

Acquiring Vocabulary through Human Robot Interaction: A Learning Architecture for Grounding Words with Multiple Meanings

Aneesh Chauhan¹ and Luís Seabra Lopes^{1,2}

Actividade Transversal em Robótica Inteligente

IEETA¹/DETI² Universidade de Aveiro

3810-193 Aveiro, Portugal

{aneesh.chauhan, lsl}@ua.pt

Abstract

This paper presents a robust methodology for grounding vocabulary in robots. A social language grounding experiment is designed, where, a human instructor teaches a robotic agent the names of the objects present in a visually shared environment. Any system for grounding vocabulary has to incorporate the properties of gradual evolution and lifelong learning. The learning model of the robot is adopted from an ongoing work on developing systems that conform to these properties. Significant modifications have been introduced to the adopted model, especially to handle words with multiple meanings. A novel classification strategy has been developed for improving the performance of each classifier for each learned category. A set of six new nearest-neighbor based classifiers have also been integrated into the agent architecture. A series of experiments were conducted to test the performance of the new model on vocabulary acquisition. The robot was shown to be robust at acquiring vocabulary and has the potential to learn a far greater number of words (with either single or multiple meanings).

Introduction

In recent years, a significant progress has been made in designing conversational robots capable of having a dialog with one or more human participants in relatively unrestricted environments (Bohus and Horwitz 2009; Gold et al. 2009; Mutlu et al. 2009). Apart from focusing on various communication challenges (e.g. speech recognition, voice synthesis), most of these approaches take cues from human-human discourse behavior (eye gaze movements, hand gestures, selective attention, target recognition and tracking etc.) to build robotic agents that can converse in a human-like manner.

Although such systems have been shown robust in a variety of real world scenarios, they lack the semantic perception of their language of communication. That is, these systems operate on the language symbols (words) and produce a reply. The meaning interpreted from these replies lies inside the head of the person interpreting them and the algorithm designer, but not inside the computer

manipulating these symbols (Harnad 1990; Searle 1980). To give conversational robots increased cognitive plausibility, it is essential that they have the capacity to ground human language in their perception. This paper is a product of one such effort in this direction.

Words are the basic tokens of our language. They are essentially symbols which do not contain an independent meaning. Their meanings lie in their association with the entities of the world (Barsalou 1999; Harnad 1990). Associating a word to its referent is an instance of the symbol grounding problem (Harnad 1990). There is also a growing view amongst linguists that language is a cultural product, which is transmitted (or spreads) socially (Cowley 2007; Love 2004). These theories have pushed forward a new set of ideas on how to approach language development in robots.

Various studies have been carried out on populations of robots to investigate language origins, transfer and evolution. Some researchers have focused on the modes of language transfer (Loreto and Steels 2007; Nowak, Plotkin and Krakauer 1999; Steels 2002), while others on the methods of internal grounding (Gold et al. 2009; Roy and Pentland 2002; Seabra Lopes and Chauhan 2007, 2008). Much of this work has been carried out in the domain of grounding vocabulary in visual perception.

Inspired by these studies, an approach to grounding vocabulary through social interaction is presented. An experiment is designed, where a human, acting as an instructor, teaches a robotic agent the names of the objects present in their visually shared environment.

Similar experiments have previously been reported, where the number of learned words ranged from 3 to 68 (Levinson et al. 2005; Roy and Pentland 2002; Seabra Lopes and Chauhan 2007, 2008; Steels and Kaplan 2002). Several of these authors have pointed out the need for scaling up the number of acquired categories in language acquisition and symbol grounding systems.

Very few vocabulary acquisition models account for words with multiple meanings. Notable being Gold's (Gold et al. 2009) social word learning model based; a child psychology inspired early word learning model of Regier

et al. (2001); and the model used in the language games of (Nowak, Plotkin and Krakauer 1999). The later two models describe plausible associative homonym formation models, but they are not models of vocabulary acquisition. The language acquisition model of Gold et al (2009) is based on dynamic decision trees, which can account for words with both single and multiple meanings, but is very limited in its vocabulary acquisition capabilities and overall performance. The system of Roy and Pentland (2002) took multiple views of objects into account when learning their names. In practice, these different views can be taken as different meanings of the learned words.

With an aim of acquiring larger vocabularies, our recent research focused on improving the learning model described in (Seabra Lopes and Chauhan 2008). While retaining the basic building blocks of this system, a new classification strategy for the existing classifiers and a set of six new classifiers have been introduced. Apart from enabling the agent to learn the regular words faster, these modifications also help in addressing words with multiple meanings.

