How Experience of the Body Shapes Language about Space

Luc Steels[1,2]
[1] Vrije Universiteit Brussel,
Pleinlaan 2, 1050 Brussels, Belgium
steels@arti.vub.ac.be

Michael Spranger[2]
[2] Sony Computer Science Laboratory Paris
6, rue Amyot, 75005 Paris, France

Abstract

Open-ended language communication remains an enormous challenge for autonomous robots. This paper argues that the notion of a language strategy is the appropriate vehicle for addressing this challenge. A language strategy packages all the procedures that are necessary for playing a language game. We present a specific example of a language strategy for playing an Action Game in which one robot asks another robot to take on a body posture (such as stand or sit), and show how it effectively allows a population of agents to self-organise a perceptually grounded ontology and a lexicon from scratch, without any human intervention. Next, we show how a new language strategy can arise by exaptation from an existing one, concretely, how the body posture strategy can be exapted to a strategy for playing language games about the spatial position of objects (as in "the bottle *stands* on the table").

1 Language Games for Embodied Agents

Over the past decade we have been investigating in our group through what mechanisms open-ended language can be grounded in situated embodied interactions by performing computer simulations and doing experiments with physical autonomous robots [Steels, 2001]. Empirical research on natural dialog [Garrod and Doherty, 1994] has shown that language is not a static system. Instead language must be viewed as a complex adaptive system that is shaped and reshaped by its users, even in the course of a single dialog, in order to remain maximally adaptive to the expressive needs of the community, while at the same time maximising communicative success and minimising cognitive effort [Hopper, 1987].

We use a whole systems approach, as pioneered in behavior-based robotics [Steels and Brooks, 1994], meaning that all aspects of the problem, from perception and action, to categorisation, meaning selection, parsing, and production, as well as learning and embodied interaction, are operationalised and integrated into a single system, so that none of the components needs to be perfect but the whole system is stronger and more reliable than each of the parts taken separately.

Our methodology has crystallized around a set of central concepts. The first one is the notion of a *language game* [Steels, 1995]. A language game is a routinised interaction between a speaker and a listener out of a population whose members have regular interactions with each other. Each individual agent in the population can be both speaker and hearer. The game has a non-linguistic goal, which is some situation that speaker and hearer want to achieve cooperatively. For example, the speaker may want to draw the attention of the hearer to some object in the world. Speaker and hearer can use bits of language but they can also use pointing gestures and non-verbal interaction, and there is a shared common ground so that not everything needs to be said explicitly. A typical example of a language game is the *Color Naming Game*, which is a game of reference, where the speaker uses a color to draw the attention of the hearer to an object in the world [Steels and Belpaeme, 2005]. Humans users are able to play thousands of different games and even a single sentence may involve different language games simultaneously. For example, if somebody says "give me the red block" there is both an Action Game (asking somebody else to do something) and a Reference Game (drawing attention to an object in the world).

The second central concept in our work is that of a *language strategy*. A language strategy is a set of procedures that will allow members of a community to become and remain effective in playing a particular language game. It includes not only the interaction script, the turn-taking, joint attention, and other non-linguistic aspects of the game, but also procedures for perceiving, conceptualising, and acting upon reality, for producing and parsing utterances, for interpreting meaning back into the world, and for acquiring both the concepts, words and grammatical constructions needed in the game. In addition, a language strategy contains procedures for diagnosing failure in an interaction, for repairing the failure and for aligning conceptual and linguistic inventories so that speakers and hearers get maximally attuned to each other. When agents start to exercise a language strategy through a series of games in concrete situated interactions, each agent progressively builds a particular *language system*, i.e. a particular ontology, lexicon and grammar. For example, a strategy for playing the Color Naming Game will

