Generalization in Languages Evolved for Mobile Robots

Ruth Schulz, Paul Stockwell, Mark Wakabayashi and Janet Wiles

School of Information Technology and Electrical Engineering,
The University of Queensland, Brisbane, QLD, 4072, Australia
{ruth,wiles}@itee.uq.edu.au

Abstract

A set of simulations are presented that investigate generalization in languages evolved for mobile robots. The mobile robot platform is RatSLAM, a model for Simultaneous Localization and Mapping based on rodent hippocampus that uses visual and odometric information to build up a map of the explored environment. The language agents use information from this system as inputs and are based on simple recurrent neural networks. This paper describes two sets of experiments exploring the nature of generalization in evolved languages. The first study investigated languages evolved from visual inputs and the second study investigated languages evolved from position representations. These studies showed that processing the input prior to the language agent affects the expressivity of the languages and the performance of the agents. Some generalization occurs in these languages. Studies are ongoing to extend these simulations using the simulated world of the robots.

Introduction

Human languages are able to generalize from one situation to another with an infinite number of meanings described using a finite number of words. The most basic type of information exchange, signaling, does not have this capability. All possible terms and structures in signaling must be predefined and known by both listener and speaker (Kirby, 2002; Oliphant & Batali, 1997). Simulations on the evolution of language suggest that languages can adapt to become more learnable and to enable generalization (Elman et al., 1996; Tonkes et al., 2000). In fact, the recursive patterns that evolve in languages providing the ability to generalize also result in languages that are easier to learn (Batali, 2002).

When humans first start to learn language, they are only exposed to a limited subset of the language, an effect known as the bottleneck of linguistic transmission (Kirby, 2000; Tonkes & Wiles, 2002). Even with minimal feedback, they almost universally are able to master language by an early age (Brown & Hanlon, 1970).

To be able to learn and produce a potentially infinite number of expressions from a finite set of examples, the structure of a language must be regular and predictable. If the learners must generalize, then the language itself must be able to generalize (Tonkes, 2001).

Simulations have investigated this bottleneck of linguistic transmission by forcing the language through the bottleneck of the agent's limited exposure (Tonkes, 2001). In these simulations, the regular structure of the language facilitates generalizing to the rest of the language without the need to be exposed to all possible combinations.

A feature of learnable simulated language is its stability through multiple generations; even though languages adapt, they still retain regularity in their structure (Kirby, 2001). Artificial language needs to be generalizable to become more learnable and able to adapt over time.

Experiments have studied the evolution of language when grounded in the environment using automated agents, however generalization was not the main focus (Marocco et al., 2003). Other research has been specifically interested in generalization but was not concerned with the effect of environmental influences (Tonkes, 2001).

In this paper, we present studies that investigate how different input representations affect the ability of language agents to generalize. Two types of input have been tested: vision and position representations. The purpose of the study was to investigate the nature of generalization and the relation between types of input, pre-processing and generalization.

The next section describes the RatChat project that these studies are part of. This is followed by the studies using vision and position representations, and a general discussion and conclusion. Further studies for the RatChat project are also presented.

RatChat

This research is part of the RatChat project (Schulz et al., in press) that builds on RatSLAM, a model of Simultaneous Localization And Mapping (SLAM) for mobile robots. SLAM is a methodology for robot map building and navigation.

The RatSLAM system is a model of SLAM based on rodent hippocampus that integrates information from external vision and internal odometry to update activity in pose cells (Milford et al., 2004). Pose cell activity represents the position and orientation of the robot. Robots using RatSLAM (see Figure 1) can use this information to navigate to places that have been visited.

RatChat uses the pose cell representation and visual input as complex representations suitable for investigating the evolution of language in mobile robots. These representations are obtained from real or simulated robots exploring an environment (see Figure 2), and are used by language agents to evolve languages.

Figure 1 The Pioneer robots used in the RatSLAM and RatChat projects have cameras, laser sensors, sonar sensors, and odometer sensors with which to explore the world. In simulations with RatSLAM, the robots have a wandering behavior to build up a map of the world.

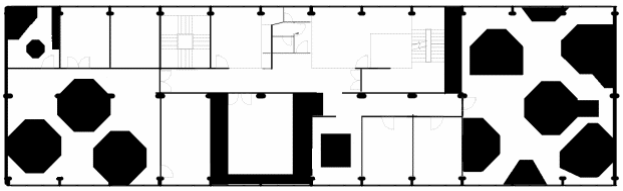


Figure 2 The robot's world comprises halls and open plan offices. A simulation world has been built to mirror the real world, with images from the real world used in constructing the views of the robot.

