Symbol Interpretation in Neural Networks: an investigation on representations in communication

Emerson Oliveira (emersonso@dcc.ufba.br)
Angelo Loula (angelocl@ecomp.uefs.br)
Cognitive and Intelligent Systems Lab (LASIC)
State University of Feira de Santana (UEFS)
Feira de Santana, BA, Brazil

Abstract

Computer simulations have been used to study various aspects about the emergence of communication. But there have been only a few works on the underlying representation processes occurring during the interpretation by an agent of a representation produced by another agent. Here we present a study on representation processes in the emergence of communication occurring in a frequently used cognitive architecture in such experiments, artificial neural networks. We investigate the neural network's activations during the emergence of communication in search for representational and referential processes. Results show that it is possible to evaluate such processes along the evolution of communication and analyze interpretation accordingly.

Keywords: Representation; Communication; Neural Networks; Artificial Intelligence; Computer Simulation.

Introduction

Computer experiments involving the simulation of interactions between agents have been used to study various aspects of communication and language (for a review of works, see Nolfi and Mirolli, 2010, Christiansen and Kirby, 2003, Wagner et al. 2003). In these experiments, communication and linguistic processes are simulated in a social context, involving multiple agents. The process in focus is not pre-defined, but it rather emerges during and by means of agents' interactions. As the main form of interaction between agents, in most of these synthetic experiments, communication has, particularly, been a significant research subject. As communication involves the production and interpretation of representations, to understand the underlying representation processes is an important issue. Even so, in computational studies on the emergence of communication, little or nothing is discussed on the representational processes taking place, therefore it remains still a rather open research trend.

To study representation processes, it is necessary to examine the interpretation process occurring in the artificial agent and thus to inspect its cognitive dynamics. A frequently used cognitive architecture to control agents during the emergence of communication is neural networks. Here we propose to investigate the activation patterns of neural networks to evaluate representational processes during the emergence of communication. As a theoretical framework to define representation, its model, constituents and varieties, we apply theoretical principles from C.S.Peirce's pragmatic theory of signs.

We reproduce the experiment on the emergence of communication as proposed by Mirolli and Parisi (2008), in which the agents are controlled by a feed-forward neural network, receiving visual and auditory inputs and producing motor actions and auditory outputs. The main objective is to compare the middle layer's activations from visual input and from auditory input and verify if an auditory input can act as a representation of an object perceived by a visual input, and determine the type of representation occurring.

In the next section, we review related work on simulation of the emergence of communication using neural networks. Next, we briefly describe the theoretical principles from Peirce theory of signs. We then describe our computational experiment to study representation and interpretation processes in communication events. Finally, we outline our results and conclusions and point out perspectives on the study of representations in the emergence of sign processes.

Related Work

There have been several works on computational experiments related to the emergence of communication in a community of artificial agents (Nolfi and Mirolli, 2010, Christiansen and Kirby, 2003, Wagner et al. 2003). However, discussions on the underlying representation processes, particularly in those using neural networks, find little space in such literature. We will review a few representative works that deal with the emergence of communication among agents controlled by neural networks that are relevant in the context of this work.

Robots were evolved by de Greeff and Nolfi (2010) to execute a navigation task in which two robots had to exchange places in two target areas. The robots could use wireless sensors for an 'explicit signal' communication or they could use their spatial position as an 'implicit signal'. At the end of an evolution process of neural networks that control the robots, de Greeff and Nolfi (2010) described that the robots were able to use 2 or 3 explicit signals to execute the proposed task, but also used one implicit signal to achieve that. They stated that explicit signals codify certain conditions in which the emitter robot finds itself and that the implicit signal is a visual perception of the position of the other robot, and that each signal produces a different reaction. Signals are said to be deictic, dependent of spatial-

temporal context, but there was no further discussion on what and how robots representationally interpret such signals.

