The Semantics of SIT, STAND, and LIE Embodied in Robots

Michael Spranger (spranger@csl.sony.fr)
SONY Computer Science Laboratory Paris, 6, Rue Amyot, 75005 Paris, France

Martin Loetzsch (martin.loetzsch@gmail.com)
VUB AI Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

Abstract

In this paper we demonstrate (1) how a group of embodied artificial agents can learn to construct abstract conceptual representations of body postures from their continuous sensorimotor interaction with the environment, (2) how they can metaphorically extend these bodily concepts to visual experiences of external objects and (3) how they can use their acquired embodied meanings for self-organizing a communication system about postures and objects. For this, we endow the agents with cognitive mechanisms and structures that are instantiations of specific ideas in cognitive linguistics (namely image schema theory) about how humans relate motor and visual space. We show that the agents are indeed able to perform well in the task and thus the experiment offers a concrete operationalization of these theories and increases their explanatory power.

Keywords: cognitive semantics; image schemas; metaphor; autonomous robots; lexicon acquistion; language games

Introduction

Speakers of Germanic languages are strongly committed to using posture verbs for describing the location of human subjects whereas for example speakers of Romance languages do not. Speakers of English for example would rather say "He sits on the couch" than "He is on the couch". Some of these languages have also extended the usage of these words to the posture of animals and even to describe the "posture" of objects. Furthermore, Dutch has an extremely productive use of these verbs in a metaphorical sense. In Dutch one *sits* in an economical crisis.

The following example from Lemmens (2002b) highlights the typological differences in the usage of *lie* across some example languages (* denotes unacceptable usage):

1.(a) There ***are*** clothes on the counter. (English)
(b) ***Il y a*** des vêtements sur le comptoir (French)
(c) Det ***är/finss*** kläder pådisken. (Swedish)
(d) * Er ***zijn*** kleren op de toonbank. (Dutch)

2.(a) The clothes are ***lying*** on the counter. (English)
(b) * Les vêtements ***couchent*** sur le comptoir (French)
(c) Det ***ligger*** kläder pådisken. (Swedish)
(d) Er ***liggen*** kleren op de toonbank. (Dutch)

In Dutch as a Germanic languages it is more or less obligatory to use a posture verb for describing the location or posture of the clothes, whereas in French it is unacceptable to use the corresponding form *coucher* ("lie"). On the other hand, Dutch speakers will feel uncomfortable about using *zijn* ("to be") to express the position of the clothes on the table, where in French the usage of *être* ("to be") or *se trouver* ("be found") is clearly the preferred coding strategy. English and Swedish are less obligatory in their use of posture verbs and allow for both coding strategies in this example.

As suggested by (Lemmens, 2002a), the use of posture verbs can be grouped into three basic scenarios: First, *postural* uses (describing human postures) are permitted in most Germanic languages and are even obligatory in some of them (Oosten, 1984). They are grouped around the three central meanings SIT, STAND, and LIE, which have been extensively analyzed with respect to their underlying semantics. Almost all authors have found visual features such as 'maximally vertically elongated' or 'resting on one's feet' to be connected to STAND, 'maximally horizontally elongated' or 'on one of the sides' to LIE, and 'more or less square' or 'on the buttocks' to SIT (Lemmens, 2002a; Borneto, 1996; Newman, 2002). Second, *locational* uses are semantic extensions of these anthropocentric perceptual schemas. The visual property of vertical/horizontal elongation, for example, has been extended to locate any entity in space (see the examples above). Another example is the use of posture verbs for conveying orientation (see Borneto, 1996 for an analysis of the German *liegen* and *stehen*). Third, *metaphorical* extensions are found in some languages for locating concrete entities in an abstract space (Lemmens, 2002a). The previously mentioned use of SIT for being in a financial crisis is an example of the latter.

