Language Games in Robotic Agents. In Harvey, I. et.al. (eds.) *Proceedings of ECAL 97*, Brighton UK, July 1997. The MIT Press, Cambridge Ma.

[34] Ullman, S. (1996) High-level Vision. Object Recognition and Visual Cognition. The MIT Press, Cambridge Ma.