Mirolli and Parisi (2008) studied the problem of how communication can emerge without cooperation given that the production and response to signals are adaptively neutral, taken in isolation. In their simulation, two creatures stand in a corridor, with one standing in one end, in front of a mushroom (edible or poisonous), and the other one standing in the other end. The two creatures are controlled by a feed forward neural network and can perceive the mushroom and listen to signals, being able to move forward or stand still and to produce signals. Their results show that a reliable communication system does not stabilize since individuals are competing, thus the communication system alternates between a reliable one and a deceiving one. Along results discussions, the authors defined (internal) representations as activation patterns in the hidden units of the neural network and used them to evaluate categorization patterns in the neural network. Yet nothing is discussed about what the communicated signal would represent and how it could represent something else.

We have previously simulated the emergence of interpretation of different types of representations in communicative interactions (Loula et al., 2010), and have also studied further the cognitive conditions to the emergence of such interpretation processes (Loula et al, 2011). However, these previously done experiments used finite state machines as cognitive architectures and focused on the emergence of interpretation with fixed production of a single representation with only one referent. We have also evaluated representation processes in the emergence of both interpretation and production of multiple representations, with multiple referents, with agents using feed-forward neural networks as cognitive architecture (Loula et al., 2013). Agents were evolved for a resource collecting task, and the neural network architecture could contain a direct connection between auditory inputs and motor actions or auditory inputs could be associated with visual categories establishing an associate memory used for representation interpretation. The activation of the neural network was studied to evaluate representation processes and to categorize different types of representations. The neural network, however, used a localist activation with only one neuron activating at each time, with auditory inputs and visual inputs connected to different neurons that could be connected, in turn, forming an explicit associative memory.

Up to now, we have not found other works that have studied representations in the emergence of communication. Although not related to the emergence of communication, Mirolli (2012) conducted an analysis of representations in a recurrent neural network in a minimum cognition experiment by Beer (2003). The experiment consisted of a single agent with limited vision in a scenario where two different types of object would fall from the top. The agent task was to get close to one type of object and to move away from the other one. Mirolli (2012) reproduced the experiment investigating if, when and in what circumstances the agent uses representations. He defined representations as an internal state that can be correlated to an external feature, with this correlation being functional to the agent. To search for representations, he inspected the middle layer of the neural network, but in the original experiment, he did not identify representations. In an alternative experimental configuration, the task was changed and the agent would need some sort of internal memory to handle it. This time, he was able to identify the use of representations.

Peirce's theory of sign

In order to study representational processes through computational experiments, it is important to clearly describe theoretical principles supporting them. In order to investigate representations in cognitive systems, it is necessary to have an appropriate theoretical framework that could explain the phenomena of interest, provide constraints in building synthetic experiments, and provide means to analyze the phenomena. There is a long history of debates around the theme of representations in cognitive science but no adequate theoretical framework has been consolidated in the research field, thus it is necessary to describe the framework in use.

In computational simulation works dealing with the emergence of communication, there is always something that is communicated from an agent to another one, and that is given various names: signal, symbol, sound, word, expression, or utterance. In most of these works, that what is communicated also seems to have representation capabilities. We have used the term representation, in the first section, to emphasize this and also to apply a more familiar word as used by the artificial intelligence community. Nevertheless, we will now use the expression 'sign', as a technical term in our theoretical background.

C.S.Peirce defined semiotics as the 'formal science of signs'. Peirce's semiotics is considered a strongly consistent theory and his theory and models for sign process have been applied and had deep impact in various research fields: philosophy, psychology, theoretical biology, and cognitive sciences.

A sign is defined, following Peirce (1958), as something that refers to something else, an object (which the sign represents in some respect) and produces an effect (interpretant) in the interpreter. For an artificial agent in the context of communication, a sign can be interpreted as being related to some object, its referent, and it produces a motor action as an outcome, due to this sign-object relation.

Signs establish a relation with the object in a variety of ways. And depending on this relation, a sign can become either a symbol, an index or an icon, and this is the 'most fundamental division of signs'. Hence, signs and symbols are not the same thing; a symbol is a special class of sign. Icons are signs that stand for their objects by a similarity or resemblance, no matter if they show any spatio-temporal physical correlation with an existent object. In this case, a sign refers to an object in virtue of a certain quality which is

shared between them. Indexes are signs which refer to their objects due to a direct physical connection between them. Since, in this case, the sign should be determined by the object (e.g. by means of a causal relationship) both must exist as actual events. Spatio-temporal co-variation is the most characteristic property of indexical processes. Symbols are signs that are related to their object through a determinative relation of law, rule or convention. A symbol becomes a sign of some object merely or mainly by the fact that it is used and understood as such by the interpreter, who establishes this connection.

