Robots That Learn to Communicate: A Developmental Approach to Personally and Physically Situated Human-Robot Conversations *

Naoto Iwahashi[1], **Komei Sugiura**[1], **Ryo Taguchi**[2], **Takayuki Nagai**[3], **Tadahiro Taniguchi**[4]
[1] National Institute of Information and Communications Technology, Seika-cho, Kyoto, Japan
[2] Nagoya Institute of Technology, Nagoya, Aichi, Japan
[3] University of Electro-Communications, Chofu, Tokyo, Japan
[4] Ritsumeikan University, Kusatsu, Shiga, Japan

Abstract

This paper summarizes the online machine learning method LCore, which enables robots to learn to communicate with users from scratch through verbal and behavioral interaction in the physical world. LCore combines speech, visual, and tactile information obtained through the interaction, and enables robots to learn beliefs regarding speech units, words, the concepts of objects, motions, grammar, and pragmatic and communicative capabilities. The overall belief system is represented by a dynamic graphical model in an integrated way. Experimental results show that through a small, practical number of learning episodes with a user, the robot was eventually able to understand even fragmental and ambiguous utterances, respond to them with confirmation questions and/or actions, generate directive utterances, and answer questions, appropriately for the given situation. This paper discusses the importance of a developmental approach to realize personally and physically situated human-robot conversations.

Introduction

In order to support human activities in everyday life, robots should adapt their behavior in response to situations which differ from user to user. One of the essential features of such adaptation is the ability of a robot to share experiences with the user in the physical world. This ability should be considered in terms of spoken language communication, which is one of the most natural interfaces.

The process of human communication is based on certain beliefs shared by those communicating (Sperber & Wilson 1995). Language is one such shared belief that is used to convey meanings based on its relevance to other shared beliefs. These shared beliefs are formed based on sharing interactive experiences with the environment and with other people, and the meanings of utterances are embedded in these shared experiences. Therefore, language processing methods must make it possible to reflect shared experiences.

However, existing language processing methods, which are characterized by fixed linguistic knowledge, do not make this possible (Allen *et al.* 2001). In these methods, information is represented and processed by symbols whose meaning has been predefined by the machines' developers. Therefore, experiences shared by a user and a machine under personal situations in the physical world can neither be expressed nor interpreted. As a result, users and robots fail to interact in a way that accurately reflects shared experiences.

To overcome this problem and to achieve natural dialog between humans and robots, we should use methods that satisfy the following requirements in terms of beliefs the robots have:

1) **Grounding:** Beliefs that robots have must be grounded in the personal physical world. Hence, linguistic beliefs should be represented by an integrated system, including other cognitive beliefs regarding perception, actions, and so on. The theoretical framework for grounding language was presented in the Reference (Roy 2005). Several computational studies have explored the grounding of the meanings of utterances in conversations in the physical world (Winograd 1972; Shapiro *et al.* 2000). These previous works, however, have not pursued the learning of new grounded beliefs.

2) **Scalability:** The situation in interactions between a user and a robot changes continuously. To enable robots to execute linguistic communication appropriately in a new situation, grounded linguistic beliefs should be transmutable and scalable. Robots themselves must be able to learn new beliefs that reflect their experiences.

* This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 20500186, 2010.

3) **Sharing:** Because utterances are processed on the basis of the beliefs assumed to be shared by a user and a robot, the grounded beliefs, which are learned, should be shared. To form such shared beliefs, the robot should possess a mechanism that enables the user and the robot to infer the state of each other's belief systems in a natural way by coordinating their utterances and actions.

All these requirements show that learning ability is essential. Cognitive activities related to *grounding*, *scalability*, and *sharing* can be observed clearly in the process of language acquisition by infants as well as in everyday conversation by adults. Therefore, we have been developing a method that enables robots to acquire linguistic communication capabilities from scratch through verbal and nonverbal interaction with users, instead of directly pursuing language processing.