Cognitive linguists have argued that these verbs are metaphorically extended from the domain of bodily posture (SIT/LIE/STAND) using notions such as image schemas (e.g. Johnson, 1987; Croft & Cruse, 2004): Image schemas are pieces of knowledge that are represented as patterns of reoccurring bodily experiences. They emerge from continuous sensorimotor activity, that is, they are developed and extended as we move through the world, direct our attentional focus, manipulate objects, orient ourselves spatially and temporally, and so on. Furthermore, they are agnostic to sensory modalities and structure not only bodily but also non-bodily experience via metaphors. Using image schema theory, Gibbs et al. (1994) considered for instance `BALANCE`, `VERTICALITY` and `RESISTANCE` as the key image schematic accounts for the polysemous meanings of the English *stand* and showed that "... people tacitly believe there are significant connections between their recurring bodily experiences and the various meanings of the polysemous word stand. We argue that theories of psychological semantics should account not only for the organization of polysemous words in the mental lexicon, but must also be capable of explaining why different senses of a word make sense to people in the way they do." Others

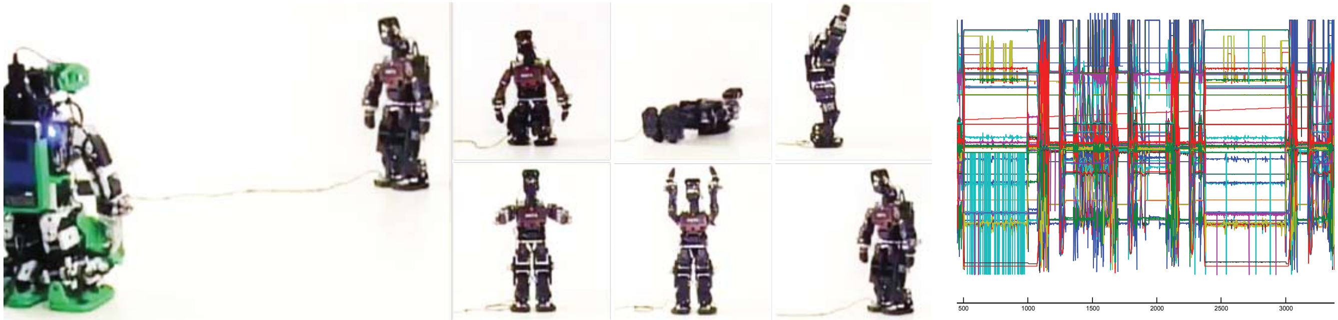

Figure 1: Experiment phase I. Left: A robot performs a series of actions in the world and is observed by the other robot on the left. Middle: Examples of the actions performed by the robot. Right: The resulting raw data stream, containing both the proprioceptive data of the performing robot and the visual features extracted by the observing robot.

like Lemmens (2004) have stressed the productivity of additional image schemas such as `‘resting on ones base’`, `CONTAINMENT` and `CONTACT` to explain some of the encountered extensions.

Although these explanations are often very detailed and closely linked to questions of representation and processing, only few have tried to operationalize the underlying mechanisms and to turn them into a computational model (e.g. Amant et al., 2006) and nobody has tackled the question of grounding in actual physical robots – which brings us to the main topic of this paper. We will show how robotic agents can construct image schematic categories from sensorimotor data and extend them to other objects. In order to validate whether the emerging representations are indeed meaningful, we will give them the task to agree on a set of names for these meanings in series of communicative interactions. This methodology is very similar to other experiments of our group on the emergence of linguistic communication systems in physical robots, of which recent examples are on flexible lexicon formation (Wellens, Loetzsch, & Steels, 2008), marking of spatial perspective (Steels & Loetzsch, 2009), learning of case grammars (van Trijp, 2008), or the relation of visual and motor space through language (Steels & Spranger, 2008). Throughout the next sections we will outline the experimental setup, perception and categorization mechanisms, the communication task and finally the results.

Experimental setup

We use the ”A-series” humanoid robots developed in the AI lab of the Humboldt University Berlin as an experimental platform. They are equipped with pan/tilt cameras, servo motors and acceleration sensors. For onboard processing and thus autonomous operation, the robots feature a PDA on the back and distributed sensor and processing boards spread out across the body, linked via a system bus. The robot's software architecture provides integrated mechanisms for balancing motion control, vision and behavior.