The experiment

In order to study representations processes, we start from the experiment proposed by Mirolli and Parisi (2008). They evolved a population of artificial creatures controlled by feed forward neural networks, which are able to perceive edible and poisonous mushroom and to communicate with each other.

The scenario consists of a one-dimensional corridor with 12 positions where one mushroom and two creatures are placed at a time. There are 420 different types of mushrooms, with visual features codified in a 10 bit vector, being half of those mushrooms edible and half poisonous. Only when a creature is standing in a position right next to the mushroom, it can perceive its visual features. To simulate communication, the signal output from the creature standing next to the mushroom is copied to the signal input in the other creature.

During simulation, two creatures are placed in the corridor, the first (speaker) in one end, next to the mushroom, and the other (hearer) in the opposite end. In the proposed task, the hearer creature should go forward and collect the edible mushroom and to stand still if it is a poisonous one, so every creature in a population of 100 creatures is evaluated as a hearer 420 times, one for each type of mushroom, with a random speaker each time. Note that, at the beginning, the hearer cannot perceive the type of mushroom present in the environment, so to achieve success it could first move next to the mushroom and then visually perceive it, or it could use the auditory output from the speaker to decide which action it should take. Creatures are evaluated positively according to the number of edible mushrooms collected, and negatively according to the number of poisonous mushroom collected and also by the number of forward movements. The creatures are not evaluated as speakers. After all creatures are evaluated, an artificial evolution process occurs: 50 creatures are selected proportionally to their task evaluation, and these selected creatures become parents of the next generation of 100 new creatures (with variation of parents' weights with 0.1% of chance). The simulation continues for 2000 generations.

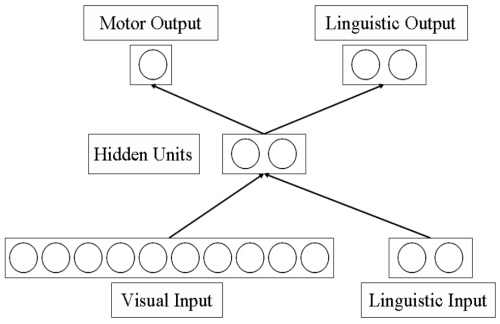

Figure 1: The architecture of the neural network controlling creatures.

Creatures are controlled by a three-layer neural network, with visual input for the mushrooms' visual features, communicative signal input for communication, motor output to determine if it moves forward or stands still, and a communicative signal output (Figure 1). It is important to notice that visual inputs and auditory inputs are connected to the same middle layer's neurons and only these neurons produce motor and auditory outputs. In this topology, consequently, the hearer's behavior and the speaker's behavior of each creature are connected by the middle layer. Actually, Mirolli and Parisi (2008) found out that the speaker produces signal patterns that can be exploited by the hearers because the auditory output layer is connected to the same middle layer neurons that are used for visual input categorization.

To follow the evolution of the population during the experiment, mean population fitness is measured at every generation and also communication quality, as proposed by Mirolli and Parisi (2008). Communication quality measures how distinct signals produced for edible mushrooms are from signals for poisonous ones and how similar signals for the same type of mushroom are. Mirolli and Parisi (2008) also proposed a measure of 'representation quality' taking a similar approach but evaluating activation patterns in the middle layer due to input from signals, i.e., if signals for different mushrooms produce distinct activation patterns

