The final publication is available at
http://dl.acm.org/citation.cfm?id=2974814

Alignment Approach Comparison between Implicit and Explicit Suggestions in Object Reference Conversations

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

The recognition of an indicated object by an interacting person is an essential function for a robot that acts in daily environments. To improve recognition accuracy, clarifying the goal of the indicating behaviors is needed. For this purpose, we experimentally compared two kinds of interaction strategies: a robot that explicitly provides instructions to people about how to refer to objects or a robot that implicitly aligns with the people's indicating behaviors. Even though our results showed that participants evaluated the implicit approach to be more natural than the explicit approach, the recognition performances of the two approaches were not significantly different.

Author Keywords

Human-robot interaction; object recognition; alignment

ACM Classification Keywords

H.5.2. User Interfaces – *Interaction styles*.

INTRODUCTION

Social robots in human society need to recognize the objects indicated by users (Figure 1). Various approaches have been proposed to recognize such objects based on user utterances and pointing gestures [1,2,3,4]. Nickel et al. used the 3D positions of a head and hands as well as the head's orientation to recognize pointing gestures in object references [1]. Schauerte et al. integrated speech and pointing gesture recognition by image processing [2].

Even though such techniques improve the sensing capabilities of robots, recognizing the objects indicated by users remains difficult because user references are often ambiguous during conversations. People enjoy enormous variability in their lexical choices in conversations [5]. Such variability degrades the recognition performance because they might not always use the words contained in a database that stores an object's characteristics, and they also do not always use enough words to identify an object [6]. Even if robots can perfectly recognize speech or pointing gestures, they might not distinguish an object indicated by humans from other objects.

How does a robot encourage a user to clarify his references? We consider two approaches: explicit and implicit. In the first approach, a robot explicitly instructs the person how to refer to objects. We call this the *explicit instruction approach*. For example, the robot explicitly asks a user, "Please describe the object's name, color, and size when pointing at it." The ambiguity of the user's references is expected to decrease if she accepts that request.

In the other approach, the robots implicitly aligns with the people's indicating behaviors. We call this the *implicit alignment approach*. For example, the robot implicitly elicits pointing gestures and lexical expressions contained in the robot's database from a user. When a person is talking with an addressee, she tends to repeat the same lexical [7], syntax [8], expression choice [9], and body movements [10,11] as the addressee. This phenomenon, called alignment, occurs in interactions not only between humans but also between a human and artificial media like spoken dialogue systems [13,14,15,16,17] and robots [18,19]. Through alignment, humans narrow down huge lexical choices and elicit terms, indications, or iconic gestures to naturally identify objects for their interlocutors. Inspired by these alignment findings, some research proposed robotic system that aimed to improve the recognition performance of objects indicated by a user by eliciting the user's references by alignment [20].

However, it remains unknown which approach more effectively decreases ambiguity and improves performance.

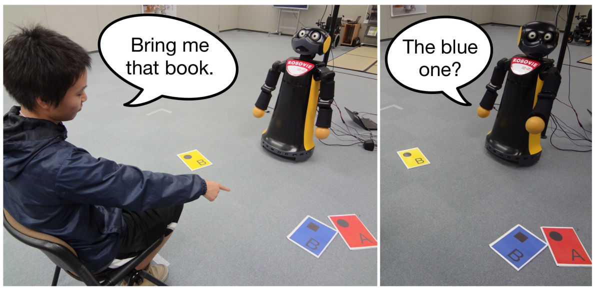

Figure 1. Robot recognizes an object indicated by a user

Moreover, social robots should choose a more appropriate approach by considering not only the recognition performance of the indicated objects but also the user's impression of the interaction. Even though the performance might be increased by either approach, such performance gain becomes worthless if the users hesitate to interact with a robot by the chosen approach and vice versa.

This paper addresses whether the explicit instruction approach or the implicit alignment approach is better for object recognition contexts in conversations with people. We developed a robot system that recognizes objects indicated by a user and experimentally compared the two approaches with our system. Based on the experiment results, we discuss the effectiveness of the approaches.

INTERACTION DESIGN

Object Reference Conversation

To investigate the effect of these two approaches on object reference recognition, we used an interaction called object reference conversations (Figure 2). Such conversations focus on *confirmation behavior*, which is often observed in human-human communication. For example, when person A asks person B to bring a magazine, she is referring to a specific magazine: "Bring that magazine to me." If person B cannot confidently understand which magazine was being referenced, she is likely to ask for confirmation: "This one?" Such conversations are already being used in human-robot interaction research fields to explore several research purposes, including lexical entrainment in human-robot interaction [18] and the implementation of the implicit alignment of reference behaviors [20].

The following are the details of object reference conversations. First, a robot asks a person to refer to an object in an environment where several objects are arranged (*Ask*). Next, she refers to an object (*Refer*), and the robot confirms the object to which she referred (*Confirm*). Then she answers whether the object confirmed by the robot is correct (*Answer*).