The languages that evolve are spatial languages, with the agents communicating about location, represented by the pose cells, and the appearance of the world, shown by the visual input. The pose cell representations refer to an absolute frame of reference that uses the internal map that is built up to refer to specific places in the world, while the visual representations, or scenes, refer to a relative frame of reference in which the robot talks about the world based on the view of the robot.

The language agents used in these simulations comprises speaker and listener agents based on simple recurrent neural networks (Elman, 1990; Tonkes, 2001). The inputs for these agents are obtained from the RatSLAM system. In this study, the language agents are simulated offline.

In previous studies (Schulz et al., in press), we found that the input and output representations for the language agents make a large difference for whether expressive languages are easy for the networks to evolve and to learn. In particular, languages are easier to learn when the input patterns are non-orthogonal. Also, agents are able to trade expressiveness and categorization, with processing of the inputs altering how expressive the languages were, and how well they could categorize inputs.

Expressive languages that categorized the world using unique words to group together input patterns were evolved using vision and pose cell representations. The studies presented in this paper aim to test whether languages can be evolved that are able to generalize.

Study 1 Vision

The different forms of input available from the RatSLAM system are vision and pose cells. In this study, vision is used. The nature of this input is such that pre-processing may be preferable to the raw visual input. This first study aimed to test whether generalization occurs in languages evolved using visual input. The generalization that may occur is the use of novel words for novel meanings, and also the ability to use these words in a consistent way that allows the world to be categorized effectively.

Methods

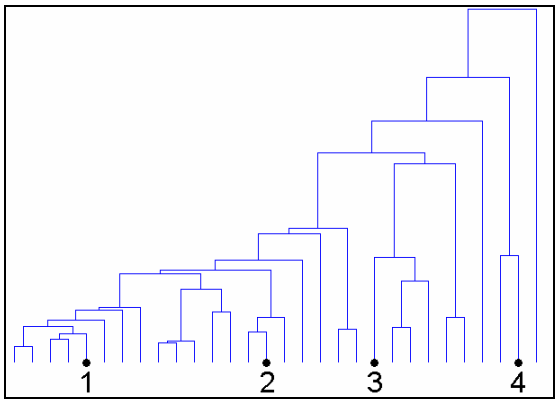
The input for this study was a series of 10000 visual scenes of 12x8 gray scale viewed by the robot exploring the simulation world. Every tenth scene of this series was chosen, forming a series of 1000 scenes for input to the language system. This series was analyzed using hierarchical clustering to determine 30 clusters of similar images (see Figure 3a). The image closest to the mean for each of the 30 clusters was chosen for evolving and training the language networks. A visual inspection of the images showed that they were dissimilar scenes (see Figure 3b) and spread throughout the robot's world (see Figure 3c).

Three techniques were used for processing the visual information for the language agents. The first technique was using the raw image. The second technique involved categorizing the input with a self organizing map (SOM) (Kohonen, 1995). In this hybrid system, a SOM was trained on the visual scenes for 1000 epochs. The output of the SOM was an array of competitive units organized in a hexagonal pattern. To give a distributed activation pattern for the language agents, the actual values of the units were scaled to values between 0 and 1. The third technique used Principal Component Analysis (PCA). The 1000 scenes were analyzed for their principal components, and the component scores were scaled to values between 0 and 1.

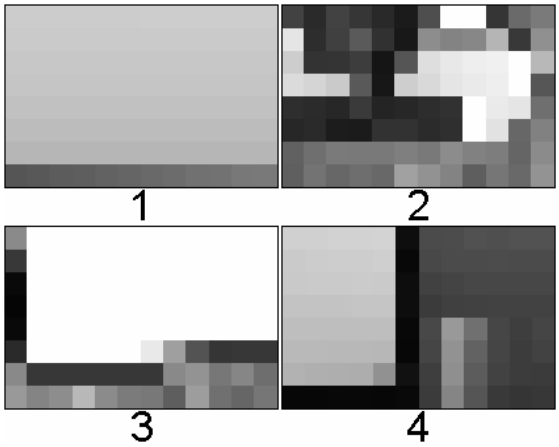
Language agents were evolved for 1000 generations with inputs of different sizes of processed images to find the smallest size for which expressive languages could evolve. For the raw image, an input scene of 12x8 pixels was used, for the SOM-based input, a SOM of size 24x16 was used, and for the PCA-based input, the first 48 components were used.