Language acquisition by machines has been attracting interest in various research fields, and several pioneering studies have developed algorithms based on inductive learning using sets of pairs, where each pair consists of a word sequence and nonlinguistic information about its meaning. In several studies, visual, rather than symbolic, information was given as nonlinguistic information (Dyer & Nenov 1993; Regier 1997). Spoken-word acquisition algorithms based on unsupervised clustering of speech tokens have also been described (Gorin et al. 1994; Nakagawa & Masukata 1995; Roy & Pentland 2002). Steels examined the socially interactive process of evolving grounded linguistic knowledge shared by communication agents from the viewpoint of game theory and a complex system (Steels 2003).

In contrast, the method (Iwahashi 2007; Iwahashi 2008) described in this paper, which is called LCore, satisfies the above-mentioned three requirements simultaneously and focuses on online learning of personally situated language use through verbal and nonverbal interaction with a user in the real physical world. LCore applies information from raw speech and visual observations and tactile reinforcement in an integrated way, and enables a robot to learn incrementally and online beliefs regarding speech units, words, concepts of objects, motions, grammar, and pragmatic and communicative capabilities.

A robot's belief system, encompassing these beliefs, is represented by a dynamic graphical model that has a structure reflecting the state of the user's belief system; therefore, learning makes it possible for the user and the robot to infer the state of each other's belief systems. Based on this belief system, LCore enables the robot to understand even fragmentary and ambiguous utterances of users, respond to them with confirmation questions and/or actions, generate directive utterances, and answer questions, appropriately for a given situation.

In particular, LCore enables the robot to learn these capabilities with relatively little interaction. This feature is also important because a typical user will not tolerate extended interaction with a robot that cannot communicate, and situations in actual everyday conversation change continuously.

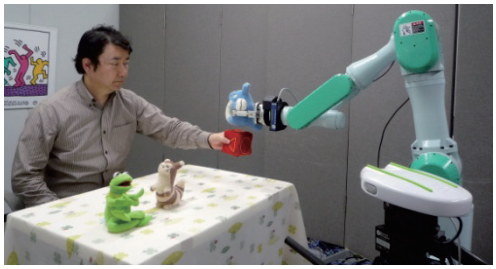


Figure 1: Robotic platform and its interaction with user.

Learning Setting

The spoken-language acquisition task discussed in this study was set up as follows: We performed experiments using the robotic platform shown in Fig. 1. The robot consisted of a manipulator with seven degrees of freedom (DOFs), a four-DOF multi-fingered grasper, a head unit, a directional microphone, a speaker, a stereo vision camera, and an infrared sensor. The head unit moved to indicate whether its gaze was directed at the user or at an object. A touch sensor was attached to the robot's hand for inputting a tactile reinforcement signal. A user and the robot looked at and moved the objects on the table shown as Fig. 1.

Initially, the robot did not possess any linguistic knowledge or concepts regarding the specific objects and the way in which they could be moved. First, to help the robot learn speech units, the user spoke for approximately one minute. Then, interactions were carried out to learn words, concepts of objects, motions, grammar, and pragmatic and communicative capabilities. Here, the communicative capability means that of understanding the type of the speech act of a user's utterance, and selecting an appropriate response from among moving an object, pointing at an object, and uttering an answer. The learning episodes for them could be carried out alternately, and were as follows:

Concepts of objects and words referring them: The user pointed to an object on the table while uttering a word describing it. The objects used included boxes, stuffed and wooden toys, and balls.

Motions and words referring them: The user moved an object while uttering a word describing the motion[1].

Grammar: The user moved an object while uttering a sentence describing the action, such as "*Place-on small frog green box*"[2]. Note that function words were not used because the learning method could not learn them.

Pragmatic capability: Using an utterance and a gesture, the user asked the robot to move an object, and the robot responded. If the robot responded incorrectly, the user slapped the robot's hand[3]. The robot also asked the user to move an object, and the user acted in response.