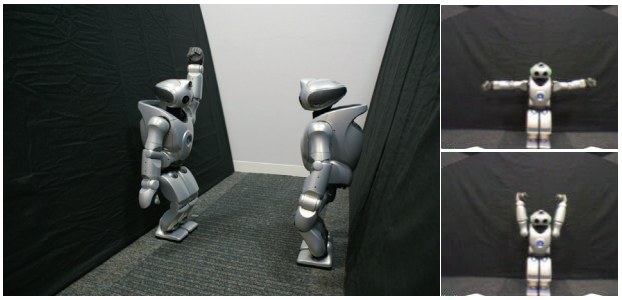

Figure 1: Experimental setup. Left: Two Sony QRIO humanoid robots face each other and play an Action Game. The speaker asks the hearer to achieve a certain posture, like raise the left arm, or stretch out both arms. Right: Two example postures as seen through the camera of the other robot.

enable each agent to build up and exercise an ontology of perceptually grounded color categories and a lexicon naming the categories. The language system of an individual agent is constrained by the situations this agent encounters and by the emerging linguistic conventions and meanings that are circulating in the population.

To explain how a shared language arises despite the absence of central control, we use a key concept from evolutionary biology, namely selectionism, but transposed to the domain of cultural evolution. Selectionism requires a mechanism to produce variants and an amplification of those variants that are more adapted to a particular set of challenges. Here we need first of all *linguistic selection*. The language systems of individuals inevitably show internal variation. Indeed, if different agents can invent new meanings and language based on their individual needs, there are unavoidably synonyms and alternative conceptualisations popping up. And because individuals have to learn meanings and language from each other, they may overgeneralise or adopt different hypotheses. The linguistic variants that are most effective should survive selection. We operationalise this in terms of alignment procedures that implement a feedback loop between communicative success and the 'score' of a particular language element (a concept, a word, a grammatical construction). If a word or grammatical construction is successful, its score is increased and hence its use re-enforced, whereas the scores of competing elements are inhibited. If the use of a linguistic element did not lead to a successful game, its score is decreased so that its further use is less likely. It has now been shown abundantly in many computer simulations and experiments that the use of this kind of lateral inhibition learning rule by individual agents leads a population to a shared language system and that this dynamics scales well with respect to meaning or population size [Baronchelli *et al.*, 2008].

But there is also variation in the possible language strategies that speakers and hearers may employ. Language strategies are assembled by recruiting and configuring generic cognitive mechanisms, such as associative memories, priming, hierarchical structuring, categorisation, perspective reversal, etc. [Steels, 2002] There are often many different ways to handle a particular class of linguistic challenges and so there is unavoidably variation within the set of strategies considered by an individual and hence in the population as a whole. To make sure that agents align their strategies we need again a selectionist process, that chooses among alternative language strategies the ones that are most effective. Language strategies are biologically implemented by networks of neuronal groups. And so we call this level of selection *neuronal selection*, to resonate with selectionist approaches to the brain [Edelman, 1987].

The criteria relevant for neuronal selection of a language strategy include whether it has been useful and successful for a particular type of language game and whether it requires less cognitive effort compared to competing strategies for the same game. There is again a cultural component because a language system can only be effective when other individuals have used the same or a similar strategy to build their own language system. And so, if a language user has persistent success with the language system built by a particular strategy, that re-enforces his usage of that strategy, whereas competing strategies get disfavored, automatically leading to convergence within the group.

We have been exploring this 'evolutionary linguistics' approach through several case studies [Steels and Loetzsch, 2009] and have developed a new formalism, called *Fluid Construction Grammar* [Steels and De Beule, 2006] that has the required flexibility and fluidity to deal with the representation, application, and acquisition of emergent grammars. In this paper we first illustrate the approach with a language game about body posture. Then we turn to the problem of the origins of language strategies. We look at one way in which a new strategy may arise, namely by exaptation. Due to space limitations, this paper can obviously not cover all the details of the hugely complex systems that are needed to operationalise language games on autonomous embodied agents. The interested reader is referred to Steels and Spranger [2008b; 2008a] and Spranger and Loetzsch [2009] for additional information about the experiments reported here.