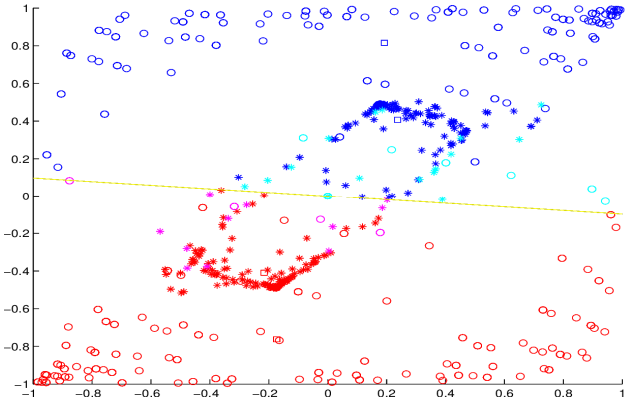

Figure 2: Activation pattern for the neural network middle layer. Circles are activation points due to visual input, asterisk are points for communicative input. Blue and cyan are activations that lead the creature to move, either for an edible mushroom (blue) or, wrongly, a poisonous one (cyan). Red and magenta are activations that lead it to stand still, either for a poisonous mushroom (red) or, wrongly, an edible one (magenta). The yellow line is the motor output

and if signals for the same mushroom produce similar activation pattern.

We will not use this originally proposed measure of 'representation quality'. Instead, we propose to evaluate representation processes as a sign interpretation process, as something that stands for something else and produces an outcome, following Peirce's definition of sign. In this experiment, the sign is the signal produced by a creature, the utterer, and communicated to the other creature, the interpreter. During sign interpretation, the interpreter creature can relate the sign as referring to a type of mushroom (the object) and, due to this relation, take an action of moving or standing still (the interpretant). In this experiment, since the communicated sign has no similarity to the object, so it cannot be an icon; it can't be related to the object by physical connection because the object cannot be perceived by the creature to establish a spatial-temporal relation. The sign can only relate to its object by using an acquired association between them, therefore the sign can only be a symbol.

If the creature uses an association that symbolically relates sign and object, how can we verify this in the neural network? Since the communicative inputs and visual inputs are connected to the same middle layer neurons, then if a communicative input due to a sign produces an activation pattern similar to the activation pattern from the visual input due to the mushroom-object, consequently producing similar motor outcomes, then sign interpretation occurs, relating sign and object. To investigate these activation patterns, we register the activation values from communicative input for each sign produced during the 420 evaluation trials, and we also register the activation values for visual inputs from each of the 420 possible mushrooms.

The middle layer has two neurons, so its activation corresponds to a bi-dimensional vector with each dimension having values in [-1; +1]. Activation patterns in the middle layer can be plotted in a bi-dimensional graph as illustrated in figure 2. The activation from middle layer is forwarded to the motor output neuron that can be either +1 and the creature walks forward, or -1 and the creature stands still. There is a the decision boundary between this two actions corresponding to a line that can be calculated and plotted along with activation values, as shown in figure 2.

Results

The experiment was simulated 15 times and the results that are representative of the overall behavior observed are as following. The obtained results on population fitness and communication quality along generations, as shown in figure 3, are similar to the ones obtained by Mirolli and Parisi (2008). A reliable communication system does not stabilize, creatures alternate between reliable and deceiving utterers because of competition for selection. Interpreter creatures exploit utterers biases due to selective pressure to categorize mushrooms, and then utterer change signals to exploit interpreters biases in a deceiving way. Creatures are competing for selection and all of them are evaluated as

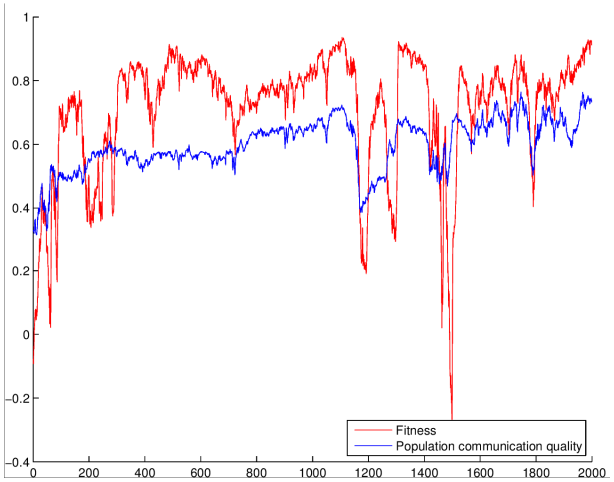

Figure 3: Population mean fitness and population communication quality.