[1] The utterance is restricted to a word in each episode for learning an object and motion..

[2] Utterances made in Japanese have been translated into English in this paper.

[3] The physical retribution was used rather than an utterance, e.g. "no that is incorrect", because the robot has not learned such utterance.

Communicative capability: The user asked the robot to do something. If the robot did not respond correctly, the user indicated the correct response by moving or pointing at an objector, or uttered an answer.

Learning Method

Speech Units

Speech is a time-continuous one-dimensional signal. A robot learns statistical models of speech units from such signals without being provided with transcriptions of phoneme sequences or boundaries between phonemes. The difficulty of learning speech units is ascribed to the difficulties involved in speech segmentation, the clustering of speech segments into speech units, and the decision on the number of the speech units. For these difficulties, the methods based on unsupervised learning of speech unit HMMs were proposed (Iwahashi 2003; Iwahashi 2008).

Words

To learn words, static or moving images of objects and speech describing them are used. The lexicon consists of a set of lexical items, and each lexical item consists of statistical models of a spoken word and a concept. The statistical models of spoken words are represented by the concatenation of the speech units. The words are categorized into the following three types: 1) those that refer to perceptual characteristics of objects, such as *blue*, *big*, *apple,* and *Kermit*, 2) those that refer to abstract meanings in terms of objects, such as *tool* and *food*, and 3) those that refer to motions, such as *place on* and *move up*.

Perceptual Characteristics of Objects. In general, the difficulty of acquiring words that refer to objects can be ascribed to the difficulties involved in specifying features and extending them to other objects.

We proposed an interactive learning method that mainly addresses the problem of extension (Iwahashi 2008). The robot decides whether the input word is one in its vocabulary (a *known* word) or not (an *unknown* word). If the robot judges that the input word is an unknown word, it registers it in its vocabulary. This decision is made using both speech and visual information about the objects.

In addition, we proposed a learning method that allows sentence utterances, which consists of unknown words, instead of words as speech input (Taguchi *et al.* 2009). In the method, in order to obtain a lexicon, a statistical model of the joint probability of a spoken utterance and an object is learned based on the minimum description length principle. This model consists of a list of the phoneme sequences of words and three statistical models, namely, a sentence acoustic model, a word bigram model, and a word meaning model. By this method, words were learned with 84% phoneme accuracy.

The model for each image category is represented by a multidimensional Gaussian function in a twelve dimensional visual feature space (in terms of shape, color, and size), and it is learned using a Bayesian method every time an object image is given. Moreover, this concept is extended to multimodal representation including visual, tactile, and acoustic features based on a *bag of features* model (Nakamura *et al.* 2007) .

Figure 2: Scene example in which utterances were made and understood.

Abstract Meanings in Terms of Objects. Here, we consider words that refer to concepts that are more abstract and that are not formed directly from perceptual information, such as "*tool*," "*food*," and "*pet*." In a study on the abstract nature of the meanings of symbols (Savage-Rumbaugh 1986), it was found that chimpanzees could learn the lexigrams (graphically represented words) that refer to both individual object categories (e.g., "*banana*," "*apple*," "*hammer*," and "*key*") and the functions ("*tool*" and "*food*") of the objects.

A method that enables robots to gain this ability has been proposed (Nakamura *et al.* 2009). In this method, the movements that are given to the objects are taken as the objects' functions. Learning is based on the statistical model selection using the variational Bayes method.

Motions. While words that refer to objects are nominal, words that refer to motions are relational. The concept of the motion of a moving object is represented by a time-varying spatial relationship between a trajector and a landmark based on cognitive linguistics (Langacker 1991). The trajector is an entity characterized as the figure within a relational profile, and the landmark is entity characterized as the ground that provides a point of reference for locating the trajectory. Therefore, the concept of the trajectory of an object depends on perspective. In Fig. 2, for example, the trajectory of the stuffed toy on the left moved by the user, as indicated by the white arrow, is understood as *move over* and *place on* when the landmarks are considered to be the stuffed toy in the middle and the box on the right, respectively.