The category names taught to the agent presented in this paper, always refer to real world (solid) objects. The choice of using solid objects is justified by analogies with early language development in children. In fact, most of the early vocabulary of children consists of common nouns (that name objects such as food items, toys etc.) (Bloom 2001; Messer 1997).

The agent is embodied with physical devices, namely a computer for interaction, visualization and internal computations, and a camera and a robotic arm to help it perceive as well as operate in its surroundings.

This paper is structured as follows: Section 2 details the Human-Robot Interaction (HRI) framework for learning words. Section 3 describes the novelties in the modified concept learning and categorization architecture of the agent. Section 4 presents the experiments and discusses the obtained results and Section 5 presents the conclusions.

Human-robot interaction for learning words

The primary purpose of a language is to communicate about the entities of the world. Meaning formation, on the other hand, is a cognitive task of representing these entities in an individual's brain. Although linked, language (as a communication tool) and the formation of meaning are two separate cognitive tasks. Any two individuals share a language when they have the same words grounded to the same entities, regardless of their respective processes of meaning formation. A robot can learn a human language if it can ground the human language symbols (words) in sensor-based descriptions.

In this work, a human instructor is used to teach the names of the objects present in their visually shared environment. These names are then grounded by the robotic agent in sensor-based descriptions, leading to a vocabulary shared with its instructor.

Shared attention between the instructor and the robot is established if the instructor, by mouse-clicking, selects an object from the robot's visible scene (camera frame) displayed on the screen, see Fig. 1a. The instructor can interact with the robot through the following instructions (using a menu-based interface):

1. Teach the category name of the selected object;
2. Ask the category name of the selected object;
3. If the category predicted in the previous case is wrong, provide the true category.
4. Provide a category name and ask the robot to locate an instance of that category;
5. If the object identified by the robot in the previous case does not belong to the requested category, provide the true category.

Besides recording information given through teach and correct actions, the robot interacts with the human by responding to the posed questions. Depending on the question, the robot can respond in either of the following ways:

1. Linguistic response: provide the categorization result;
2. Visual response: visually report the results of the "locate" task.

Simulated user agent. Using a human to teach is an extremely exhaustive task (previous experiments took weeks to accomplish). Therefore, a simulated user was designed and developed for the experiments reported in this paper. The actions of this agent are limited to the following actions of the human user: teaching, asking and correction. From many previous human-robot interactions, a database of ~7500 images (from 69 categories) has been collected (Seabra Lopes and Chauhan 2007, 2008). The extracted object in Fig. 1 can give an idea of the type of images in this database. These images and their names (the image-name database) provide accurate material for naturalistic simulations.

Concept learning and categorization

The category learning architecture has been adopted from (Seabra Lopes and Chauhan 2008), which has previously been reported to show good performance in acquiring vocabulary.

It uses an instance based approach for category representation, where categories are represented by the sets of known instances. The new instances are stored only when there is a direct intervention from the human instructor: an explicit teaching action; or a corrective feedback.

This architecture has provisions for:

- Multiple object and category representations; and
- Multiple classifiers and classifier combinations (based on majority voting and Dempster-Shafer evidence theory);

Base classifiers result from applying specific similarity measures to specific feature spaces. In the implementation

of the model, 11 base classifiers and 7 classifier combinations were included.

To be more adaptive and to improve learning performance as well as memory usage, the architecture includes a metacognitive processing component. All learning computations are carried out during the normal execution of the agent, which allows continuous monitoring of the performance of the different classifiers. The measured classification successes of the individual classifiers support an attentional selection mechanism, through which classifier combinations are dynamically reconfigured and a specific classifier is chosen to predict the category of a new unseen object.

One limitation of the previous system is its inability to handle words with multiple meanings. For example, *Stapler1*, *Stapler2* and *Stapler3* were all being taught to the agent as separate categories (Fig. 1b), while in reality they are instances of the same category, *Stapler*.

This system is also incapable of handling “broader” categories, that is, when multiple categories are contained in a broader category (Fig. 1c) labeled by a single name. This limitation is not restricted to this particular model. Most of the research on vocabulary acquisition models has focused on words that have only one meaning (simple words).