The experiment consists of two stages: In a first *developmental* phase (see Figure 1), the robots interact with their environment and form posture categories from that. For this, one robot performs a series of actions over a time span of 10 minutes. These actions include walking and turning motions, arm gestures while the robot is standing, lying, getting up after falling, some sitting motions and means to switch between these actions. During that, the robot continuously perceives proprioceptive data from its internal sensors. In order for the robot to also have access to visual appearances of the performed actions, a second robot perceives the scene through his camera and we provide the first robot with that data stream. Giving one robot access to what another one sees might seem very unnatural, but it overcomes our robots' lack of a geometric body model that would allow them to determine how it looks when they perform an action (another possibility to provide a robot with proprioceptive and visual data would be to let the robot perform actions in front of a mirror, as it was done in Steels & Spranger, 2008). We will describe further below how image schemas are extracted from

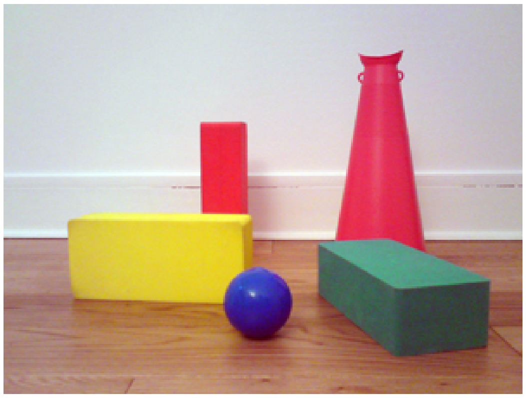

Figure 2: Example objects in the environment of the agents. The objects have different verticality and horizontality features depending on the view point of the agent. However, some always have a strong verticality component, like the red cone in the back.

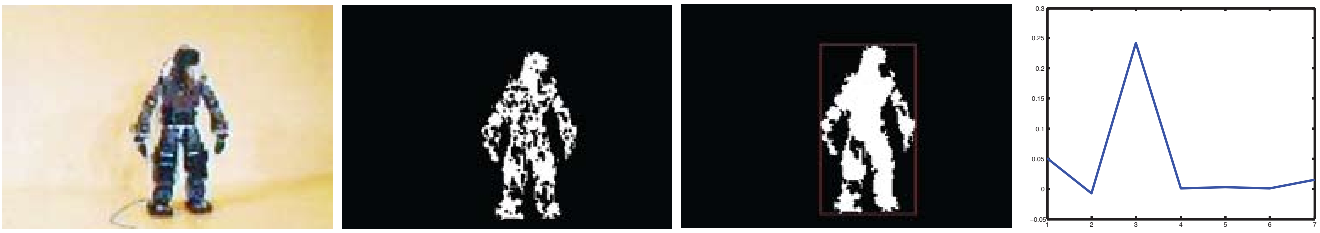

Figure 3: Extraction of visual features. First: the original image stemming from the onboard cameras of the robot. Second: foreground/background subtracted image. Third: the connected component processing unit has identified a single connected area, depicted by the bounding box. Fourth: the seven visual features computed for the connected region, shown as a line plot.

this combined data stream. Then, in a second *communication* phase, the robots test their acquired categories in two types of *language games*: first about robot postures that they perceive and then about objects that a human experimenter presents to them (see Figure 2).

Sensory experiences

We call the raw uncategorized proprioceptive and visual data stream that the robots perceive as they move around in their environment *sensory experiences*. The proprioceptive features are gathered in every time step from the acceleration sensors and the motors. The acceleration sensors report two dimensional values that not only reflect the movement of arms and the acceleration of limbs, but also gravitational forces, which allows the robot to directly sense its orientation in space. Furthermore, for each of the motors the actuated and the actual position of the motor as well as the torque applied to the motor is sensed. Actuated position is the value that was requested by the behavior control programs, and it is often different from the actual position because the motor does not necessarily reach it. Notice that humans have very similar proprioceptive capabilities. When we use muscles to reach certain positions we have means to detect whether this position was reached and we have the inner ear for orienting ourselves in space.

For the visual part of sensory experiences, we do not use the whole image but compute a set of seven translation and scale-invariant shape features for objects found in the image (see Figure 3). Every 80 ms the digital camera of the robot provides a new two dimensional image. The vision system then first uses running average background subtraction to detect robots or other objects as connected regions that sufficiently differ from background. These connected areas are, after being noise filtered with morphological operators, processed as belonging to separate objects in the world.