In general, however, information about what is a landmark is not obtained in learning data. The learning method must infer the landmark selected by a user in each scene. In addition, the type of intrinsic coordinate system in the space should also be inferred to appropriately represent the graphical model for each concept of motion. A method that can solve this problem is proposed in the References (Haoka & Iwahashi 2000; Sugiura & Iwahashi 2008). In this method, the concept of a motion is represented by a HMM. Landmarks and intrinsic coordinate systems are considered to be latent variables, and they are inferred using an expectation maximization

(EM) algorithm. By this method, each motion concept could be learned by showing about five different movements. The trajectory for the motion referred to by a motion word is generated by maximizing the output probability of the learned HMM, given the positions of a trajector and a landmark.

Grammar

To learn grammar, moving images of actions and speech describing them are used. The robot should detect the correspondence between a semantic structure in the moving image and a syntactic structure in the speech. However, such semantic and syntactic structures are not observable. While an enormous number of structures can be extracted from a moving image and speech, the method should select those with the most appropriate correspondence between them. Grammar is statistically learned using such correspondences, and is then inversely used to extract the correspondence.

It is assumed that each utterance is generated on the basis of stochastic grammar, based on a conceptual structure. The conceptual structure used here is a basic schema that is applied in cognitive linguistics. The word sequence of utterance $\boldsymbol{s}$ is interpreted as a conceptual structure $\mathbf{z} = [(z_1, W_{z_1}), (z_2, W_{z_2}), (z_3, W_{z_3})]$, where z_i represents the attribute of a phrase and has a value among {M, T, L}. W_M, W_T, W_L represent the phrases describing a motion, a trajector, and a landmark, respectively. For example, when the image is the same as that shown in Fig. 2, and the corresponding utterance is "*Place-on small frog brown box*," then the utterance is interpreted as follows: [(M, place-on), (T, small frog), (L, brown box)].

The grammar is a statistical language model that is represented by a set of occurrence probabilities for the possible orders of attributes in the conceptual structure, and is learned using Bayesian method.

Pragmatic Capability

The meanings of utterances are conveyed based on certain beliefs shared by those communicating in the situations. When a participant interprets an utterance based on his/her assumptions that certain beliefs are shared and is convinced, based on certain clues, that the interpretation is correct, he/she gains confidence that the beliefs are shared. On the other hand, because the sets of beliefs assumed to be shared by participants actually often contain discrepancies, the more beliefs a listener needs to understand an utterance, the greater is the risk that the listener will misunderstand it.

As mentioned above, a pragmatic capability of the robot relies on the capability to infer the state of a user's belief system. Therefore, the method should enable the robot to adapt its assumption of shared beliefs rapidly and robustly through verbal and nonverbal interaction. The method should also control the balance between (i) the transmission of the meanings of utterances and (ii) the transmission of information about the state of belief systems in the process of generating utterances.

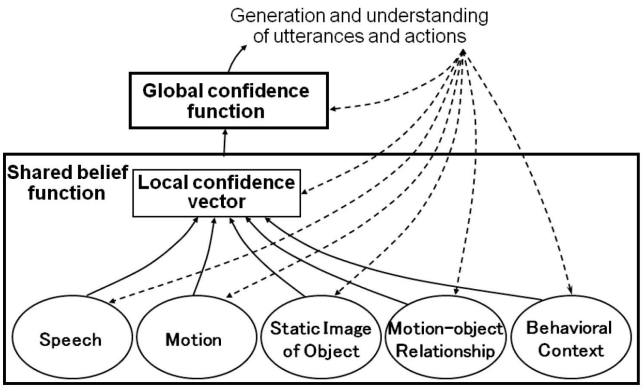


Figure 3: Belief system of robot for pragmatic capability.