One way of testing whether a language captures the underlying structure of a set of visual scenes is to test how well the concepts are mapped to the language terms. Listeners produce a prototype for each unique word. If the original scene presented to the speakers is closest to the prototype for the word used by the speaker, the scene has been correctly categorized. The measure of similarity between images and prototypes used in this analysis was sum squared error.

a)



b)



c)

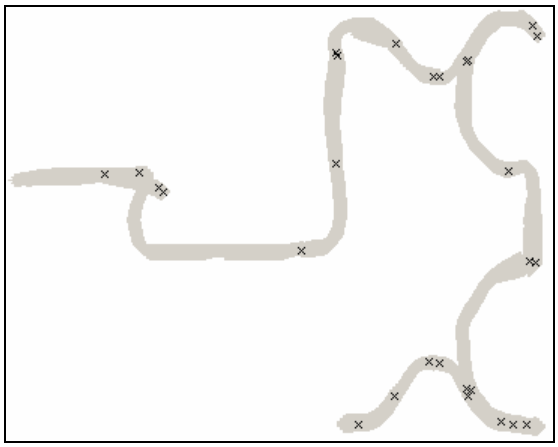


Figure 3 a) An analysis of the meaning space of 1000 scenes using hierarchical clustering. The diagram shows the 30 clusters used to choose 30 dissimilar images for evolving the language agents. b) The 12x8 images closest to the mean for the four labeled clusters. These were four of the images used to evolve the language agents. c) The position of the robot for each of the 30 images, shown over the path of the robot. Most of the images are located at times when the robot was turning.

For each pre-processing technique, ten agents were evolved for 500 generations with a selection strategy based on how well the agents categorized the world. The winner of the current champion and mutant language was the one in which the trained listener and speaker networks were able to associate the highest number of images with words correctly. The evolved languages were analyzed for generalization by testing how many words were uttered and the performance of the agents for the set of 1000 scenes.

Results

The agents with the raw image input evolved languages with an average of 22.4 unique words. With the SOM-based input they averaged 17.2 unique words, and with the PCA input they averaged 22.8 unique words (see Table 1).

Table 1 Language Sizes for Pre-Processing Techniques

	Image unique words average (std)	SOM-based unique words average (std)	PCA unique words average (std)
30 images	22.4 (8.3)	17.2 (5.3)	22.8 (3.3)
1000 images	99.9 (65.2)	43.5 (17.5)	111.2 (46.8)

A feature of generalization is the ability to produce new utterances for novel meanings. In this study, the ability to produce new utterances can be measured by comparing the number of words used for the original 30 scenes with the number used for the larger set of 1000 scenes.

When the speakers were presented with 1000 images, the raw vision speakers produced an average of 99.9 unique utterances, the SOM-based speakers averaged 43.5 unique utterances, and the PCA speakers averaged 111.2 unique utterances. These results are remarkable in the number of new words for novel images, with more than twice the number of words than for the initial 30 images.

To measure the performance of the agents, the number of visual scenes close to the prototypes used for the scenes was found. The distance between the visual scene and the prototype used was determined by treating them as vectors and calculating one minus the cosine of the included angle between them. This distance was then normalized by the standard deviation of the distances between each of the 30 scenes. The number of scenes within 0.25, 0.5, and 1.0 standard deviations of the prototype were calculated for each of the techniques for the 30 and 1000 scenes.

For the 30 scenes used for evolving the agents, the image agents averaged 14.8 scenes within 0.25 standard deviations of the prototype used, 21.9 within 0.5 standard deviations, and 28.0 within one standard deviation; the SOM-based agents averaged 21.7 within 0.25 standard deviations, 27.2 within 0.5 standard deviations, and 29.8 within one standard deviation; and the PCA based agents averaged 8.0 within 0.25 standard deviations, 11.4 within 0.5 standard deviations, and 17.7 within one standard deviation (see Table 2).

Table 2 Images Similar to Prototypes for Pre-Processing Techniques (30 Images)

Standard Deviations	Image images close to prototype average (std)	SOM-based images close to prototype average (std)	PCA images close to prototype average (std)
0.25	14.8 (9.3)	21.7 (6.4)	8.0 (7.3)
0.5	21.9 (8.0)	27.2 (6.1)	11.4 (7.9)
1.0	28.0 (4.7)	29.8 (5.3)	17.7 (9.8)

When the agents were presented with 1000 images, the image agents averaged 26.1 scenes within 0.25 standard deviations of the prototypes used, 81.2 within 0.5 standard deviations, and 399.0 within one standard deviation. The SOM-based agents averaged 334.9 within 0.25 standard deviations, 689.6 within 0.5 standard deviations, and 920.9 within one standard deviation. The PCA-based agents averaged 8.3 within 0.25 standard deviations, 13.2 within 0.5 standard deviations, and 29.2 within one standard deviation (see Table 3).