In the original system, the principal cause for failure in handling multiple meanings of words is the categorization mechanism. Categorization of a new instance involved ranking the known categories according to measures of membership of the instance to each of the categories. Each measure is computed as an average over all the instances that describe a category. Two basic measures were used, namely Euclidean Distance and Pyramid Match Score (Grauman and Darrell, 2007). From these, two category membership measures were derived. The Euclidean Membership Measure is defined as follows:

$$EuclidMem(C_i) = \frac{N}{D_i \sum_{k=1}^N (1/D_k)} \quad (1)$$

where N is the number of categories, $i, k=1, \dots, N$, and D_i and D_k are the average Euclidean distances of the target object to the known instances of categories C_i and C_k , respectively. The membership values $EuclidMem(C_i)$ sum to 1.0, allowing their use as evidence in Dempster-Shafer combinations.

Similarly, the following Pyramid Membership Measure was defined:

$$PyramidMem(C_i) = \frac{N \cdot P_i}{\sum_{k=1}^n P_k} \quad (2)$$

where P_i and P_k are the average pyramid match scores of the target object to the known instances of categories C_i and C_k , respectively, and the rest as in the previous case.

Averaging the measurements severely limits the categorization accuracy in the cases where the instances

describing a category vary significantly from each other (heterogeneous categories).

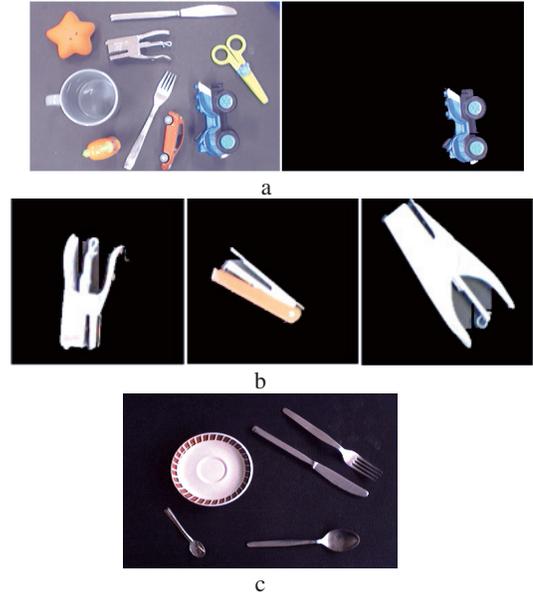


Figure 1. a) Robot’s visual scene and an extracted object as selected by the user b) Example objects of categories *Stapler1*, *Stapler2* and *Stapler3*, that differ in shape but belong to the same category (multiple meanings); c) A set of different objects that can all be said to belong to the “cutlery” category.

The following subsections will detail the modifications introduced to this model for improving its robustness in handling words with multiple meanings.

Nearest-neighbor classifiers

A set of six new classifiers, all based on the nearest-neighbor (NN) principle, were added to the existing system. Given an object to be classified, it is compared with all the instances stored in memory and the category containing the instance most similar to the input object is predicted as its category.

As in previous classifiers, category membership measures are based on Euclidean Distance and Pyramid Match Score. The Euclidean Membership Measure of the target object to a given category C_i is computed by inverting Euclidean distances:

$$EuclidMemNN(C_i) = \frac{1}{\min D_i \sum_{k=1}^N (1/\min D_k)} \quad (3)$$

where N is the number of categories, $i, k=1, \dots, N$, and $\min D_i$ and $\min D_k$ are the minimum Euclidean distances of the target object to the known instances of categories C_i and C_k , respectively.

Derived from the pyramid match kernel, the Pyramid Membership Measure for a particular target object and category C_i can be computed as:

$$PyramidMemNN(C_i) = \frac{maxP_i}{\sum_{k=1}^N maxP_k} \quad (4)$$

where $maxP_i$ and $maxP_k$ are the maximum pyramid match scores of the target object to the instances of categories C_i and C_k , respectively, and the rest as before.

The Pyramid Membership (PM) and Euclidean Membership (EM) measures are used for the new NN classifiers. These classifiers work on different sets of shape features. Following features (Seabra Lopes and Chauhan 2008) and membership measures were used to implement the classifiers:

- “shape slices normalized radii averages” (EM and PM);
- “shape layers histogram” (EM and PM);
- “shape slices histogram” (PM); and
- “shape slices normalized radii standard deviations” (PM)

Nearest-cluster classifiers

An original classification method has been developed that facilitates and improves the classification performance of each of the base classifiers inherited from the previous version of the system (with exception of a color-based classifier, excluded due to poor performance). The approach involves locating the nearest neighbor of each instance and clustering the instances that are connected to each other through their nearest neighbors (see Fig 2). Thus, each category will be represented by sets of clusters of known instances (i.e. stored in memory).