2 A language strategy for body postures

All languages have words such as "stand", "sit" and "lie" which can be used to ask others to take on a certain body position (as in "sit down please") or to describe a body posture (as in "Katja sits down") [Newman, 2002]. In order to evolve this kind of body posture language, a number of deep fundamental problems need to be solved: Where do categories for body postures come from? How are actions to reach body postures learned? How is the visual recognition of a body posture learned? How can one acquire which visual body image is related to the action to reach that body posture? How do postures get shared between speaker and hearer? How can a language to talk about body postures grounded in the sensorimotor experiences of agents emerge and become shared? In line with the 'whole systems' approach, we argue that all these problems hang together. Each problem taken in isolation is a mystery, but language can be helpful both in acquiring relations between visual image schemata and body actions *and* in coordinating the ontologies and lexicons of dif-

Figure 2: 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 represented as a graph connecting the seven centralised normalised moments.

ferent individuals.

To evolve an ontology and lexicon for body postures and the visual image schemata they generate, we use the experimental setup shown in figure 1. Two humanoid robots face each other and play an Action Game: One robot (the speaker) asks another robot (the hearer) to perform an action. The speaker then observes the body posture achieved by the action and declares the game a success if the body posture is the desired one. Otherwise the speaker provides feedback by doing the action himself. For example, the speaker asks "raise left arm" and if the hearer indeed raises the left arm (as observed by the speaker) the game is a success. Otherwise the speaker may himself raise his left arm to show what he intended. Achieving this game on real robots with a pre-programmed ontology and lexicon is difficult enough, but we add the extra difficulty that the robots start without any pre-programmed notion of posture image schemata, actions to reach postures, nor words to ask for or recognise postures.

In a first experiment [Steels and Spranger, 2008b], we used kinesthetic teaching so that the robot can acquire the right motor commands to achieve a particular body posture. Kinesthetic teaching means that the experimenter moves the robot's body from a given position to a target body posture. The robot records the critical angles with which to construct the motor control program, the expected proprioception, and possibly additional feedback loops and information needed to replay the same action path in order to reach the same body posture later.

Learning and recognising visual image schemata involves three steps (figure 2). (i) The robot body must be segmented against the background, which is done here with classical running average foreground/background segmentation techniques. (ii) Features must be extracted that are shift and scale invariant. The features used here focus on the upper torso of the robot and are based on a binary version of the original image. They rely on another standard pattern recognition technique, namely normalised centralised moments [Hu, 1962]. Moments are a global description of shape, capturing the statistical regularities of its pixels for area, center of mass, and orientation. Centralised moments are invariant to translation and normalised moments invariant to scale. There are seven moments, so that an image schema of the upper torso is captured in terms of a feature vector with seven data points, represented as a graph (figure 2, right). (iii) These feature vectors are then classified using a standard prototype-based approach. A prototype of an image schema consists of typical points for the seven moments, as well as a minimum and maximum deviation from each point. The best matching prototype is found by nearest neighbor computation and the prototype is then adapted to integrate better the new instance.

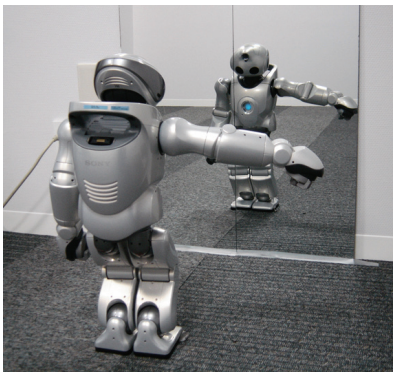

Figure 3: A humanoid robot stands before a mirror and performs various motor behaviors thus observing what visual body-images these behaviors generate.

To be able to learn the relation between body posture and visual image schemata, a robot stands before a mirror (figure 3), makes a gesture to reach a given body posture, and then gets a visual image which corresponds to a particular body posture. This generates the data for learning the association between body postures and image schemata. Because all robots have the same body shape, a robot can use visual body image schemata of himself in order to categorise the body image of another robot, after perspective reversal. Perspective reversal implies that the robot is able to detect the position of the other robot and is able to perform a geometric transformation to map the visual image of the other robot onto the canonical body position of itself [Steels and Loetzsch, 2009].