Table 3 Images Similar to Prototypes for Pre-Processing Techniques (1000 Images)

Standard Deviations	Image images close to prototype average (std)	SOM-based images close to prototype average (std)	PCA images close to prototype average (std)
0.25	26.1 (15.8)	334.9 (123.5)	8.3 (8.5)
0.5	81.2 (38.8)	689.6 (194.3)	13.2 (11.1)
1.0	399.0 (111.0)	920.9 (205.0)	29.2 (19.7)

Discussion

Agents can trade expressiveness and categorization or generalization, so for a comparison of generalization between different techniques, languages must be of a similar size. The expressiveness of the image, SOM-based and PCA agents is similar, with an average of 22.4, 17.2 and 22.8 unique words for the 30 images.

One feature of generalization is the ability of agents to utter new words when presented with new meanings. In this study, the agents were able to produce between 2.5 (agents with SOM-based inputs) and 4.9 (agents with PCA inputs) times the number of words for 30 images when presented with the larger set of 1000 images.

The performance of the agent is the next feature to consider. For this, the similarity of an image to the prototype for the word associated with the image was used. For the 30 images, the image and SOM-based agents performed well, with almost all images within one standard deviation of the prototypes. The PCA agents had an average of just fewer than 18 of the 30 images within one standard deviation of the prototype.

When generalizing to the 1000 images, the SOM-based agents performed better than both other techniques, with almost all images within one standard deviation. The image agents performed moderately well, with just under 400 within one standard deviation, while the PCA had only 29.2 within one standard deviation.

These generalization results are skewed by the input representations. The SOM-based representations have many patterns within one standard deviation of the average SOM-based pattern; the image representations have more patterns between one and two standard deviations from the average image; and PCA based representations have more patterns between 1.5 and 2.5 standard deviations from the average PCA pattern.

When the majority of patterns are less similar to the average pattern, languages that effectively cover the space are more difficult to find. This is in part due to the difficulty in constructing appropriate prototypes for a group of diverse input patterns

Study 2 Pose Cells

The second study aimed to test whether generalization occurs in languages evolved using pose cells, again looking at whether new words were used for new input patterns, and at the performance of the agents.

Methods

The input for this study was a series of 10000 pose cell representations, obtained from the run of the robot used in the previous study. Again, every tenth pattern was included for input to the language agents.

The number of pose cells was reduced from 440640 to 947 by reducing the resolution of the pose cells (180x68x36 cells to 45x17x9 cells) and by discarding cells that were inactive for the entire run (6885 cells to 947 cells).

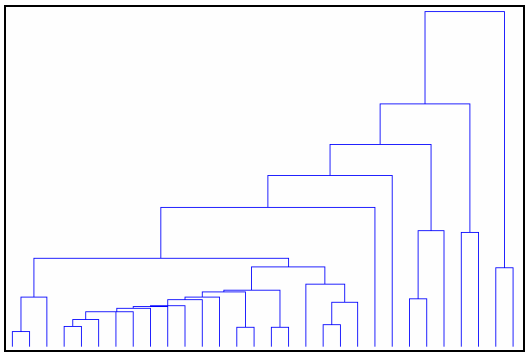
These inputs were analyzed using hierarchical clustering to find 30 pose cell representations for presenting to the language agents (see Figure 4a). The position of the robot for each of the 30 pose cell patterns is shown in Figure 4b.

Again, three techniques were used for processing the input. The raw pose cell representation, a SOM-based technique, and PCA were used.

To determine appropriate sizes for the inputs, language agents were evolved for 1000 generations with variability as the fitness function to find the smallest size that allowed expressive languages to evolve. For the pose cells, 947 input units were used, for the SOM-based input, a SOM of size 12x8 was used, and for the PCA-based input, the first 120 components were used.

