

# TOWARDS A SPATIAL LANGUAGE FOR MOBILE ROBOTS

RUTH SCHULZ, PAUL STOCKWELL, MARK WAKABAYASHI, JANET WILES

*School of Information Technology and Electrical Engineering,  
The University of Queensland,  
Brisbane, QLD, 4072, Australia*

We present a framework and first set of simulations for evolving a language for communicating about space. The framework comprises two components: (1) An established mobile robot platform, RatSLAM, which has a "brain" architecture based on rodent hippocampus with the ability to integrate visual and odometric cues to create internal maps of its environment. (2) A language learning system based on a neural network architecture that has been designed and implemented with the ability to evolve generalizable languages which can be learned by naive learners. A study using visual scenes and internal maps streamed from the simulated world of the robots to evolve languages is presented. This study investigated the structure of the evolved languages showing that with these inputs, expressive languages can effectively categorize the world. Ongoing studies are extending these investigations to evolve languages that use the full power of the robots representations in populations of agents.

## 1. Introduction

While all human languages can describe spatial representations, people speaking different languages will use different frames of reference: intrinsic (from the point of view of the object), relative (from the point of view of the speaker or some other viewer) or absolute (e.g. North, South, East and West) (Levinson, 1996). These frames of reference can be used to construct or describe spatial relationships in the world. The use of different frames of reference in different languages indicates that language may restructure the spatial representations of the language speaker, rather than the existence of innate and universal spatial concepts (Majid, Bowerman, Kita, Haun, & Levinson, 2004).

Computational modeling of language evolution provides a means of investigating ontology, grounding, learnability, and generalization in languages that evolve in populations of agents (See Steels, 2005 for an outline of the major stages in the evolution of language using computational models). The use of simulation techniques can add to the debate on the origins and evolution of language by determining factors that are important for evolving communication systems. Language games are a possible framework for language models in which agents engage in tasks requiring communication. These games have been used to evolve lexicons (Hutchins & Hazlehurst, 1995), categories (Cangelosi & Harnad, 2001), and grammars (Batali, 2002) in populations of agents.

The symbol grounding problem (Harnad, 1990) is a major issue for computational models of language. Without the grounding of meanings in the world, symbols refer only to other symbols with no association between the symbols and the world. One way to address the symbol grounding problem in computational models of language is to conduct language research with real or simulated robots (Marocco, Cangelosi, & Nolfi, 2003; Roy, 2001; Steels, 1999; Vogt, 2000).

In robot language research, the environments are often simplified and idealized compared to the real world. In the Talking Heads Experiment (Steels, 1999) geometric shapes were used rather than ‘real world’ objects such as tables and chairs. The languages evolved in the Talking Heads Experiment used a relative frame of reference to talk about the different shapes in the scene using meanings such as ‘left’ and ‘right’.

One way to extend robot language research is to use mobile robots that interact with a real world environment, using navigation systems to build up internal maps of the world. The use of mobile autonomous agents that move in a real environment enables the evolution of spatial languages using both relative and absolute frames of reference. The visual input of the robot would be used in a relative frame of reference, where the scenes can be categorized with respect to what the world looks like from the perspective of the robot. The internal maps would be used in an absolute frame of reference. The languages evolved could provide a methodology to investigate the structure of languages that describe space.

This paper introduces RatChat, a project that uses RatSLAM, an established mobile robot platform, to develop a framework for the robots to evolve a language describing their environment. The RatChat and RatSLAM projects are described in Section 2. A study using this platform to evolve spatial languages is presented in Section 3, followed by a general discussion and conclusion.

## **2. RatChat**

Simultaneous Localisation and Mapping (SLAM) is a methodology for robot map building and navigation. RatSLAM is a model of SLAM, based on the hippocampal complex in rodents, that uses a combination of the properties of grid based, topological, and landmark representations to keep a sense of space while adding robustness and adaptability (Milford, Wyeth, & Prasser, 2004).