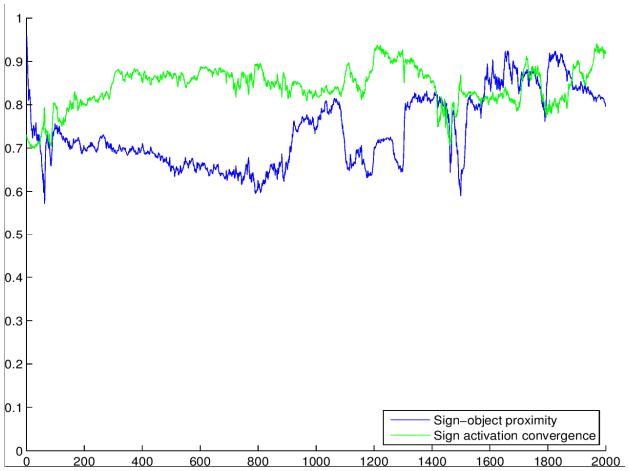

Figure 4: Sign-object proximity and sign activation convergence.

interpreters. But a creature can negatively influence other creatures' evaluation if it (utterer) deceives other creatures (interpreters), thus lowering competitors evaluation score (for further details, see Mirolli and Parisi, 2008).

In order to evaluate representation processes underlying communication, we measure the distance between middle layer activation for communicated sign inputs and for visual inputs from mushrooms. As can been seen in figure 2, it is possible to distinguish four point clouds: one for blue and cyan circles corresponding to activations from visual inputs that lead to forward movement (VM cloud); one for blue and cyan asterisks for communicated sign input with the same motor outcome (CM cloud); one for red and magenta circles corresponding to activations for visual inputs leading to standing still (VS cloud); and one for red and magenta asterisks for communicated sign inputs with the same motor outcome (CS cloud). A possible evaluation of sign-object reference relation is to estimate sign-object proximity as the complement of the normalized mean value of the distances between the center of VM and the center of CM and

between the center of VS and the center of CS. That gives a rough valuation of referential association between sign and object, as shown in figure 4, the higher the value is, the closer is the sign from the object.

Notice that we separate activation points according to the interpreter perspective, according to its action, and not according to the actual mushroom type. By identifying the sign interpretant, i.e. the effect the sign produces in the interpreter, its motor action, we are able to properly group activations according to interpretation. A misinterpretation according to external observer is still an interpretation, so even if, for example, the interpreter moves for a poisonous mushroom, for the interpreter it is a mushroom to move forward to. Moreover, there is always an interpretation for every sign, being it an interpretation that contributes positively or negatively to fitness.

Sign input activation clouds can have varying distribution of activation value points. Due to alternating quality of the communication system, utterers may change the sign produced for a given mushroom and interpreters tend not to stabilize sign-referent associations with sign input activations scattering. To better trail these changes in the interpreter, we also calculated the sign activation convergence as the complement of the normalized distance from points in a communicative sign cloud to cloud center, as shown in figure 4, the higher the value is, the closer the points inside the cloud are to each other.

From sign-object proximity and sign activation convergence, we can evaluate the interpretation and representation processes underlying the communication evolution in the proposed scenario. It is possible for example to assess if sign-object proximity is due to a great dispersion of activation points (as in the start of the experiments with random neural nets) or because communicative and visual clouds are compact and close to each other. Before we proceed, it is important to explain that the distance measures were normalized by the largest possible distance, but actual point distances tend to be much smaller than this. Even at simulation start, distances can be small compared to the largest distance, so it is important to observe changes in proximity and convergence values, and not their absolute values.

Examining sign-object proximity and sign activation convergence in figure 4, it can be noticed that sign-object proximity starts with a value near 1,0 but at the same time sign activation convergence has its lowest values. This is due to the random neural nets at the beginning of the simulation, with almost uniform distribution and cloud centers near the center of the graph. After a few generations, sign activation convergence increases with sign-object proximity as activation clouds get better defined but cloud centers are not so close to each other.