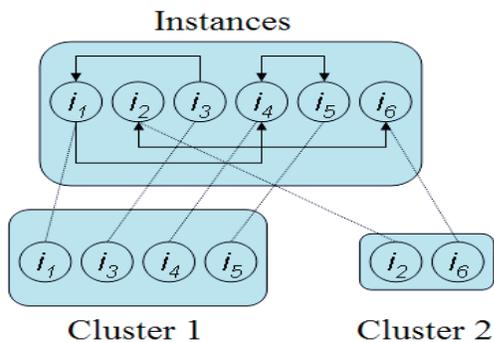


Figure 2. Cluster formation for a set of 6 instances (the nearest neighbor of an instance is pointed by the head of the arrow originating from that instance)

Since each base classifier applies a specific membership measure to a specific feature space, the instances within a category will be organized into different sets of clusters as appropriate for the different classifiers (Fig. 3). Each time there is a change in the set of instances of a category, the clustering process for this category is run once more, producing a new set of clusters of this category.

For each of the nearest-cluster classifiers, instead of computing average similarity measures over all the instances of a category description, as in the original

system, these measures are computed for the instances in each cluster. Thus, equations 1 and 2 are now applied to clusters rather than whole categories. For each category, the cluster with the highest average similarity to the target object will provide the membership score of the category.

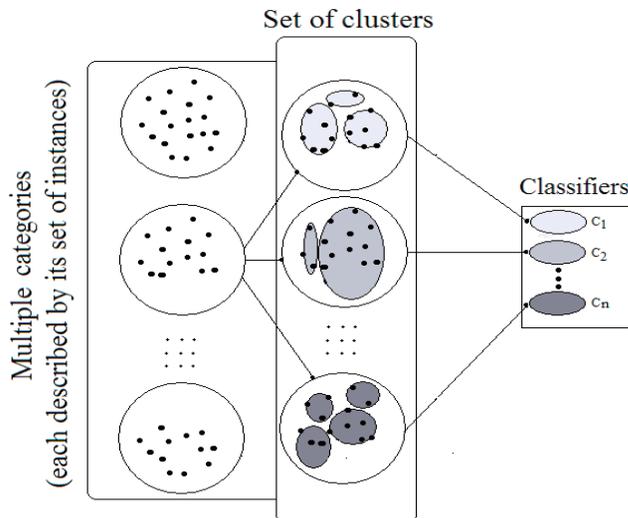


Figure 3. A conceptual illustration of the set of clusters of a category description, for a set of “n” classifiers (the instances are symbolized by dots).

This classification approach has significant implications on the robustness and flexibility of the learning model. If the feature space and similarity measures of a classifier are good, the instances of a category with similar poses will cluster together for that specific classifier. Similarly, by bringing the similar instances together, the cluster organization can account for words with multiple meanings, i.e., that have more than one subcategory associated to them. In general, this strategy will improve the classification performance of each classifier as well as the overall vocabulary acquisition capacity of the robotic agent.

Experimental evaluation

The objective of the experiments reported here is to evaluate the developed system with respect to vocabulary acquisition performance and robustness at handling words with multiple meanings. The performance of the learning model was evaluated using the teaching protocol and the classification precision measure initially proposed in (Seabra Lopes and Chauhan 2007) and used in other more recent papers (e.g. Seabra Lopes and Chauhan 2008). The precision measure is used to analyze the impact of the introduction of a new category on a learning system, from a possible initial instability to the final recovery. A new category is introduced only if the precision measure is above a certain threshold. Breakpoint is reached when the learning system stops showing signs of evolution or recovery.

The experiments were carried out using the simulated user agent. The categories and objects are selected randomly from the image-name database (of 69 categories), hence preserving the essence of natural interactions. When this agent runs out of images of a particular category, the human user is called to show a new object.

The performance of the learning model is evaluated over 20 experiments, which are divided into two sets:

1. First 10 experiments were conducted to assess the agent’s vocabulary acquisition performance for simple words (words with single meanings, including some names invented for subcategories without a natural name) at various precision thresholds;
2. Next 10 experiments were performed to evaluate the learning performance on acquisition of real vocabulary (consisting of both words with single and multiple meanings, and no invented names).

