A Method for Forming Mutual Beliefs for Communication through Human-robot Multi-modal Interaction

Naoto Iwahashi
Sony Computer Science Labs.
Tokyo, Japan
iwahashi@csl.sony.co.jp

Abstract

This paper describes a method of multi-modal language processing that reflects experiences shared by people and robots. Through incremental online optimization in the process of interaction, the user and the robot form mutual beliefs represented by a stochastic model. Based on these mutual beliefs, the robot can interpret even fragmental and ambiguous utterances, and can act and generate utterances appropriate for a given situation.

1 Introduction

The process of human communication is based on certain beliefs shared by those who are communicating. Language is one such mutual belief, and it is used to convey meaning based on its relevance to other mutual beliefs (Sperber and Wilson, 1995). These mutual beliefs are formed through interaction with the environment and other people, and the meaning of utterances is embedded in such shared experiences.

If those communicating want to logically convince each other that proposition p is a mutual belief, they must prove that the infinitely nested proposition "They have information that they have information that . . . that they have information that p" also holds. However, in reality, all we can do is assume, based on a few clues, that our beliefs are identical to those of other people we are talking to. That is, it can never be guaranteed that our beliefs are identical to those of other people.

The processes of utterance generation and understanding rely on a system of mutual beliefs assumed by each person, and this system changes autonomously and recursively through these processes. The listener interprets utterances based on their relevance to their system of assumed mutual beliefs. The listener also receives information for updating their system of assumed mutual beliefs through this process. In addition, the speaker can receive similar information through the response of the listener. Through utterances, people simultaneously send and receive information about one another's system of assumed mutual beliefs. In this sense, we can say that a mutual belief system assumed by one person couples with mutual belief systems assumed by other people they are communicating with.

To enable humans and robots to communicate with one another in a physical environment the way people do, spoken-language processing methods must emphasize mutual understanding, and they must have a mechanism that would enable the mutual belief systems couple with one another. Moreover, language, perception, and behavior have to be processed integratively in order for humans and robots to physically share their environment as the basis for the formation of common experiences and in order for linguistic and nonlinguistic beliefs to combine in the process of human-robot interaction. Previous language processing methods, which are characterized by fixed language knowledge, do not satisfy these requirements because they cannot dynamically reflect experiences in the communication process in a real environment.

I have been working on methods for forming linguistic beliefs, such as beliefs about phonemes, lexicon, and grammar, based on common perceptual ex-

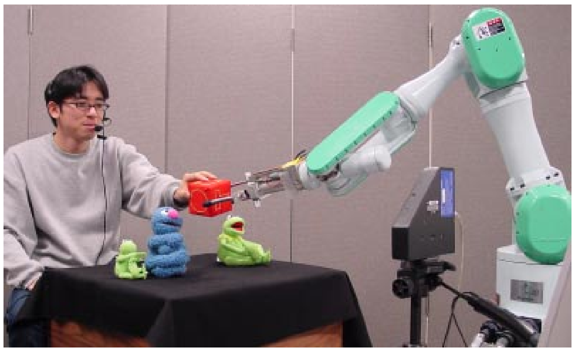

Figure 1: Interaction between a user and a robot

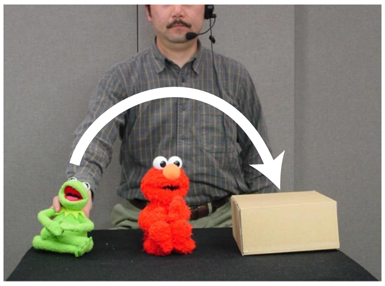

Figure 2: A scene during which utterances were made and understood

periences between people and robots (for further detail, see (Iwahashi, 2001; Iwahashi, 2003)). This paper describes a method that enables robots to learn a system of mutual beliefs including nonlinguistic ones through multi-modal language interaction with people. The learning is based on incremental online optimization, and it uses information from raw speech and visual observations as well as behavioral reinforcement, which is integrated in a probabilistic framework.