As interpreter sign clouds do not stabilize due to changing use of signs by utterers, they tend to stay close to the action decision line and far from the visual activation clouds, even if fitness is high (Figure 5). Besides, sign activation point for deceiving interpretation points also tends to be closer to

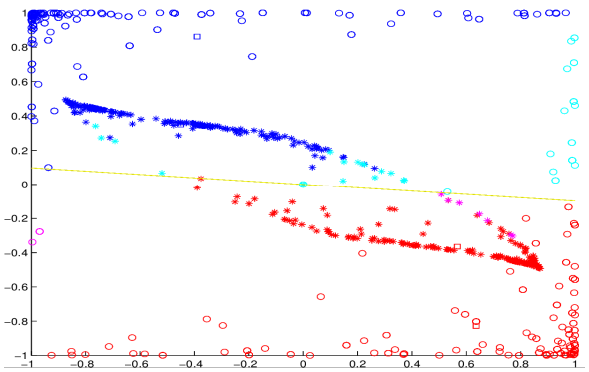

Figure 5: Activations for the best creature in generation 1030 with high population fitness in the original experiment.

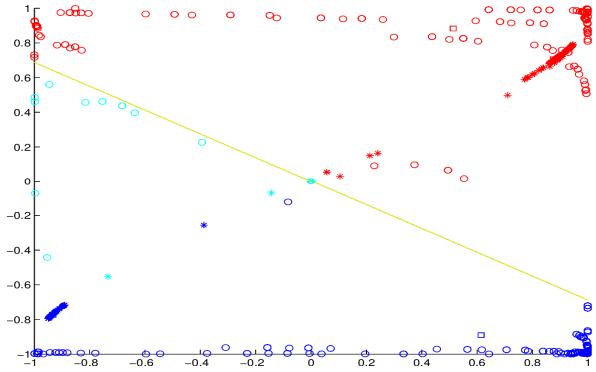

Figure 6: Activations for the best creature in generation 1275 with high population fitness in the modified experiment.

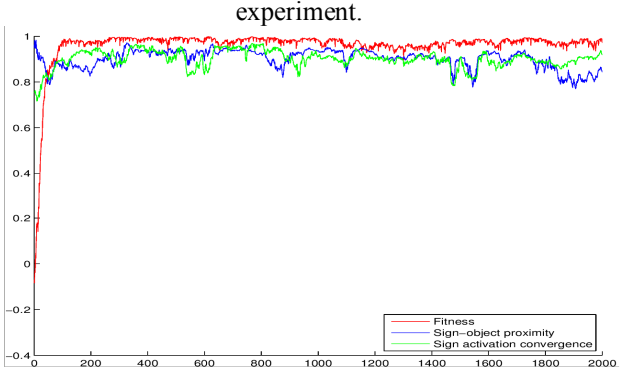

Figure 7: Sign-object proximity and sign activation convergence.

the motor decision line, evidencing that the interpreter is in the process of adaptation to the communication system. Activation points for visual inputs tend to have a completely different distribution pattern, concentrating on extreme values, since mushrooms do not change along generations allowing for a better adaptation.

To compare these results were utterers and interpreters are competing with each other, we modified the experiment for creatures to communicatively cooperate. In this modified experiment, to obtain cooperation, the interpreter will always have a kin creature as utterer, i.e. a creature with the same neural connection weights (see Floreano, 2007). In

this experiment's simulation, creatures' fitness rapidly increases and stays near maximum for the rest of the simulation (figure 7) and referential relations are also quite different. Figure 6 shows the activation points for a creature in this second experiment. Notice that sign activation clouds are much more compact and farther from the motor decision line then in the original experiment. They are also closer to the visual activation cloud, with signs closer to referents. This convergence of sign activations and high proximity of sign-object is observed after a few generations and stays that way in the following generations as shown in figure 7, where the values are consistently higher than in the original experiment.