Once the robots have a reliable mapping between image schemata of postures and body movements, the lateral inhibition dynamics described earlier can easily solve the task of evolving a shared vocabulary. The score of the association between a word and a body posture is increased in a successful game and synonyms decreased. In an unsuccessful game, the score of the used association is decreased. Figure 4 shows the global behavior of a population of 10 agents after each individual has coordinated motor behavior and visual body-image through the mirror for 10 postures. 100 % success is reached easily after about 2000 games. After 1000 games, which means 200 games per agent, there is already 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.

In a second experiment [Steels and Spranger, 2008b] the robots no longer use a mirror to learn about the relation between a visual body image schema and their own bodily action (and vice versa) but coordinate this relationship through language. Language will enforce coordination because if a speaking robot R_1 asks R_2 to achieve a posture P using a word W, the game will only be successful if for R_2, W is as-

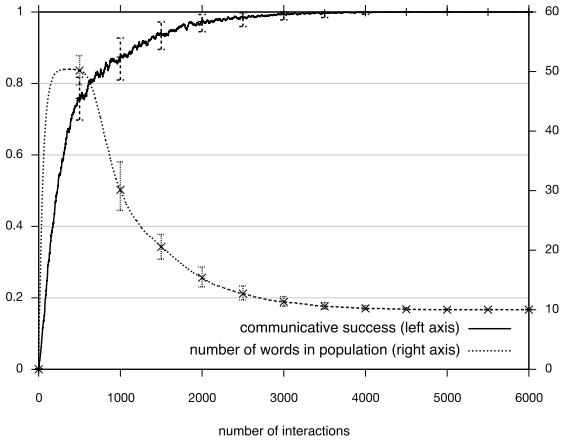

Figure 4: Results of the Action Game played by a population of 10 agents for 10 postures after coordinating image schemata and motor control through a mirror. The x-axis plots the number of language games. The y-axis shows the running average of communicative success and average lexicon size.

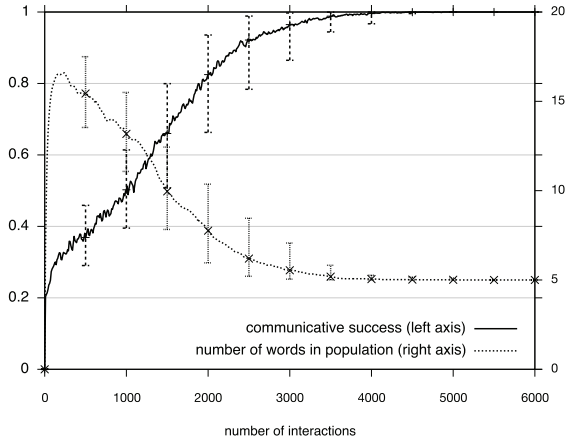

Figure 5: Results of the Action Game played by a population of 5 agents for 5 postures whereby robots no longer use a mirror for learning which image schema relates to which action. The x-axis plots the number of language games and the y-axis the running average of communicative success and average lexicon size.

sociated with an action A so that P is indeed achieved. Results of this second experiment for a population of 5 agents and 5 postures show that the coordination indeed happens (figure 5). 100 % success is reached after about 4000 games (1600 games per agent) and stays stable. The speed of convergence can be improved significantly if the hearer uses his own self-body model (a stick-figure simulation of the impact of motor commands on body parts) in order to guess which actions have to be performed in order to reach the body posture that is shown by the speaker [Steels and Spranger, 2008b]. These results are remarkable because there is a kind of chicken and egg situation. A population of agents can only develop a shared language by having shared categories for bodily image schemata and knowledge about which motor control programs generate which bodily images. But at the same time, language is used here as a mechanism to get categories to be shared and to acquire the perception/action mirror system.