The inputs to the RatSLAM system include odometry and vision with the resulting map represented by pose cells. Active pose cells represent the current location and orientation of the robot, and are arranged in  $(x,y,\theta)$  for ease of

visualization. With RatSLAM, robots use the appearance of an image to aid localization by learning to associate the appearance of a scene and its position estimate (Prasser, Wyeth, & Milford, 2004).

RatChat aims to evolve a shared lexicon between robots grounded in perceptions, local views, and behaviors using a language game framework (see Figure 1). The evolution of languages for locations will be explored, later extending the vocabulary of the robots to include objects. The challenge is for the robots to categorize their internal representations and label these with appropriate generalization and variability. The shared lexicon should allow the robots to agree on words for categories while including sufficient diversity for different categories to have different labels. As the language is expanded to include objects, more emphasis will be on the visual inputs of the robots (see Figure 2).

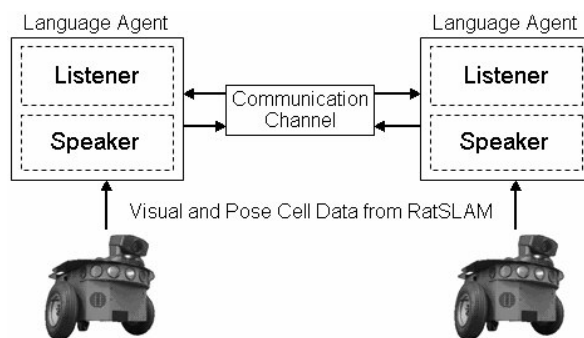


Figure 1 The framework for a language game. Each language agent obtains visual and pose cell data from the RatSLAM system. A communication channel is set up between the agents, allowing the speaker for each agent to produce utterances, and the listener for each agent to receive utterances for comprehension.



Figure 2 The robot's world comprises halls and open plan offices. A simulated world has been built to mirror the real world. The features of the environment shown in the visual images seen by the robot include the floor, walls, desks, chairs, and filing cabinets. The left image is from the robot's camera and the right image is the same location in the simulated world.

The RatChat language agents consist of a speaker and a listener based on simple recurrent neural networks (Elman, 1990; Tonkes, Blair, & Wiles, 2000). Speaker networks are extended to include the output of the network in the context for the next time step. Preliminary simulations showed that languages are easier to learn when the meaning space patterns are non-orthogonal and that distributed representations in signal space enable expressive languages to be found more easily than if localist representations are used.

### **3. A Spatial Language**

This study investigated the evolution of spatial languages using the visual and pose cell representations of the robot, looking at the expressivity of the languages evolved, and how the languages categorized the world of the robot.

*Methods:* The visual input for this study was every 100<sup>th</sup> scene in a series of 10000 visual scenes of 12x8 gray scale arrays obtained from a run of the robot in the simulated world. The pose cell input for the study was every 100<sup>th</sup> pattern in a series of 10000 pose cell patterns from the same run. The number of cells was reduced from 440640 to 610 by reducing the resolution of the pose cells (4x4x4 pose cells to 1 pose cell), and by discarding cells that are inactive in every pattern. For a third representation, the pose cells were processed using a hybrid system based on Self Organizing Maps (SOMs) (Kohonen, 1995). In the processing system, a SOM was trained on the input series for 1000 epochs. The output of the SOM was a 12x8 set of competitive units organized in a hexagonal pattern. To construct a distributed activation the actual output values of the units were converted to values between 0 and 1.

For the signal representation, utterances consisted of a sequence of three syllables. Each syllable was represented by a ten unit binary vector in which the two most active units were set to one, with all other units set to zero.

One way to measure understanding is to test how well an agent has categorized the world. The representations of the world are presented to the speaker, resulting in words associated with each pattern. Listeners produce a prototype for each unique utterance. If the original input pattern presented to the speaker is closest to the prototype for the utterance used by the speaker, this pattern has been correctly categorized. When many of the patterns are associated with one word, the agents will categorize more patterns correctly, but the language does not divide the meaning space effectively. A more appropriate measure of understanding is the number of patterns correctly categorized divided

by the largest category size, indicating how well the language divides up the meaning space, and how well the agent understands the language.