Final Remarks

In the experiments presented, an investigation on representation and referential processes involved in sign-object relation during the evolution of communication has been started. This is only an initial study that evidences that there is the possibility of analysis of such processes by evaluating neural networks activation and bringing forth preliminary conclusions.

The creatures had an artificial neural network as their cognitive architecture. Since communicated signals are connected to the same neurons as visual inputs from mushrooms, such signals can produce an activation pattern similar to the ones produced by visual inputs, and produce an equivalent motor outcome. This elicits a referential association between signals and mushrooms established only by the interpreter, determining a sign-object symbolic relation.

Establishing a strong sign-object relation with signs, producing activations that are close to the activations from the object, appears to be adaptively beneficial to creatures. Interpreters search for this quality in referential relations during evolution process, but since the communicated signals used by utterers changed constantly in the first experiment, it was not possible for interpreters to better improve sign-object relations, but even in such a competitive scenario a referential relation was created. In the second experiment, however, cooperation aided this adaptation process, and interpreters could establish better quality referential associations for signs.

This article presents only initial findings on the investigation of representational and referential processes in neural networks. Initial results were presented but further scrutiny and discussion of these simulation results will be done in the future. Nevertheless, the proposed methodology is promising and may open up a new perspective on the studies on representations.

The fact that the neural network used had only feed forward connections and no recurrent ones was important for our investigation. It was easier to determine causal relations between input neurons, intermediary neurons and output neurons, and then to point out and compare activation patterns. An open issue is how to conduct such analysis on recurrent networks with time-dependent activations.

References

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4), 209–243.

Cangelosi, A. (2001). Evolution of communication and language using signals, symbols, and words. IEEE *Transactions on Evolutionary Computation*, 5(2), 93–101.

Christiansen, M.H., & Kirby, S. (2003). Language evolution: consensus and controversies. *Trends in Cognitive Sciences*, 7 (7), 300-307.

De Greeff, J. & Nolfi, S. (2010). Evolution of implicit and explicit communication in mobile robots. *In Evolution of Communication and Language in Embodied Agents,* 179–214. Springer Verlag.

Floreano, D., Mitri, S., Magnenat, S., & Keller, L. (2007). Evolutionary conditions for the emergence of communication in robots. *Current Biology*, 17, 514-519.

Loula, A., Gudwin, R. & Queiroz, J. (2013) Studying sign processes in the emergence of communication. In *Proceedings of 35th Annual Meeting of the Cognitive Science Society*, *CogSci* 2013, 2013, Berlin.

Loula, A., Gudwin, R. & Queiroz, J. (2011) Cognitive conditions to the emergence of sign interpretation in artificial creatures. In *Proceedings of the 11th European Conference of Artificial Life*, ECAL'11, 2011, France. (p. 497-504)

Loula, A., Gudwin, R., & Queiroz, J. (2010) On the emergence of indexical and symbolic interpretation in artificial creatures, or What is this I hear? In Fellermann, H., et al., editors, *Artificial Life XII*, pages 862-868. MIT Press.

Marocco, D. & Nolfi, S. (2007). Emergence of communication in embodied agents evolved for the ability to solve a collective navigation problem. *Connection Science*, 19(1), 53–74.

Mirolli, M. (2012). Representations in Dynamical Embodied Agents: Re-Analyzing a Minimally Cognitive Model Agent. *Cognitive Science*, 36, 870-895.

Mirolli, M., Parisi, D. (2008): How producer biases can favor the evolution of communication: An analysis of evolutionary dynamics. *Adaptive Behavior*, 16(1): 27-52.

Mirolli, M., Parisi, D. (2005): How can we explain the emergence of a language that benefits the hearer but not the speaker? *Connection Science*, 17(3-4): 307-324.

Nolfi, S. & Mirolli, M., Eds. (2010). Evolution of Communication and Language in Embodied Agents. *Springer*.

Peirce, C. S. (1958). The Collected Papers of Charles Sanders Peirce. Cambridge, Mass., USA: Harvard University Press.

Wagner, K., Reggia, J., Uriagereka, J., & Wilkinson, G. (2003) Progress in the simulation of emergent communication and language. *Adaptive Behavior*, 11(1), 37-69.