3 Exaptation of language strategies

Clearly a population of agents can be shown to bootstrap remarkably quickly a shared grounded language system, once they all use the same language strategy. This is possible because a language strategy contains not only all the relevant mechanisms to conceptualise and verbalise meaning, but also ways to efficiently get data to learn from, learning algorithms optimised for the relevant competence, and alignment procedures. A language strategy comes with diagnostic and repair strategies so that failure in a game is never catastrophic but rather an opportunity for learning. But there is a price for this efficiency. To start approaching the complexity of human language thousands of strategies will be needed, and this raises the question where language strategies come from. How do they come into existence, how do they survive neuronal selection within the individual, and how they become shared in the group? We use again inspiration from biology. It is well established that many biological traits first develop for one purpose and then get exapted for another purpose. For example, insect wings developed first for regulating heat before they became exapted for flight, lungs developed in fish first for catching oxygen before they became exapted as swim bladders. There is plenty of evidence from all languages of the world that language strategies developed for one semantic domain and set of tasks may become exapted for another domain. A very nice and well documented example is the exaptation of body posture language for describing the spatial stance of an object, as in ("the bottle it stands on the table" or "the paper *lies* on the table") [Lemmens, 2002]. This kind of metaphorical transfer is by no means common to all languages. For example in French the phrase "le papier est *couché* sur la table" (literally "the paper lies on the table") sounds ridiculous. And even if languages adopt metaphorical transfer of body posture language, there is still considerable variation in how the transfer is enacted, suggesting that the prototypical associations of body postures are culturally shaped. For example, in Dutch you can say "Het stadhuis ligt aan de markt", literally "The Town Hall *lies* on the market", to mean "The Town Hall is located in market square". Body posture language is often extended even further to entirely abstract domains such as language about the economy, as in Dutch: "Het land zit in een recessie." (literally: The country *sits* in a recession). Interestingly, the metaphorical transfer does not always work in the other direction. Words for describing or requesting a given body posture (like "stand", "sit" and "lie") are reused for describing the position of an object but not the other way round.

We now briefly report an experiment, carried out on the Humboldt A-series robots [Spranger and Loetzsch, 2009; Spranger *et al.*, 2008], that exapts the body posture strategy effective in Action Games to a spatial position strategy for Reference Games. In a reference game, agents try to draw attention to an object in the scene before them. This object is called the topic. Reference games require that the speaker first finds a property that is distinctive for the topic with re-

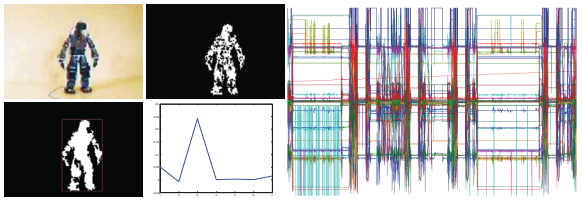

Figure 6: Left: Aspects of visual processing of body postures. We see (top) the source image and the image after segmentation, and (bottom) the binary image and feature signature (similar to figure 2). Right: Snapshot of 86-dimensional sensori-motor streams obtained while motor babbling.

Figure 7: Reuse of visual categories developed for body postures for categorising object positions. This image shows from left to right: the source image, object contours in a binary image, and feature signature of the centralised normalised moments of the two objects in the scene.

spect to other objects in the context, such as a color, a shape, a texture, or a spatial position. The speaker then names this property and the hearer decodes the name and applies the property to his own perception of the scene in order to identify the intended topic. The game is a success if the hearer points to the topic originally chosen by the speaker. We show how the visual image schemata of postures and the names for these schemata as developed in the context of the Action Game can be exapted to be successful in the reference game *without any need for additional learning*.

A language strategy can be exapted by reusing not only all procedures for perception, conceptualisation, production and comprehension but also by reusing the language system, i.e. the perceptually grounded ontology, the words, and the grammatical constructions, that have already been built with this strategy for another domain and another purpose. Research on the metaphorical reuse of posture verbs, [Lemmens, 2002] has shown that different prototypical associations of a posture may play a role to reuse a posture word: the visual image schema of the posture, the stability of the posture, the relation to the horizontal, the effort needed to sustain it, what the posture is needed for (e.g. walking, sleeping, eating), etc. We here focus only on the visual features of a posture, although other associations like needed effort, stability, or relation to horizontal could also potentially be operationalised on physical robots.