Theoretical research (Clark, 1996) and its computational modelling (Traum, 1994) focused on the formation of mutual beliefs that is a direct target of communication, and aimed at representing the formation of mutual beliefs as a procedure- and rule-driven process. In contrast, this study focuses on a system of mutual beliefs that is used in the process of utterance generation and understanding in a physical environment, and aims at representing the formation of this system by a mathematical model of coupling systems.

2 Task for Forming Mutual Beliefs

The task for forming mutual beliefs was set up as follows. A robot was sat at a table so that the robot and the user sitting at the table could see and move the objects on the table (Fig. 1) The user and the robot initially shared certain basic linguistic beliefs, including a lexicon with a small number of items and a simple grammar, and the robot could understand some utterances [1]. The user asked the robot to move an object by making an utterance and a gesture, and the robot acted in response. If the robot responded incorrectly, the user slapped the robot's hand. The robot also asked the user to move an object, and the user acted in response. Mutual beliefs were formed incrementally, online, through such interaction.

Figure 2 shows an example of utterance generation and understanding using mutual beliefs. In the scene shown in Fig. 2, the object on the left, *Kermit*, has just been put on the table.

If the user in the figure wants to ask the robot to move Kermit onto the box, he may say *"Kermit box move-onto"*. 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 moved onto, he might just say *"move-onto"*. To understand this fragmental utterance, the robot has to have similar beliefs. Inversely, when the robot wants to ask the user to do something, mutual beliefs are used in the same way.

3 Algorithm for Learning a System of Mutual Beliefs

3.1 Setting

The robot has an arm with a hand, and a stereo camera unit. A close-talk microphone is used for speech input, and the speech is represented by a time-sequence of Mel-scale cepstrum coefficients. Visual observation o_i of object i is represented by using such features as color (three-dimensional: L*a*b* parameters), size (one-dimensional), and shape (two-dimensional). Trajectory u of the object's motion is represented by a time-sequence of its positions. A touch sensor is attached to the robot's hand.

[1] No function words were included in the lexicon.

3.2 Representation of a system of mutual beliefs

In the algorithm I developed, the system of mutual beliefs consists of two parts: 1) a decision function, composed of a set of beliefs with values representing the degree of confidence that each belief is shared by the robot and the user, and 2) a global confidence function, which represents the degree of confidence for the decision function. The beliefs I used are those concerning lexicon, grammar, behavioral context, and motion-object relationship. The degree of confidence for each belief is represented by a scalar value. The beliefs are represented by stochastic models as follows:

Lexicon L

The lexicon is represented by a set of pairs, each with probability density function (pdf) $p(s|c_i)$ of feature s of a spoken word and a pdf for representing the image concept of lexical item c_i, $i = 1, \ldots, N$. Two types of image concepts are used. One is a concept for the static observation of an object. Conditional pdf $p(o|c)$ of feature o of an object given lexical item c is represented by a Gaussian pdf. The other is a concept for the movement of an object. The concept is viewed as the process of change in the relationship between the trajector and the landmark. Here, given lexical item c, position $o_{t,p}$ of trajector object t, and position $o_{l,p}$ of landmark object l, conditional pdf $p(u|o_{t,p}, o_{l,p}, c)$ of trajectory u is represented by a hidden Markov Model (HMM). Pdf $p(s|c)$ of the features of a spoken word is also represented by an HMM.

Grammar G

I assume that an utterance can be understood based on its conceptual structure z, which has three attributes: *landmark*, *trajector*, and *motion*, each of which contains certain elements of the utterance. Grammar G is represented by a set of occurrence probabilities for the orders of these attributes in an utterance and a statistical bigram model of the lexical items for each of the three attributes.

Effect of behavioral context $B_1(i, q; H)$

Effect of behavioral context represents the belief that the current utterance refers to object i, given behavioral context q. q includes information on whether object i was a trajector or a landmark in the previous action and whether the user's current gesture is referring to object i. This belief is represented by a parameter set $H = \{h_c, h_g, h_p\}$, and it takes h_c as its value if object i was involved in the previous action, h_g if the object is being held, h_p if it is being pointed to, and 0 in all other cases.