Each experiment is evaluated using two measures: externally observable performance of the agent and average classification success. Both evaluation measures are computed once an experiment has concluded. Externally observable performance is given as the percentage of correct predictions made during an experiment. Average classification success is an average of the classification precision values internally computed over all the question-correction iterations.

Experiments on words with single meanings

The first set of experiments (words with single meanings) was carried out with different values of precision thresholds. For each threshold two experiments were conducted. Table I provides the summary of these experiments.

Table I Summary of experiments on words with single meanings

Exp #	Prec. threshold (%)	# iterations	# cats	Avg. # instances /cat	Ext. obsv. perf. (%)	Avg. class. success (%)
1	50	2415	69	10.01	74.24	76.92
2	50	2415	69	9.41	75.98	79.74
3	66.67	2554	69	10.54	74.24	73.57
4	66.67	2532	69	10.75	73.42	76.28
5	70	2515	69	9.54	76.58	82.44
6	70	3076	69	12.72	73.70	72.89
7	80	3009	69	10.77	77.60	81.39
8	80	3398	69	12.39	76.87	80.13
9	100	614	9	7.44	90.55	89.78
10	100	384	9	4.56	91.67	90.72

The learning agent was able to learn the names of all 69 categories in the first 8 experiments with an average system precision of 77.9% (± 3.56). This means, irrespective of the precision threshold used in the first 8 experiments, the agent was able to successfully learn 69 category names with a consistent externally observable performance roughly between 74% and 81%. Similar

values were obtained for average classification success measure.

Increasing the precision threshold forces the agent to form better descriptions of the existing categories before next category can be introduced. As a consequence, the number of question-correction iterations also increases. There is also a tradeoff between the precision threshold and the number of categories learned. In the experiments 9 and 10, the system precision is $\sim 90\%$ and prediction accuracy is $\sim 91\%$. But the target precision threshold of 100% was not achieved after the introduction of the 9th category.

Figure 4 shows the evolution of classification performance of the 8th experiment. The depression in the graph indicates the periods after the introduction of a new category. In general, the introduction of a new category will affect the prediction of the existing categories. The categories that have been affected will be predicted incorrectly, hence reducing the classification precision of the system. Each incorrect prediction will lead the simulated user to send a correction (hence modifying the description of that category). This process is continued till the classification precision reaches the precision threshold. Note that the points where the values in the graph are zero are an indicator that after the introduction of a new category, the first category prediction was incorrect (following the teaching protocol).

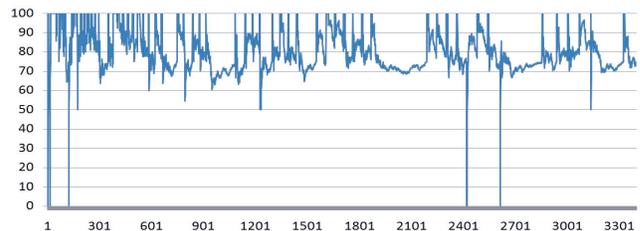


Figure 4. Evolution of classification precision versus number of question/correction iterations in experiment 8.

Overall, each of the first 7 experiments followed the same evolution pattern as that of the 8th experiment. For the experiments 7-8, the number of iterations is higher because of the higher precision threshold (80%).

Experiments on real vocabulary

The second set of experiments was conducted on a real vocabulary. The image-name database was modified to reflect real human categorization of the 69 simple categories. After the modifications, seven words had 2 meanings, two words contained 3 meanings, one word had 4 meanings, one word with 6 meanings and thirty eight words with 1 meaning each. This led to a total of 49 categories. The precision threshold was set to 77%.

Fig. 5 shows the evolution of classification precision for the 5th experiment on real vocabulary acquisition (rest of the experiments follow a similar pattern). The learning system successfully learned all the 49 categories in each of these experiments.

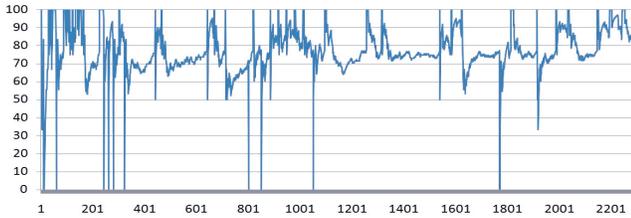


Figure 5. Evolution of classification precision versus number of question/correction iterations for the 5th experiment on real vocabulary acquisition.