For each pre-processing technique, ten language agents were evolved for 500 generations with a selection strategy based on how well the language categorizes the world. The languages were then analyzed for generalization by testing how many words were uttered and the performance of the agents for the larger set of 1000 patterns.

a)



b)

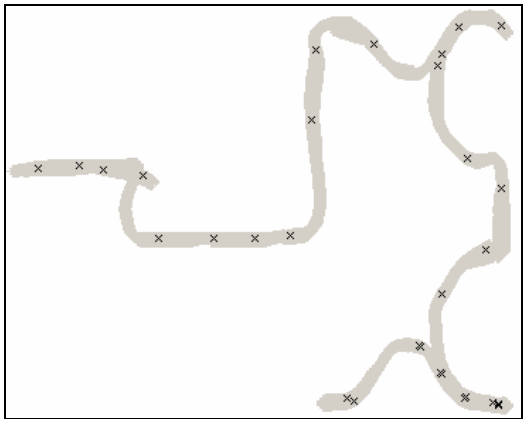


Figure 4 a) An analysis of the meaning space of 1000 pose cell patterns using hierarchical clustering. The diagram shows the 30 clusters used to choose 30 pose cell patterns spread across the space for evolving the language agents. b) The 30 positions chosen for the pose cell patterns, shown over the path of the robot. In this figure, some of the crosses are very close to each other. These are pose cell patterns that have a similar position, but different orientation.

Results

The speakers presented with the raw pose cell representation evolved languages with an average of 18.9 unique words. With the SOM-based input they averaged 18.5 unique words and with the PCA input they averaged 21.7 unique words (see Table 4).

Table 4 Language Sizes Generated for Pre-Processing Techniques

	Pose Cell words spoken average (std)	SOM-based words spoken average (std)	PCA words spoken average (std)
30 patterns	18.9 (7.7)	18.5 (6.1)	21.7 (5.3)
1000 patterns	112.6 (86.7)	34.4 (19.2)	93.1 (46.5)

When the speakers were presented with 1000 patterns, the raw vision speakers produced an average of 112.6 unique utterances, the SOM-based speakers averaged 34.4 unique utterances, and the PCA speakers averaged 93.1 unique words.

For the 30 patterns used for evolving the agents, the agents with pose cell inputs had no input patterns with in one standard deviation of the prototype used for that pattern. The SOM-based inputs averaged 18.1 patterns within 0.25 standard deviations of the prototype, 23.2 within 0.5 standard deviations and 28.3 within one standard deviation; and the PCA based inputs averaged 7.7 patterns within 0.25 standard deviations, 8.3 within 0.5 standard deviations and 14.5 within one standard deviation (see Table 5).

Table 5 Patterns Similar to Prototypes for Pre-Processing Techniques (30 Patterns)

Standard Deviations	Pose Cell patterns close to prototype average (std)	SOM-based patterns close to prototype average (std)	PCA patterns close to prototype average (std)
0.25	0 (0)	18.1 (3.5)	7.7 (7.1)
0.5	0 (0)	23.2 (2.7)	8.3 (6.8)
1.0	0 (0)	28.3 (1.3)	14.5 (8.0)

When the agents were presented with 1000 images, the pose cell agents again had no input patterns within one standard deviation of the prototype. The SOM-based agents averaged 558.4 within 0.25 standard deviations, 767.0 within 0.5 standard deviations, and 915.5 within one standard deviation; and the PCA-based agents averaged 16.6 within 0.25 standard deviations, 42.8 within 0.5 standard deviations, and 131.6 within one standard deviation (see Table 6).

Table 6 Patterns Similar to Prototypes for Pre-Processing Techniques (1000 Patterns)

Standard Deviations	Pose Cell patterns close to prototype average (std)	SOM-based patterns close to prototype average (std)	PCA patterns close to prototype average (std)
0.25	0 (0)	558.4 (40.6)	16.6 (38.5)
0.5	0 (0)	767.0 (36.1)	42.8 (59.9)
1.0	0 (0)	915.5 (41.4)	131.6 (74.6)

Discussion

The expressiveness of the pose cell, SOM-based and PCA agents is similar, with an average of 18.9, 18.5 and 21.7 unique words for each of the 30 patterns, meaning that comparisons can be made for generalization in these languages.

Considering the ability of agents to utter new words when presented with new meanings, in this study, the agents were able to produce between 1.8 (agents with SOM-based inputs) and 6.0 (agents with raw pose cell inputs) times the number of words for 30 patterns when presented with the larger set of 1000 patterns.

For the performance of the agents, the closeness of a pattern to the prototype for the word associated with the pattern was used. For the 30 patterns, the SOM-based agents performed well, with almost all patterns within one standard deviation of the prototypes. The PCA agents had an average of just under half of the patterns within one standard deviation of the prototype.