The following is an example of generating and understanding utterances based on the assumption of shared beliefs. Suppose that in the scene shown in Fig. 2, the frog on the left has just been put on a table. If the user in the figure wants to ask the robot to place a frog on the box, he may say, "*Place-on frog box*." In this situation, if the user assumes that the robot shares the belief that the object moved in the previous action is likely to be the next target for movement and the belief that the box is likely to be something for the object to be placed on, he might just say "*Place-on* [1]." To understand this fragmentary and ambiguous utterance, the robot must possess similar beliefs. If the user knows that the robot has responded by doing what he asked it to, this knowledge would strengthen his confidence that the beliefs he assumed to be shared are really shared. It can be understood that the former utterance is more effective than the latter utterance in transmitting the meaning of the utterance, while the latter utterance is more effective than the former utterance in transmitting information about the state of belief systems. Conversely, when the robot wants to ask the user to do something, the beliefs that it assumes to be shared are used in the same way.

We have proposed the method which copes with the above difficulty (Iwahashi 2003; Iwahashi 2007). In the method, robot's belief system has a structure that reflects the state of the user's belief system so that the user and the robot infer the state of each other's belief systems (Fig. 3). This structure consists of the following two parts:

1) **Shared belief function (SBF)**, which models the assumption of shared beliefs and consists of a set of belief modules with values (the local confidence vector) that represent the degree of confidence that each belief is shared by the robot and the user. The beliefs used are those that concern speech, motions, static images of objects, behavioral context, and motion-object relationships. The output of this function is the sum of the outputs of all belief modules weighted by the local

[1] Although use of a pronoun might be more natural than deletion of noun phrases in some languages, the same ambiguity in meaning exists in both such expressions.

confidence vector, and it represents the degree of correspondence between an utterance and an action.

2) **Global confidence function (GCF),** which represents the estimation of the difference between the shared beliefs assumed by the robot, i.e., SBF, and the shared beliefs assumed by the user. GCF is represented by a sigmoid function, and it outputs an estimate of the probability that the speaker's utterance will be correctly understood by the listener.

Through the interactive episodes, the robot can incrementally learn all the parameters of this belief system online. The parallel structure of SBF with minimum classification error and Bayesian learning methods can make the learning of the belief system fast. Using these functions, the robot can understand the user's fragmental and ambiguous directive utterances, respond to them by acting, and generate confirmation questions, if necessary.

In addition, a method that detects users' robot-directed speech has been proposed based on pragmatic capability (Zuo *et al.* 2010).

Utterance understanding by robot. Given an utterance, the optimum action is that which maximizes the output of SBF. When the user utters directives to get the robot to move an object, the robot can move in response even if the utterances are fragmental and ambiguous. In experiments, approximately 80% in correct understanding rate was achieved for such fragmental utterances through about ninety episodes. An example of action generated as a result of a correct understanding of the user's utterance "*Place-on*" is shown in Fig. 4, along with the second action candidates. Output log-probabilities obtained from the belief modules weighted by the local confidence vector are also shown. It was found that nonlinguistic beliefs were used appropriately in understanding the utterance based on its relevance to the situation.

Utterance generation by robot. Because GCF outputs the estimate of the probability that a speaker's utterance will be correctly understood by a listener, the robot can control the ambiguity of its utterances. Hence, the robot can facilitate the formation of shared beliefs between a user and itself by adjusting the risk of being misunderstood in order to enable the user and the robot to infer each other's inner state (Nakamura *et al.* 2009).

In addition, the robot can generate confirmation utterances if it decides that the user's utterances are too ambiguous to execute an action immediately. This decision-making is based on expected utility (Sugiura *et al.* 2009). The optimum response can be selected based on the threshold θ of the output of the GCF of the optimum action.