Shape descriptor features are then computed from the image pixels contained in a connected area using *centralized normalized moments features*: The central moment of order $p+q$ is computed as follows (Hu, 1962):

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

with $\bar{x}$ and $\bar{y}$ following from the raw moments M_{pq}

$$\bar{x} = \frac{M_{10}}{M_{00}},\ \bar{y} = \frac{M_{01}}{M_{00}}$$

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

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

where in all formulas I is a function equal to 1, if the pixel x,y is part of the shape the feature is computed for (or 0 otherwise). Given these definitions, the normalized central moment η_{pq} of order $(p+q)$ is computed (Mukundan & Ramakrishnan, 1998)

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{1+\frac{p+q}{2}}}$$

The shape descriptor is consequently agnostic of the exact position of the object in the image and in the visual field, due to being centralized with respect to the center of the shape and normalized with respect to the total number of pixels (mass of the area). Nevertheless these features are powerful for quantifying basic relationships such as the correlation between vertical and horizontal elongation η_{20}, η_{02}.

A sensory experience s at time t is a vector $s_t = \begin{pmatrix} f_1 & \cdots & f_n \end{pmatrix}^T$, where n is the dimensionality of the sensory experience, and $f_1, \ldots, f_n$ are the values of the proprioceptive sensors and the computed visual features.

Categorization

To organize the continuous flow of sensory experiences into categories of bodily related meaning, we employ an unsupervised machine learning clustering technique called K-means (Lloyd, 1982). Unsupervised means that K-means autonomously finds clusters in the incoming data without requiring labeled sets of training data created by a human experimenter. The computed categories are hence grounded in the preconceptual, raw stream of sensory experiences and in nothing else.

K-means receives as input the number k of clusters that one would like to find and m unlabeled, unclassified data points.

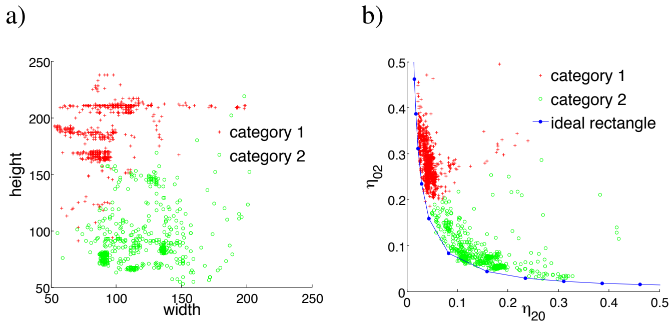

Figure 4: Categorization results K-means $k = 2$ of the raw sensory experience stream (see Figure 1). We focus on four dimensions: bounding rectangle width and height (which are not part of the centroid, Figure a), as well as two visual features, that quantify vertical and horizontal elongation (Figure b).

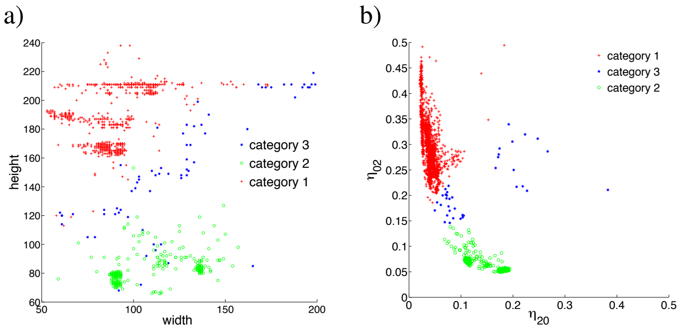

Figure 5: Categorization. K-means of the raw sensory experience stream (see Figure 1) for $k = 3$. We center on four dimensions: bounding rectangle width and height (which are not part of the centroid, Figure a), as well as two visual features, that quantify vertical and horizontal elongation (Figure b).