Motion-object relationship $B_2(o_{t,f}, o_{l,f}, W_M; R)$

Motion-object relationship represents the belief that in the motion corresponing to lexical item W_M, feature $o_{t,f}$ of object t and feature $o_{l,f}$ of object l are typical for a trajector and a landmark, respectively. This belief is represented by a conditional multivariate Gaussian pdf, $p(o_{t,f}, o_{l,f}|W_M; R)$, where R is its parameter set.

3.3 Decision function

The beliefs described above are organized and assigned confidence values to obtain the decision function used in the process of utterance generation and understanding. This decision function is written as

$$\begin{aligned}
&\Psi(s, a, O, q, L, G, R, H, \Gamma)\\
&= \max_{l,z}\Big(\gamma_1 \log p(s|z;\ L, G) \qquad \text{[Speech]}\\
&\qquad +\gamma_2 \log p(u|o_{t,p}, o_{l,p}, W_M;\ L) \qquad \text{[Motion]}\\
&\qquad +\gamma_2\Big(\log p(o_{t,f}|W_T;\ \ L) + \log p(o_{l,f}|W_L;\ \ L)\Big)\\
&\qquad\qquad \text{[Object]}\\
&\qquad +\gamma_3 \log p(o_{t,f}, o_{l,f}, |W_M;\ R)\\
&\qquad\qquad \text{[Motion-Object Relationship]}\\
&\qquad +\gamma_4\Big(B_1(t, q;\ \ H) + B_1(l, q;\ H)\Big)\Big)\\
&\qquad\qquad \text{[Behavioral Context]}
\end{aligned}$$

where $\Gamma = \{\gamma_1, \ldots, \gamma_4\}$ is a set of confidence values for beliefs corresponding to the speech, motion, object, motion-object relationship, and behavioral context; a denotes the action, and it is represented by a pair (t, u) of trajector object t and trajectory u of its movement; O denotes the scene, which includes the positions and features of all the objects in the scene; and W_T and W_L denotes the sequences of the lexical items in the utterances for the trajector and landmark, respectively. Given O, q, L, G, R, H, and Γ, the corresponding action, $\tilde{a} = (\tilde{u}, \tilde{t})$, understood to be the meaning of utterance s is determined by maximizing the decision function as

$$\tilde{a} \quad = \quad \arg\max_a \Psi(s, a, O, q, L, G, R, H, \Gamma).$$

3.4 Global confidence function

Global confidence function f outputs an estimate of the probability that the robot's utterance s will be correctly understood by the user, and it is written as

$$f(x) = \frac{1}{\pi} \arctan\left(\frac{x - \lambda_1}{\lambda_2}\right) + 0.5,$$

where λ_1 and λ_2 are the parameters of this function. A margin in the value of the output of the decision function in the process of generating an utterance is used for input x of this function. Margin d obtained in the process of generating utterance s that means action a in scene O under behavioral context q is defined as

$$\begin{aligned} & d(s, a, O, q, L, G, R, H, \Gamma) \\ & \quad = \min_{A \neq a} \Big(\Psi(s, a, O, q, L, G, R, H, \Gamma) \\ & \qquad\qquad - \Psi(s, A, O, q, L, G, R, H, \Gamma)\Big). \end{aligned}$$

We can easily see that a large margin increases the probability of the robot being understood correctly by the user. If there is a high probability of the robot's utterances being understood correctly even when the margin is small, we can say that the robot's beliefs are consistent with those of the user. When the robot asks for action a in scene O under behavioral context q, the robot generates utterance $\tilde{s}$ so as to make the value of the output of f as close as possible to value of parameter ξ, which represents the taget probability of the robot's utterance being understood correctly. This utterance can be represented as

$$\tilde{s} = \arg\min_s \Big(f(d(s, a, O, q, L, G, R, H, \Gamma)) - \xi\Big).$$

The robot can increase the chance of being understood correctly by using more words. On the other hand, if the robot can predict correct understanding with a sufficiently high probability, the robot can manage with a fragmental utterance using a small number of words.