In this study, ten agents were evolved individually for 100 generations to produce languages based on each set of inputs (vision, pose cells and processed pose cells). A simple (1+1)-evolutionary strategy (Beyer & Schwefel, 2002) was used to evolve the agent’s speaker, introducing variability in the language. At each step, the agent’s speaker was evolved and the agent’s listener was trained on the language from the speaker for 500 epochs using the Back Propagation Through Time algorithm (Rumelhart, Widrow, & Lehr, 1994). The agents were evaluated with a fitness function based on the measure of understanding described above. If the listener trained on the mutant languages were better at categorizing the input patterns than the listener trained on the current champion language, then the mutant became the champion. The languages produced by the agents for each set of inputs were compared for expressiveness, categorization and how the meaning space was divided.

*Results:* The agents evolved with visual scenes as inputs produced languages with an average of 24.2 words (see Table 1). The average number of scenes correctly categorized by the agents was 53.4 out of 100. One highly expressive language had 67 unique words of which 47 were associated with single scenes. Words often appeared to group several different types of images together, with the resulting prototype visual scene for the word a combination of these scenes. One set of similar scenes were those in which the robot faced a white wall with a strip of black next to the floor. All of the languages other than the most expressive language grouped together some of these scenes (see Figure 3).

Table 1 Properties of the languages evolved with different sets of input

	<i>Number of Unique Words (avg (std))</i>	<i>Number of Patterns Correctly Categorized (avg (std))</i>
Vision	24.2 (17.3)	53.4 (13.5)
Pose Cells	23.2 (12.4)	22.6 (10.4)
Processed Pose Cells	10.9 (6.4)	58.7 (10.4)

The agents evolved with pose cells as inputs produced languages with an average of 23.2 words. The average number of scenes correctly categorized by the agents was 22.6 out of 100. The majority of the words were associated with single input patterns or a small number of input patterns, scattered across the space. Some words group together input patterns that are close together in space, but these words are also generally associated with a small number of input patterns from other areas.

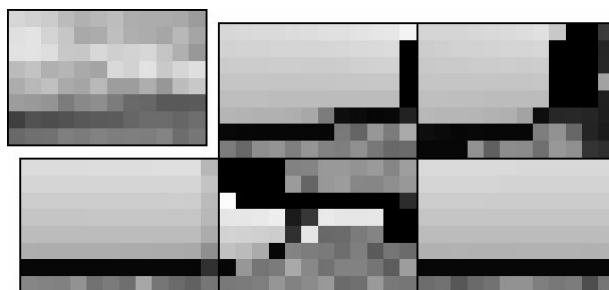


Figure 3 The prototype for the word 'kufufu' (top left) and the five scenes that are associated with this word in a language with 27 unique words. Most of the scenes associated with 'kufufu' show a white wall with a black strip, although the bottom middle scene has different features.

The agents evolved with processed pose cells as inputs produced languages with an average of 10.9 words. The average number of scenes correctly categorized by the agents was 58.7 out of 100. These languages had less words associated with single input patterns and more words associated with many input patterns spread across the entire space. The larger languages had more words associated with groups of input patterns that were close together in space.

*Discussion:* Expressivity is an important feature of language, where unique words are used for unique meanings. In this simulation, expressivity is indicated by the number of unique words. The vision and pose cell representations resulted in languages with an average of over 20 unique words for the 100 input patterns, while the processed pose cell representation resulted in languages with an average of 10.9 unique words. This reduction in expressivity for the processed pose cell representation indicates that the unique information in the input patterns may be lost when the pose cell representation is processed.