The result is a set of k categories, which are represented as centroids (the mean/prototype of a cluster) and that can be used to partition the data in a metric space. The algorithm starts by first randomly selecting k seed centroids. All data points of the same class are closest to the same centroid. In a second phase the algorithm iterates until convergence. Unlabeled data points are classified as belonging to the centroid with the smallest Euclidian distance and the centroids are updated by shifting the mean value given all data points of the same class. The algorithm terminates as soon as centroids are not moved anymore, which means that the class of every data point in the unlabeled data is not subject to change. The outcome of the algorithm is a set of k centroids, which cluster the sensorimotor space into k disjoint sets of experience. In other words, every point in the sensorimotor space, every sensory experience belongs to exactly one category.

The computed centroids can immediately be used to categorize new incoming sensory experiences that were previously not encountered by the system (see Figures 4 and 5). Given a set of k centroids $C_1,..,C_k$, a new sensory experience can be classified using a Minimum Euclidean Distance Classifier:

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

where the C_i are the k centroids computed by the K-means algorithm and s denotes a new sensory experience. Notice that partial experiences/stimuli can be classified as well. That is, given the centroids $C_i = \begin{pmatrix} f_1^i & .. & f_n^i \end{pmatrix}$, where n is the dimensionality of the centroid and f_j^i is the value of the feature channel j of centroid i, we can, for instance, categorize a visual stimulus $s = \begin{pmatrix} f'_{l_1} & .. & f'_{l_{\hat{n}}} \end{pmatrix}$ with $\hat{n} \leq n$ and $l_1,..,l_{\hat{n}} \in \{1,..,n\}$ being the index subset of visual features of the d feature channels of the centroids, using the feature dimensions of the centroids relevant to that stimulus. Similarly to the n dimensional stimulus case, we define

$$\text{class}(s') = \arg\min_i ||C'_i - s'||,$$

where C'_i is constructed from the i-th centroid by selecting the feature dimensions present in the stimulus $C'_i = \begin{pmatrix} f_{l_1}^i & .. & f_{l_{\hat{n}}}^i \end{pmatrix}$. In other words, focussing only on the dimensions present in the stimulus it is possible to infer the closest centroid with respect to those dimensions. However, since centroids are in fact central points in the complete sensorimotor space this also activates the other features not included in the stimulus. For example a robot perceiving another robot performing bodily actions, can classify the visual stimulus with respect to his inventory, thereby effectively activating the proprioceptive part of the best matching centroid. The perceiving robot ergo has a sense of what the other robot is doing or what the internal proprioception of the performing robot could be like.

Each agent of the population (for the rest of the paper we deal with 10 agent populations) is presented with different but similar sensorimotor streams. Consequently, the categories constructed by agents are similar but not identical. The categorization results for K-means clustering ($k = 2$) of the raw sensory data stream for a single agent are displayed in Figure 4. The graphs clearly show that the resulting categories (centroids) establish a width-height correlation, linearly separable well above the square dimension line. The blue line in Figure 4b shows an ideal rectangle in the visual field of the robot changing its width and height while keeping its area constant, for illustration purposes. The point in the cusp of that curve is an ideal square. Category 2 (which we would call LYING) ranges from strong horizontal elongation to square, to a little vertical elongation. Category 1 covers sensory experiences which are strongly vertically elongated, and thus visually corresponds to STANDING.

For $k = 3$ (see Figure 5, categories are further split by cat-

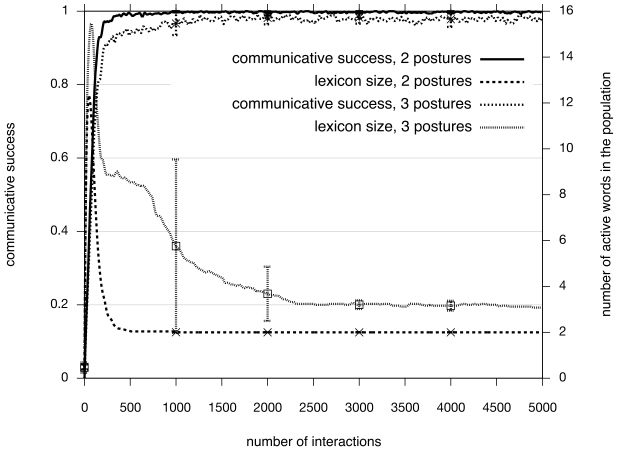

Figure 6: Experimental results for a population of 10 agents. Average communicative success and lexicon size for series of 5000 language games, averaged over 25 experimental runs. In the first 1500 interaction agents play language games about postures. Starting from interaction 1500 every second language games is about objects.