3.5 Learning

The decision function and the global confidence function are learned separately in the utterance understanding and utterance generation processes, respectively.

The decision function is learned incrementally, online, through a sequence of episodes each of which consists of the following steps. 1) Through an utterance and a gesture, the user asks the robot to move an object. 2) The robot acts on its understanding of the utterance. 3) If the robot acts correctly, the process is terminated. Otherwise, the user slaps the robot's hand. 4) The robot acts in a different way. 5) If the robot acts incorrectly, the user slaps the robot's hand.

When the robot acts correctly in the first or second trial in an episode, the robot associates utterance s, action a, scene O, and behavioral context q with each other, and makes these associations a learning sample. Then the robot adapts the values of parameter set R for the belief about the motion-object relationship, parameter set H for the belief about the effect of the behavioral context, and a set of weighting parameters, Γ. R is learned by using the Bayesian learning method. H and Γ are learned based on the minimum error criterion (Juang and Katagiri, 1992). Lexicon L and grammar G are given beforehand and are fixed. When the ith sample (s_i, a_i, O_i, q_i) is obtained based on this process of association, H_i and Γ_i are adapted to minimize the probability of misunderstanding based on the minimum error criterion as

$$\sum_{j=i-K}^{i} w_{i-j}\, g(d\,(s_j, a_j, O_j, q_j, L, G, R_i, H_i, \Gamma_i)) \\ \rightarrow \min,$$

where $g(x)$ is $-x$ if $x < 0$ and 0 otherwise, and K and w_{i-j} represent the number of latest samples used in the learning process and the weights for each sample, respectively.

Global confidence function f is learned incrementally, online, through a sequence of episodes which consist of the following steps. 1) The robot generates an utterance to ask the user to move an object. 2) The user acts according to their understanding of the robot's utterance. 3) The robot determines whether the user's action is correct or not.

In each episode, the robot generates an utterance that makes the value of the output of global confidence function f as close to ξ as possible. After each episode, the value of margin d in the utterance generation process is associated with information about whether the utterance was understood correctly or not, and this sample of associations is used for learning. The learning is done so as to approximate the probability that an utterance will be understood correctly by minimizing the weighted sum of squared errors in the latest episodes. After the ith episode, parameters λ_1 and λ_2 are adapted as

$$[\lambda_{1,i}, \lambda_{2,i}] \leftarrow (1-\delta)[\lambda_{1,i-1}, \lambda_{2,i-1}] + \delta[\tilde{\lambda}_{1,i}, \tilde{\lambda}_{2,i}],$$

where

$$\begin{aligned}&(\tilde{\lambda}_{1,i}, \tilde{\lambda}_{2,i})\\ &\quad = \arg\min_{\lambda_1,\lambda_2} \sum_{j=i-K}^{i} w_{i-j}(f(d_j;\ \ \lambda_1,\lambda_2) - e_j)^2,\end{aligned}$$

where e_i is 1 if the user's understanding is correct and 0 if it is not, and δ is the value that determines the learning speed.

4 Experiments

4.1 Conditions

The lexicon used in the experiments included eleven items for the static image and six items to describe the motions. Eleven stuffed toys and four boxes were used as objects in the interaction between a user and a robot. In the learning process, the interaction was simulated by using software.

4.2 Learning of decision function

Sequence X of quadruplets (s_i, a_i, O_i, q_i) consisting of the user's utterance s_i, scene O_i, behavioral context q_i, and action a_i that the user wanted to ask the robot to perform ($i = 1, \ldots, n_d$), was used for the interaction. At the beginning of the sequence, the sentences were relatively complete (e.g., "*green kermit red box move-onto*"). Then the length of the sentences was gradually reduced (e.g., "*move-onto*").