The number of categories correct indicates how well the language categorizes the world. The processed pose cell languages were most successful at clustering input patterns that were close together in space, with distinct clusters associated with single words. The unprocessed pose cell languages were not as successful at categorizing the patterns, which may be due to the size and sparseness of the pose cell representation, and can be addressed by processing the pose cell representation.

Some of the agents using languages evolved with vision were successful in grouping together similar scenes, however many of the words in the vision languages grouped together images that were dissimilar as well as similar, or were associated with single images. In this study, raw vision as an input provided a structure that allowed some languages to evolve to successfully categorize the

world. Processing the scenes prior to the language agent may extract the important information from each scene that is necessary for languages to consistently evolve with expressivity and categorization.

#### **4. General discussion and conclusion**

The RatChat project aims to explore the structure of languages that describe space using mobile robots. The simulations presented in this paper represent agents developing their internal representations of the world prior to playing naming games in populations of agents, and have provided insight into the expressivity, categorization, and structure of languages that can evolve from visual and pose cell representations.

There is a tradeoff between expressivity, with unique words for unique meanings, and categorization, with the use of one word for a group of similar meanings. The degree of expressivity and categorization can be altered by processing the inputs, as can be seen with the pose cell representation: the unprocessed languages are more expressive, while the processed languages are better at categorizing the world.

We are currently running simulations to scale up these results with further studies into processing the robot representations prior to the language networks and evolving languages in populations of agents.

#### **Acknowledgements**

We thank members of the RatSLAM team Michael Milford, David Prasser, Shervin Emami, and Gordon Wyeth. This research is funded in part by a grant from the Australian Research Council.

#### **References**

- Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In E. J. Briscoe (Ed.), *Linguistic Evolution Through Language Acquisition: Formal and Computational Models*. Cambridge, UK: Cambridge University Press.
- Beyer, H.-G., & Schwefel, H.-P. (2002). Evolution Strategies: A comprehensive introduction. *Natural Computing*, 1, 3-52.
- Cangelosi, A., & Harnad, S. (2001). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, 4(1), 117-142.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14, 179-211.

- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42, 335-346.
- Hutchins, E., & Hazlehurst, B. (1995). How to invent a lexicon: The development of shared symbols in interaction. In N. Gilbert & R. Conte (Eds.), *Artificial Societies: The Computer Simulation of Social Life*. London: UCL Press.
- Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.
- Levinson, S. C. (1996). Language and Space. *Annual Review of Anthropology*, 25, 353-382.
- Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Science*, 8(3), 108-114.
- Marocco, D., Cangelosi, A., & Nolfi, S. (2003). The role of social and cognitive factors in the emergence of communication: experiments in evolutionary robotics. *Philosophical Transactions of the Royal Society London - A*, 361, 2397-2421.
- Milford, M. J., Wyeth, G. F., & Prasser, D. (2004). RatSLAM: a hippocampal model for simultaneous localization and mapping. In *IEEE International Conference on Robotics and Automation (ICRA 2004)*: IEEE Press.
- Prasser, D., Wyeth, G. F., & Milford, M. J. (2004). *Biologically inspired visual landmark processing for simultaneous localization and mapping*. Paper presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai.
- Roy, D. (2001). Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication*, 4(1), 33-56.
- Rumelhart, D. E., Widrow, B., & Lehr, M. A. (1994). The basic ideas in neural networks. *Communications of the ACM*, 37(3), 87-92.
- Steels, L. (1999). *The Talking Heads Experiment* (Vol. I. Words and Meanings). Brussels: Best of Publishing.
- Steels, L. (2005). The emergence and evolution of linguistic structure: from lexical to grammatical communication systems. *Connection Science*, 17(3-4), 213-230.
- Tonkes, B., Blair, A., & Wiles, J. (2000). Evolving learnable languages. In S. A. Solla, T. K. Leen & K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems 12*. Boston: MIT Press.
- Vogt, P. (2000). Bootstrapping grounded symbols by minimal autonomous robots. *Evolution of Communication*, 4(1), 87-116.