The prediction accuracy and the system precision over these experiments were 76.3% (± 0.64) and 75.9% (± 1.15) respectively. At the precision threshold of 77%, the externally observable performance and the average classification success are consistently $\sim 77\%$. The consistent equivalence between the system threshold, the externally observable performance and the average classification success demonstrates the learning stability achieved in the final experiments. In other words, at the precision threshold of 77%, the learning system has the potential to learn many more real categories, independent of the type of words being taught.

Conclusion

This paper presented a social language grounding scenario, where a human instructor teaches a robotic agent the names of the objects present in a shared environment. The agent grounds these names in sensor-based category descriptions. The learning model of the agent is adopted from (Seabra Lopes and Chauhan 2008). The primary aim of the paper was to modify this model such that words with multiple meanings could be learned more easily.

To tackle this problem, a set of six new classifiers, based on the nearest-neighbor principle, were introduced in the learning model implementation. In addition, novel cluster-based categorization strategy was introduced to improve other classifiers already existing in the previous version of the system. The classification strategy takes into the account the similarity between the instances describing a category with respect to each classifier. This strategy can account for the incremental changes in the class descriptions (e.g. addition of new instances to existing categories or introduction of previously unknown categories). A clustering strategy based on chaining nearest-neighbors was also presented to identify multiple subcategories in a single category description.

Two sets of experiments were conducted to evaluate the vocabulary acquisition performance of the new learning model. 10 experiments each were carried out to assess the model's performance on learning words with single meanings and the real vocabulary respectively. The robotic agent successfully acquired all the vocabulary in most of the experiments (experiments with the precision threshold set to 100% could only learn 9 categories).

References

- Barsalou, L. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–609.
- Bloom, P. 2001. Word learning. *Current Biology*, 11:5-6.
- Bohus, D.; and Horvitz, E. 2009. Dialog in the Open World: Platform and Applications. ICM1'09, MA.
- Cowley, S. J. 2007. Distributed language: Biomechanics, functions and the origins of talk. In C. Lyon, C. Nehaniv & A. Cangelosi (Eds.), *Emergence of communication and language*. Springer, 105-127.
- Gold, K.; Doniec, M.; Christopher, C.; and Scassellati, B. 2009. Robotic Vocabulary Building Using Extension Inference and Implicit Contrast. *Artificial Intelligence* 173(1):145-166.
- Grauman, K.; and Darrell, T. 2007. The Pyramid Match Kernel: Efficient Learning with Sets of Features. *Journal of Machine Learning Research*, 8:725-760.
- Harnad, S. 1990. The symbol grounding problem. *Physica D*, 42:335-346.
- Levinson, S. E.; Squire, K.; Lin, R. S.; and McClain, M. 2005. Automatic language acquisition by an autonomous robot, *Proceedings of the AAAI Spring Symposium on Developmental Robotics*.
- Loreto, V.; and Steels, L. 2007. Social dynamics: Emergence of language. *Nature Physics*, 3(11):758-760.
- Love, N. 2004. Cognition and the language myth. *Language Sciences*, 26:525-544.
- Messer, D. J. 1994. *The Development of Communication*. West Sussex, England: Wiley, 1994.
- Mutlu, B.; Shiwa, T.; Kanda, T.; Ishiguro, H.; and Hagita, N. 2009. Footing in Human-Robot Conversations: How Robots Might Shape Participant Roles Using Gaze Cues. In *Proceedings of HRI'09, San Diego, CA*.
- Nowak, M. A.; Plotkin, J.; and Krakauer, D. 1999. The evolutionary language game. *Journal of Theoretical Biology*, 200:147-162.
- Regier, T.; Corrigan, B.; Cabasaan, R.; Woodward, A.; Gasser, M.; and Smith, L. 2001. The Emergence of Words. *Proceedings of CogSci 2001*, 815-820.
- Roy, D.; and Pentland, A. 2002. Learning words from sights and sounds: A computational model. *Cognitive Science*, 26:113-146.
- Seabra Lopes, L.; and Chauhan, A. 2007. How many Words can my Robot learn? An Approach and Experiments with One-Class Learning. *Interaction Studies*, 8(1):53-81.
- Seabra Lopes, L.; and Chauhan, A. 2008. Open-ended category learning for language acquisition. *Connection Science*, 20(4):277-297.
- Searle, J.R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3:417-457.
- Steels, L. 2002. Language games for autonomous robots. *IEEE Intelligent Systems*, 16(5):16-22.
- Steels, L.; and Kaplan, F. 2002. AIBO's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1):3-32.