R could be estimated with high accuracy during the episodes in which relatively complete utterances were understood correctly. H and Γ could be effectively estimated based on the estimation of R when fragmental utterances were given. Figure 3 shows changes in the values of γ_1, γ_2, γ_3, and $\gamma_4 h_c$. We can see that each value was adapted according to the ambiguity of a given sentence. Figure 4 shows the decision error (misunderstanding) rates obtained during the course of the interaction, along with the error rates obtained in the same data, X, by keeping the values of the parameters of the decision function fixed to their initial values.

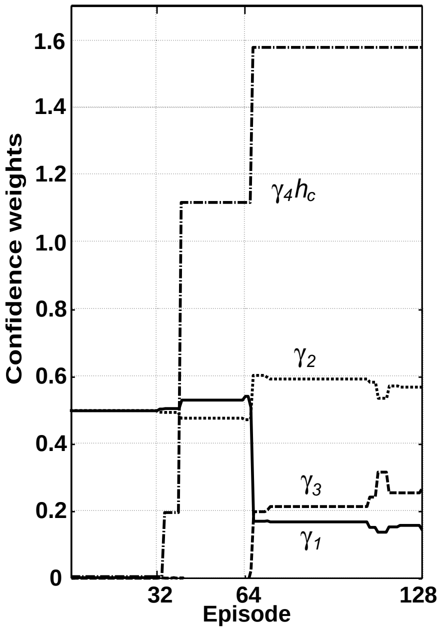

Figure 3: Changes in the confidence values

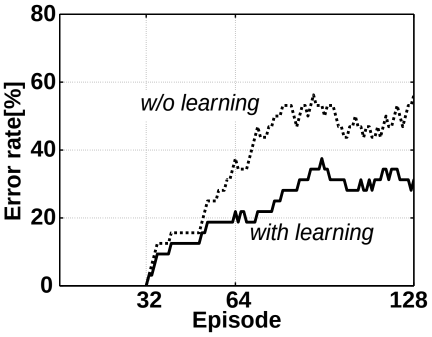

Figure 4: The change in decision error rate

Examples of actions generated as a result of correct understanding are shown together with the output log probabilities from the weighted beliefs in Figs. 5 (a) and (b), along with the second and third action candidates, which led to incorrect actions. We can see that each nonlinguistic belief was used appropriately in understanding the utterances. Beliefs