egory 3 (which resembles SIT) acting as a separator between the two basic postures LIE and STAND. Category 3, which lends itself to the SITTING interpretation covers a narrow margin with a little bit stronger vertical elongation than horizontal elongation, which stems from the fact that indeed sitting for this robot looks like a thick vertical elongated rectangle. However, there are certain problems with category 3 (in the $k = 3$ case) not shown here. The SIT category is almost useless since it is cramped between the two major and clearly separable categories LIE and STAND. Moreover, when looking at the postures categorized with that category in the proprioceptive space, quite a number of postures are not what we would call SIT, but are closer to lying or even standing.

Language Games

Categorization itself is useless unless the categories provide some benefit for the agent in its interaction with the environment. The testbed we are going to use here is linguistic communication: the agents learn to use words that denote their acquired categories in a communicative scenario, which could be for example useful for commanding each other to perform a certain action or for drawing attention to an object in the environment. In order to be able to do that effectively, agents have to develop a shared lexicon, linking categories and words in similar ways across all agents. How this can be achieved is nowadays well known (Steels, 1995, 2001): Populations of agents engage in series of *language games*, which are local communicative interaction with a routinized dialogue pattern. Each agent in the population maintains its own private (initially empty) lexicon, which is learnt and updated as a side effect of a game.

At the beginning of an interaction, two agents are randomly drawn from the population and assigned the roles of speaker and hearer. Both robots are shown either a robot in a specific posture or another object (one out of a set of colored bricks of different sizes and each time in a different orientation, see Figure 2 for examples). The speaker then uses the Minimum Euclidian Distance Classifier (as described above) to find the category (which was learnt in the developmental phase) that is closest the sensory experience of the robot/object and retrieves the name for that category with the highest score from his lexicon and speaks it out to the hearer. When he does not have a name yet for this meaning, he invents a new random word for it and stores it in his lexicon. The hearer looks up that word in his own lexicon and checks whether the category that he associates to the word (with the highest score) is the same as the category that is for him closest to the sensory experience. If that is the case, he signals agreement with the description and the interaction is a communicative success, otherwise it is a failure. When the hearer does not know the word, he also signals a communicative failure and associates the new word to his conceptualization of the scene. Depending on the outcome of the game, both speaker and hearer increase the score of the word used by 0.1 in case of success and decrease it by 0.2 on failure. Words with higher scores are preferred by the agents and words with a score of 0 are removed from the lexicon, which leads to a conventionalization of names in the population because words that are successfully used by many agents will ‘win’ over other words with the same meaning.

Since it would be impractical to do hundreds of language games with real robots and in order to be able to do repeatable and controlled experiments, we pre-recorded data sets of visual experiences and feed one of them to the agents in each interaction. But in principle it is possible to do the language games on-line on the robots – it would just take very long (in the range of hours) before they reach convergence.

Agents first play series of 1500 language games about body postures (see Figure 6). The graph shows that indeed the agents can reach a consensus on how to name posture categories stemming from the categorization processes and communicate successfully after a period of invention and alignment. Depending on different categorizations (two or three bodily posture categories), alignment in the population is reached on different time scales. In the two posture case the population converges much faster than in the three posture case because in the latter the number of meanings is higher (3 instead of 2) and thus there is a longer phase of word invention. But more importantly the cramped nature of the ”sit” category (category 3 in Figure 5) leads to a significant increase in time to alignment. Points that lie on the border of the category might be conceptualized differently by the interacting robots, in turn leading to a different word used and leading to communicative failure, which eventually decreases the score of that word-category link in the lexicon of the agent. Nevertheless both in the $k = 2$, as well as in the $k = 3$ cases the population reaches agreement and well above

95% communicative success.