Robots first generate rich sensorimotor experiences by motor babbling and behavior exploration before a mirror (as in figure 3). Motor babbling means that the robot executes its behavioral programs including random body movements constrained by the physical limitations of the robot, and records at the same time motor commands, proprioceptive feedback, as well as the visual features of the body in terms of the centralised normalised moments discussed earlier. This activity generates trajectories in an 86-dimensional space (figure 6).

The target body postures are no longer scaffolded by an experimenter through kinesthetic teaching. Instead robots use a K-means clustering algorithm to divide the sensori-motor space into areas around a centroid prototype. Each prototype categorises a particular bodily action, not only in terms of joint angles and expected proprioceptive feedback but also in terms of the visual features of the action involved, so that they can re-enact a motor control sequence to reach a body posture which has the desired visual features. The visual features are the same as used earlier, i.e. centralised normalised moments (figure 6, left). Given this mapping from visual image schemata to motor control, the robots can use the same lateral inhibition dynamics as discussed earlier to coordinate an inventory of body postures and self-organise a lexicon for them. Results shown in figure 8 demonstrate that the agents indeed arrive again at a sufficiently shared vocabulary and ontology to have total success in the game.

After 1000 language games, agents progressively start to play reference games about scenes consisting of several objects that stand or lie on the ground. Figure 7 shows (left) an example scene with a block lying down and a cone standing upright. The speaker selects one object to draw the hearer's attention to, chooses a description that is distinctive for the topic and names the description. The description could be the color, or the shape, but here we are interested in the stance of the object. A reference game is a success if the hearer has correctly pointed to the object, otherwise the speaker points to the object. The scenes are analysed by each robot using the same algorithms as used earlier: there is color-based segmentation, transposition to a binary image, identification of the object contour, and computation of the centralised normalised moments (figure 7, right). Based on the feature signature for each object, they can then use their existing ontology of prototypical image schemata for body postures to categorise each object and attempt to find a distinctive description. In the present case, the left object fits with the prototype of a lying body posture and the right object with that of a standing body posture. Given this categorisation the same lexicon can be reused as for body postures. So the speaker now says the equivalent of "stand" to describe the right-most object in the scene of figure 7. The hearer retrieves the image schema associated with this word and can use it to interpret the scene and thus identify the topic originally intended by the hearer.

In the experiment shown in figure 8, the probability of playing reference games increases after 1000 body posture games. By that point an ontology and lexicon for talking about body action is already firmly in place. From 1500 games onwards agents have an equal chance to play a reference game or an action game. Results (figure 8) show that switching to a reference game has no effect whatsoever on communicative success, which implies that the agents are entirely successful in exapting the ontology and lexicon evolved for body postures to the description of the spatial position of objects.

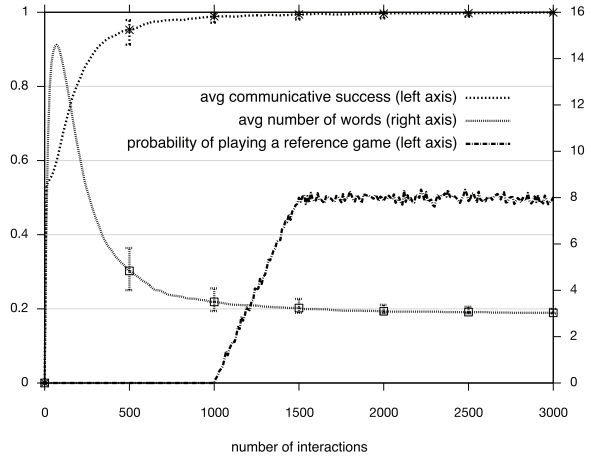

Figure 8: The first series of 2000 interactions concern a body posture game, followed by a series of Reference Games based on spatial position. The x-axis shows the series of games and the y-axis communicative success and lexicon size.