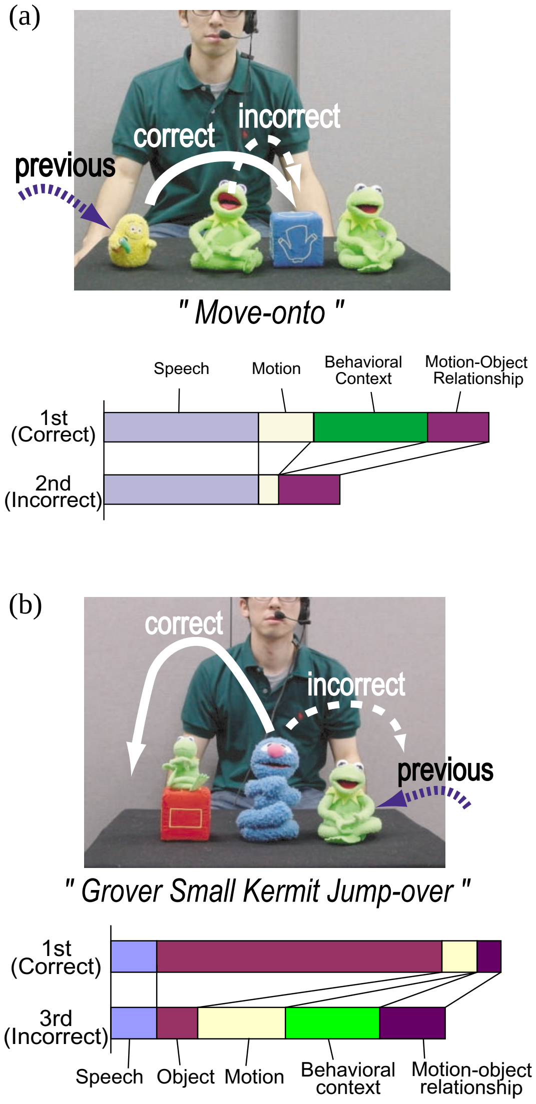

Figure 5: Examples of fragmental utterances understood correctly by the robot

about the behavioral context were more effective in Fig. 5 (a), while in Fig. 5 (b), beliefs about the object concepts were more effective than other nonlinguistic beliefs in leading to the correct understanding. This learning process is described in greater detail in (Miyata et al., 2001)

4.3 Learning of the global confidence function

In the experiments with the learning of the global confidence function, The robot's utterances were expressed through text on a display instead of oral speech, and they included one word describing the motion and either no words or one to several words describing the trajector and landmark objects or just the trajector object.

A sequence of triplets (a, O, q) consisting of scene O, behavioral context q, and action a that the robot needed to ask the user to perform, was used for the interaction. In each episode, the robot generated an utterance to bring the global confidence function as close to 0.75 as possible.

The changes in $f(d)$ are shown in Fig. 5 (a), where three lines are drawn for $d_{0.5} = f^{-1}(0.5)$, $d_{0.75} = f^{-1}(0.75)$, and $d_{0.9} = f^{-1}(0.9)$ in order to make the shape of $f(d)$ easily recognizable. The changes in the number of words used to describe the objects in each utterance are shown in Fig. 5 (b), along with the changes obtained in the case when $f(d)$ was not learned, which are shown for comparison. The initial values were set at $d_{0.9} = 161$, $d_{0.75} = 120$, and $d_{0.5} = 100$, which means that a large margin was necessary for an utterance to be understood correctly. Note that when all the values are close to 0, the slope in the middle of f is steep, and the robot makes a decision that a small margin is enough for its utterances to be understood correctly. After the learning began, these values rapidly approached 0, and the number of words decreased. The slope became temporarily smooth at around the 15th episode. Then, the number of words became too small, which sometimes lead to misunderstanding. Finally, the slope became steep again at around the 35th episode.

5 Discussion

The above experiments illustrate the importance of misunderstanding and clarification, *i.e.* error and repair, in the formation of mutual beliefs between people and machines. In the learning period for utterance understanding by the robot, the values of the parameters of the decision function changed significantly when the robot acted incorrectly in the first trial, and correctly in the second trial. In the learning period for utterance generation by the robot, in the experiment in which the target value of the global confidence function was set to 0.95, which was larger than 0.75 and closer to 1, the global confidence function was not properly estimated because almost all utterances were understood correctly (The results of this experiment are not presented in de-

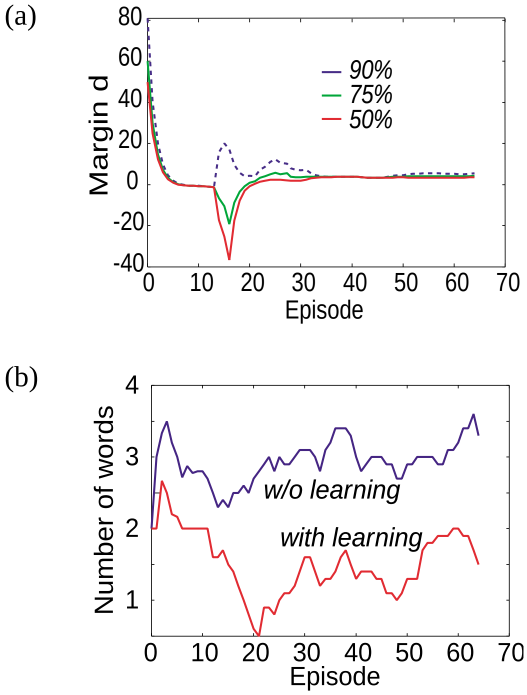

Figure 6: Changes in the global confidence function (a) and the number of words needed to describe the objects in each utterance (b)

tail in this paper). These results show that occasional errors enhance the formation of mutual beliefs in both the utterance generation and utterance understanding processes. This implies that in order to obtain information about mutual beliefs, both the robot and the user must face the risk of not being understood correctly. The importance of error and repair to learning in general has been seen as an exploration-exploitation trade-off in the area of reinforcement learning by machines (e.g. (Dayan and Sejnowski, 1996)).