Then, after 1500 interactions, the agents are presented with sensory experiences of objects instead of robots. None of the agents has seen any of the objects before. As shown Figure 6 (last 3500 games), they continue to have the same communicative success as in the previous games and the size of their lexicons does not change, which indicates that the agents readily extend their previously learnt posture words to additional objects. In fact, there is no difference in performance of the agents, when confronted with objects or postures. This shows that the visual features used for categorization are sufficient for the extensional use on objects.

Conclusion

In this paper we have presented a concrete operationalization of image schema theory in a computational embodied model. Processes hypothesized by cognitive linguists about how speakers of Germanic languages extend the use of posture verbs to non-living objects have been implemented in humanoid robots. We showed how semantics for postures can emerge from recurrent and repeated interactions of the agents with their environment and how these semantics can be used in repeated interactions between intelligent agents as the basis for successful communication. One particularly interesting explanation offered by the model is the account for the high cross-linguistic variety and vagueness in the usage of “sit” (when compared to ”lie” and ”stand”, which also seems to be the reason for its broad semantic extension, see Lemmens, 2002a).

We see this work as a support for image schema theory and as an example of how cognitive modeling can be substantiated with formal methods and thus provide insights for theories of cognition.

Acknowledgments

This research has been carried out at the Sony Computer Science Laboratory in Paris and at the Neurorobotics Research Laboratory at the Computer Science Department of the Humboldt-Universität zu Berlin with partial support from the ALEAR project, funded by the EU Cognitive Systems program. We would like to thank Christian Thiele for the help with the robotic setup.

References

Amant, R., Morrison, C., Chang, Y., Cohen, P., & Beal, C. (2006). An image schema language. In *7th international conference on cognitive modelling (ICCM 2006)* (pp. 292–297).

Borneto, S. (1996). Liegen and stehen in German: A study in horizontality and verticality. *Cognitive linguistics in the redwoods: The expansion of a new paradigm in linguistics*, 459–506.

Croft, W., & Cruse, D. (2004). *Cognitive Linguistics*. Cambridge University Press.

Gibbs, R. W., Beitel, D., Harrington, M., & Sanders, P. (1994). Taking a stand on the meanings of stand: Bodily experience as motivation for polysemy. *Journal of Semantics*(11), 231–251.

Hu, M. (1962). Visual Pattern Recognition by Moment Invariants. *Information Theory, IEEE Transactions on*, *8*(2), 179–187.

Johnson, M. (1987). *The body in the mind: the bodily basis of meaning, imagination and reason.* Chicago: University of Chicago Press.

Lemmens, M. (2002a). The semantic network of dutch posture verbs. In J. Newman (Ed.), *The linguistics of sitting, standing and lying.* Amsterdam/Philadelphia: John Benjamins.

Lemmens, M. (2002b). Tracing referent location in oral picture descriptions. In A. W. et al. (Ed.), *A rainbow of corpora-corpus linguistics and the languages of the world.* Lincom-Europa.

Lemmens, M. (2004). *Metaphor, image schema and grammaticalisation: a cognitive lexical-semantic study.*

Lloyd, S. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, *28*(2), 129–137.

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

Newman, J. (Ed.). (2002). *The linguistics of sitting, standing and lying*. Amsterdam and Philadelphia: John Benjamins.

Oosten, J. V. (1984). Sitting, standing and lying in dutch: A cognitive approach to the distribution of the verbs zitten, staand and liggen. In *Dutch linguistics at berkeley.* CA: UCB.

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

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent Systems*, *sept-oct 2001*, 17-22.

Steels, L., & Loetzsch, M. (2009). Perspective alignment in spatial language. In K. R. Coventry, T. Tenbrink, & J. A. Bateman (Eds.), *Spatial language and dialogue.* Oxford University Press. (to appear)

Steels, L., & Spranger, M. (2008). The robot in the mirror. *Connection Science*, *20*(4).

van Trijp, R. (2008). The emergence of semantic roles in fluid construction grammar. In A. D.M. Smith, K. Smith, & R. Ferrer i Cancho (Eds.), *Proceedings of the 7th international conference on the evolution of language (evolang 7)* (pp. 346–353). Singapore: World Scientific Publishing.

Wellens, P., Loetzsch, M., & Steels, L. (2008, June). Flexible word meaning in embodied agents. *Connection Science*, *20*(2 & 3), 173–191.