4 Conclusions

This paper has illustrated how a population of embodied agents can use language strategies to build a language system for being successful in a particular language game, in this case an Action Game about body postures, and how a new language strategy may arise by exaptation from an existing one, in this case the metaphorical use of body postures for describing the stance of objects. We make no specific claims about the specific solutions for each component that have been adopted. There are other ways to do segmentation, image recognition, motor control learning, associative memory, and so on, and for other language games other cognitive mechanisms will be needed. The experiment illustrates our general methodology to achieve open-ended grounded language on autonomous robots and many more experiments are needed to better understand the mechanisms behind the recruitment of generic cognitive mechanisms for building strategies and for the linguistic and neuronal selection that drives a population towards effective coherent language systems.

Acknowledgments

The research described here was sponsored by the Sony Computer Science Laboratory in Paris with additional support from the EU Cognition project ALEAR of the European Commission (IST-FET). We thank Masahiro Fujita and Hideki Shimomura from Sony's Systems Technologies Laboratory in Tokyo for providing access to the QRIO robot and Manfred Hild and his team at Humbold University for providing access to the Humbold A-series robots.

References

[Baronchelli *et al.*, 2008] A. Baronchelli, V. Loreto, and L.x Steels. In-depth analysis of the Naming Game dynamics: the homogeneous mixing case. *International Journal of Modern Physics C*, 19(5):785–812, 2008.

[Edelman, 1987] G.M. Edelman. *Neural Darwinism: The theory of neuronal group selection*. Basic Books New York, 1987.

[Garrod and Doherty, 1994] S. Garrod and G. Doherty. Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, 53(3):181–215, 1994.

[Hopper, 1987] P. Hopper. Emergent grammar. *Papers of the 13th Annual Meeting, Berkeley Linguistics Society*, pages 139–157, 1987.

[Hu, 1962] M.K. Hu. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179–187, 1962.

[Lemmens, 2002] M. Lemmens. The semantic network of dutch posture verbs. In *The linguistics of sitting, standing and lying*. John Benjamins, 2002.

[Newman, 2002] J. Newman, editor. *The linguistics of sitting, standing and lying*. John Benjamins, 2002.

[Spranger and Loetzsch, 2009] M. Spranger and M. Loetzsch. The semantics of sit, stand, and lie embodied in robots. *CogScie-2009, Amsterdam*, 2009.

[Spranger *et al.*, 2008] M. Spranger, C. Thiele, and M. Hild. A modular architecture for the integration of high and low level cognitive systems of autonomous robots. *IEEE/RSJ 2008 IROS, Workshop on Current Software frameworks in Cognitive Robotics*, 2008.

[Steels and Belpaeme, 2005] L. Steels and T. Belpaeme. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, 28(04):469–489, 2005.

[Steels and Brooks, 1994] L. Steels and R. Brooks. *The artificial life route to artificial intelligence: Building situated, embodied agents*. Lawrence Erlbaum Associates, 1994.

[Steels and De Beule, 2006] L. Steels and J. De Beule. Unify and merge in fluid construction grammar. In *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication*, pages 197–223. Springer, 2006.

[Steels and Loetzsch, 2009] L. Steels and M. Loetzsch. Perspective alignment in spatial language. In *Spatial Language and Dialogue*. Oxford University Press, 2009.

[Steels and Spranger, 2008a] L. Steels and M. Spranger. Can Body Language Shape Body Image? *Artificial Life XI*, 11:577–584, 2008.

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

[Steels, 1995] L. Steels. A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332, 1995.

[Steels, 2001] L. Steels. Language games for autonomous robots. *IEEE Intelligent systems*, 16(5):16–22, 2001.

[Steels, 2002] L. Steels. The recruitment theory of language origins. In *Emergence of Communication and Language*, pages 129–151. Springer, 2002.