The experimental results showed that the robot could learn its system of the beliefs the robot assumed the user had. Because the user came to understand the robot's fragmental and ambiguous utterances, the user and the robot must have shared similar beliefs, and must have been aware of that. It would be interesting to investigate by experiment the dynamics of sharing beliefs between a user and a robot.

6 Related Works

(Winograd, 1972) and (Shapiro et al., 2000) explored the grounding of the meanings of utterances in conversation onto the physical world by using logic, but the researchers did not investigate the processing of information from the real physical world. (Matsui et al., 2000) focused on enabling robots to work in the real world, and integrated language with information from robot's sensors by using pattern recognition. (Inamura et al., 2000) investigated an autonomous mobile robot that controlled its actions and conversations with a user based on a Bayesian network. The use of Bayesian networks in the interpretation and generation of dialogue was also investigated by (Lemon et al., 2002). In (Singh et al., 2000), the learning of dialogue strategies using reinforcement learning was investigated. Some of these works looked at beliefs ”held by” the machines themselves, but none focused on the formation of mutual beliefs between humans and machines through interaction, based on common experiences.

7 Conclusion

The presented method enables the formation of mutual beliefs between people and robots through interaction in physical environments, and it facilitates the process of human-machine communication. In the future, I want to focus on the generalization of learning of mutual beliefs and the learning of dialogue control.

References

H. Clark. 1996. *Using Language*. Cambridge University Press.

P. Dayan and T. J. Sejnowski. 1996. Exploration Bonuses and Dual Control. *Machine Learning*, 25:5–22.

T. Inamura, M. Inaba and H. Inoue. 2000. Integration Model of Learning Mechanism and Dialogue Strategy based on Stochastic Experience Representation using Bayesian Network. *Proceedings of International Workshop on Robot and Human Interactive Communication*, 27–29.

N. Iwahashi. 2001. Language Acquisition by Robots. *The Institute of Electronics, Information, and Communication Engineers Technical Report* SP2001-96.

N. Iwahashi. 2003. Language Acquisition by Robots: Towards New Paradigm of Language Processing. *Journal of Japanese Society for Artificial Intelligence*, 18(1):49-58.

B.-H. Juang and S. Katagiri. 1992. Discriminative Learning for Minimum Error Classification. *IEEE Transactions on Signal Processing*, 40(12):3043–3054.

O. Lemon, P. Parikh and S. Peters. 2002. Probabilistic Dialogue Management. *Proceedings of Third SIGdial Workshop on Discourse and Dialogue*, 125–128.

T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T Kurita and N. Otsu. 1999. Integrated Natural Spoken Dialogue System of Jijo-2 Mobile Robot for Office Services. *Proceedings of 15th National Conference on Artificial Intelligence*.

A. Miyata, N. Iwahashi and A. Kurematsu. 2001. Mutual belief forming by robots based on the process of utterance comprehension. *Technical Report of The Institute of Electronics, Information, and Communication Engineers*, SP2001-98.

C. S. Shapiro, H. O. Ismail, and J. F. Santore. 2000. Our Dinner with Cassie. *Working Notes for AAAI 2000 Spring Symposium on Natural Dialogues with Practical Robotic Devices*, 57-61.

S. Singh, M. Kearns, D. J. Litman and M. A. Malker. 2000. Empirical Evaluation of a Reinforce Learning Spoken Dialogue System. *Proc. 16th National Conference on Artificial Intelligence*, 645–651.

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

D. R. Traum. 1994. *A computational theory of grounding in natural language conversation*. Unpublished doctoral dissertation, University of Rochester.

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