When generalizing to the 1000 patterns, the SOM-based agents performed better than the other techniques, with almost all patterns within one standard deviation. The PCA averaged 131.6 within one standard deviation.

The pose cell agents had no patterns within one standard deviation of the prototype for the word associated with the pattern. This lack of similarity is almost certainly due to the sparseness of the input patterns. In the pose cell representation, most of the inputs are not active, with an average of 11 out of 947 cells active. This representation means that the listeners do not learn to associate the words with the pose cell representations, and the prototypes have all cells set close to zero.

These results are skewed by the input representations in a similar way to the visual inputs, where the SOM-based representations have many patterns within one standard deviation of the average SOM-based pattern, and the PCA based representations have more patterns between 1.5 and 2.5 standard deviations from the average PCA pattern.

General Discussion and Conclusions

Generalization is an important question for simulations of the evolution of language. For generalizable languages, the first requirement is that the languages have the potential for expressing novelty. In these studies, the language agents produced novel words for novel scenes, which can be seen as constructing new "words" by recombining known "morphemes" in different ways. This expressiveness occurred in each type of processing for both vision and pose cell representations, with an average of between 1.8 (agents with pose cell SOM-based inputs) and 6.0 (agents with pose cell inputs) times the number of words produced for 1000 patterns compared to the original 30 patterns.

The next feature of generalization to consider is whether the language agents perform well in producing meaningful utterances that they are able to understand. The measure of meaningful generalization used in these studies was the number of images that were similar to the prototype for the word used for the image. This measure was altered by the features in the input representation, including the sparseness of the input patterns, and how the input patterns are grouped. If there are natural clusters of input patterns, such as in the SOM-based representations, it is easier for the agents to create prototypes that are closer to a cluster of input patterns. If, however, there are no natural clusters of input patterns, such as with the pose cell representations where most of the input patterns are dissimilar to each other, it is difficult for the agents to create effective prototypes.

The different pre-processing techniques and sizes of inputs to the language speakers affect the expressivity of the languages produced and the success of categorization of the scenes. Other techniques may be more successful at extracting the important information from the raw input for evolving expressive, comprehensible languages.

Further Studies

The studies presented here provide first steps towards an analysis of generalization. Further analysis is required to determine whether the agents are able to categorize similar scenes together using the evolved languages and how the structure of the input space affects the languages that are evolved and the ability of the agents to learn the scene-word associations.

In addition to the raw visual and pose cell inputs, a SOM-based, and a PCA processing technique were the types of pre-processing used in this study. Other forms of processing may enable the interesting information to be extracted from the robot representations while reducing the dimensionality of the input. An ideal pre-processing system would result in an input structure which provides sufficient information for expressive languages to evolve, while also allowing the networks to be small enough to provide real time language processing. Variations on the SOM and PCA based techniques are being investigated.

With a better understanding of an analysis of generalization and techniques more appropriate for processing the robots representations, future studies will involve language agents implemented online in the simulated world of the robots.

Acknowledgements

We thank members of the RatSLAM team Michael Milford, David Prasser, Shervin Emami, and Gordon Wyeth.

This research is funded 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.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes

(Ed.), *Cognition and the development of language*. New York: Wiley.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A connectionist perspective on development*. Boston: MIT Press.

Kirby, S. (2000). Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In C. Knight, J. R. Hurford & M. Studdert-Kennedy (Eds.), *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form* (pp. 303-323). Cambridge, UK: Cambridge University Press.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life, 8*(2), 185-215.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.

Marocco, D., Cangelosi, A., & Nolfi, S. (2003). The emergence of communication in evolutionary robots. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences, 361*(1811), 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.

Oliphant, M., & Batali, J. (1997). Learning and the emergence of coordinated communication. *The newsletter of the Center for Research in Language, 11*(1).

Schulz, R., Stockwell, P., Wakabayashi, M., & Wiles, J. (in press). Towards a spatial language for mobile robots. In A. Cangelosi, A. D. M. Smith & K. Smith (Eds.), *Proceedings of the 6th International Conference on the Evolution of Language*. Singapore: World Scientific.

Tonkes, B. (2001). *On the origins of linguistic structure: computational models of the evolution of language*. Unpublished PhD dissertation, School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane.

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.

Tonkes, B., & Wiles, J. (2002). Methodological issues in simulating the emergence of language. In A. Wray (Ed.), *The Transition to Language*. Oxford: Oxford University Press.