An example of executed dialogue is shown in Fig. 5. Because the output of the GCF of the optimum action was less than θ, a confirmation utterance was the optimum response. Therefore, the robot first asked whether "*green box*" was the trajector. Here, the word "*green*" was used to describe the major difference between Object 2 (green box) and Object 3 (blue box). In the second confirmation utterance, the word "*blue*" was inserted into the segment

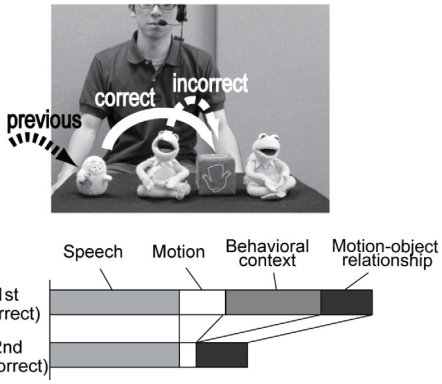
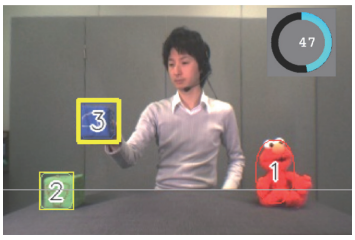


Figure 4: An example of action generated as a result of correct understanding of utterance "*Place-on*" and weighted output log-probability from belief modules, along with second choices.



[Situation: Object 2 was manipulated most recently]
U: "*Move-closer box Elmo.*"
R: "*Move-closer green box?*"
U: "*No.*"
R: "*Move-closer blue box?*"
U: "*Yes.*"
R: (The robot moves Object 3 closer to Object 1.)

Figure 5: Dialog example. Motion executed with confirmation utterances. The correct action is to move Object 3 (blue box) closer to Object 1 (red stuffed toy).

W_T. In contrast, the landmark was not mentioned in either generated utterance, because no word insertion to W_L had a large influence on the GCF output.

Communicative Capability

In previous subsections, we have described the processing of directive utterances. So that the robot can communicate with users more naturally, however, it should classify users' utterances into one of the types of speech acts in order to respond to them appropriately. The proposed learning method (Taguchi *et al.* 2009) enables the robot to return suitable utterances to a human or to perform actions by learning the meanings of interrogative words, such as "what" and "which." The meanings of these words are grounded in communication and stimulate specific responses from a listener. The method learns the relationship between the user utterances and the robot responses. Hence, the robot can select any of the following responses to users' utterances: moving an object, pointing at an object, or answering by an utterance. For example, a

user asks "*What place-on?*" after he has placed an apple on a box. In this case, the robot can answer "*Apple.*"

Conclusion and Future Work

This paper described our developmental approach toward personally and physically situated human-robot conversations. The communication learning method LCore based on the approach satisfies three major requirements that existing language processing methods cannot, namely, *grounding*, *scalability*, and *sharing*.

Finally, we suggest solving two challenging problems for future work. The first is bridging the gap between nonlinguistic and linguistic computational processes in communication, which at present are completely different. A key to solving this problem could be role reversal imitation (Carpenter *et al.* 2005; Taniguchi *et al.* 2010), which is a basis of communication learning. The second problem is to enable learning of the displaced language in addition to language grounded in the physical world. To solve this problem, developing a computational mechanism for metaphors is essential.

References

Allen, J. et al. 2001. Toward Conversational Human-Computer Interaction. *AI Magazine* 22(4): 27–38.

Carpenter, M. *et al.* 2005. Role Reversal Imitation and Language in Typically Developing Infants and Children with Autism. *Infancy* 8(3): 253–278.

Dyer, M. G.; and Nenov, V. I. 1993. Learning Language via Perceptual/Motor Experiences. *Proc. Annual Conf. of Cog. Sci. Society*, 400–405.

Gorin, A. *et al.* 1994. An Experiment in Spoken Language Acquisition. *IEEE Trans. Speech and Audio Processing* 2(1): 224–240.

Haoka, T.; and Iwahashi, N. 2000. Learning of the Reference-Point-Dependent Concepts on Movement for Language Acquisition. *IEICE Tech. Rep. PRMU2000-105*, 39–45. (in Japanese)

Iwahashi, N. 2003. Language Acquisition Through a Human-Robot Interface by Combining Speech, Visual, and Behavioral Information. *Information Sciences* 156: 109–121.

Iwahashi, N. 2007. Robots That Learn Language: A Developmental Approach to Situated Human-Robot Conversations. In *Human-Robot Interaction,* Nilanjan Sankar, (Ed), I-Tech Education and Publishing, 95–118.

Iwahashi, N. 2008. Interactive Learning of Spoken Words and Their Meanings Through an Audio-Visual Interface. *IEICE Trans. Information and Systems*, E91D(2): 312–321.

Langacker, R. 1991. *Foundation of Cognitive Grammar.* Stanford University Press, CA.

Nakagawa, S.; and Masukata, M. 1995. An Acquisition System of Concept and Grammar Based on Combining Visual and Auditory Information. *Trans. Information Society of Japan* 10(4): 129–137.

Nakamura, T. *et al.* 2007. Multimodal Object Categorization by a Robot. *Proc. Int. Conf. Intelligent Robots and Systems*, 2415–2420.

Nakamura, S. *et al.* 2009. Learning of Abstract Concepts and Words Based on Model Structure Selection Using Variational Bayes, *IEICE Trans. Information and Systems* . J92D(4): 467–479. (in Japanese)

Nakamura, S. *et al.* 2009. Mutually Adaptive Generation of Utterances Based on the Inference of Beliefs Shared by Humans and Robots. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics* 12(5): 37–56. (in Japanese)

Regier, T. 1997. *Human Semantic Potential.* MIT Press.

Roy, D. 2005. Semiotic Schemas: A Framework for Grounding Language in Action and Perception. *Artificial Intelligence* 167(1): 170-205.

Roy, D.; and Pentland, A. 2002. Learning Words from Sights and Sounds: A Computational Model. *Cognitive Science* 26(1): 113–146.

Savage-Rumbaugh, E. S. 1986. *Ape Language – From Conditional Response to Symbol*. Columbia Univ. Press.

Shapiro, C. S.; et al. 2000. Our Dinner with Cassie. *Proc. AAAI Symposium on Natural Dialogues with Practical Robotic Devices*, 57–61.

Sperber, D.; and Wilson, D. 1995. *Relevance (2nd Edition)*, Blackwell.

Steels, L. 2003. Evolving Grounded Communication for Robots. *Trends in Cognitive Science* 7(7): 308–312.

Sugiura, K.; and Iwahashi, N. 2008. Motion Recognition and Generation by Combining Reference-Point-Dependent Probabilistic Models. Proc. Int. Conf. Intelligent Robots and Systems, 852–857.

Sugiura, K. *et al.* 2009. Bayesian Learning of Confidence Measure Function for Generation of Utterances and Motions in Object Manipulation Dialogue Task. *Proc. Interspeech*, 2483-2486.

Taguchi, R. *et al.* 2009. Learning Communicative Meanings of Utterances by Robots. In Hattori, H., et al., (Eds), *New Frontiers in Artificial Intelligence*, LNCS/ LNAI 5447, Springer, 62–72.

Taguchi, R. *et al.* 2009. Learning Lexicons from Spoken Utterances Based on Statistical Model Selection. *Proc. Interspeech*, 2731–2734.

Taniguchi, T. *et al.* 2010. Simultaneous Estimation of Role and Response Strategy in Human-Robot Role-Reversal Imitation Learning. *Proc. Sym. Analysis, Design, and Evaluation of Human-Machine Systems (to appear).*

Winograd, T. 1972. *Understanding Natural Language.* Academic Press New York.

Zuo, X. *et al.* 2010. Robot-Directed Speech Detection Using Multimodal Semantic Confidence Based on Speech, Image, and Motion. *Proc. Sym. Robot and Human Interactive Communication (to appear).*