Explicit Instruction Approach

Explicit instruction requires the information that the robot wants the interlocutors to use for recognizing the indicated object, because it limits the references and reduces unexpected references. If an interlocutor refers to an object as the robot instructed, the robot will probably recognize it

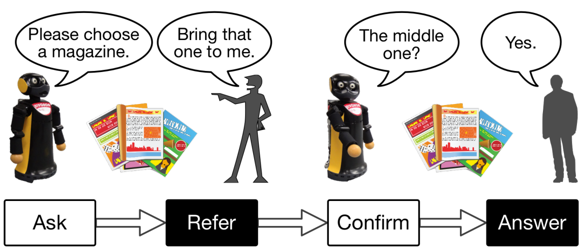

Figure 2. Object reference conversation: white and black boxes denote robot and human turns

Figure 3. Example of explicit robot instructions: robot explicitly instructs how to refer to objects by encouraging interlocutor to use information that was missing in previous references

with high performance. A robot should also give instructions about how to refer to objects in such a way that asks an interlocutor to make a reference that includes as much information as possible that the robot can recognize. If the robot's recognition fails partly because of noise, insufficient speech volume, unclear pointing gestures, and so on, references that include sufficient information increase the chances that the robot will correctly recognize the referenced object.

In addition, if the interlocutor does not follow the robot instructions, the robot should request that she use all of the instructed information for the object references. This suggestion reminds the interlocutor of the instructions and encourages her to use all of the information in subsequent conversations.

Based on these considerations, we designed the explicit instructions of the reference behavior as follows. A robot provides instructions about how to refer to objects in a way that asks the interlocutors to make a reference that includes as much information as possible and requests that the interlocutor use the information that was missing from the previous references. Figure 3 shows an example of explicit instructions from a robot.

Implicit Alignment Approach

We adopted the implicit alignment approach proposed by Kimoto et al. [20]. This approach exploits alignment in object reference conversations. Based on these three alignment phenomena, lexical alignment, gestural alignment, and alignment inhibition, they designed robot behavior where the robot should use minimum information for distinguishing among objects in the environment. Alignment inhibition is the phenomenon of alignment becoming substandard in conversations in some cases. Through this design, humans learn to use references that include enough information to identify the objects by reducing the alignment inhibitions. They implemented this design in the *Confirm* part in the object reference conversation described in the previous section. Their experimental results suggested the possibility of improving

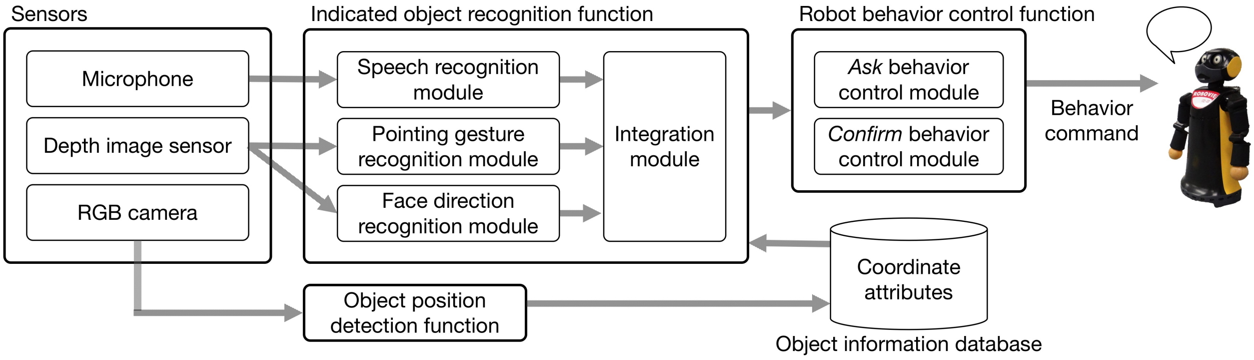

Figure 5. System architecture to recognize objects indicated by user

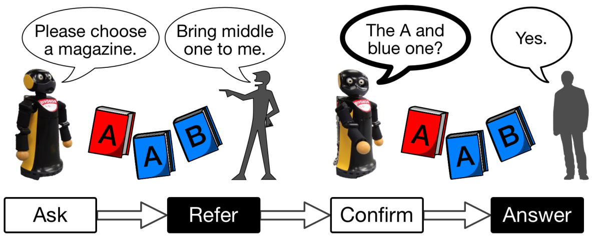

Figure 4. Example of object reference conversation with implicit alignment approach: robot confirmed indicated object using minimum information for distinguishing objects

the recognition performance by the implicit alignment of reference behavior. However, even though they discussed the effectiveness of reference behavior, they didn't evaluate user's impression of the interaction.

Therefore, we used this design of robot behavior as an implicit alignment of reference behavior; a robot should make confirmations that contain minimum information for distinguishing objects. Figure 4 shows an example of an object reference conversation with an implicit alignment approach.

SYSTEM

Figure 5 shows the architecture of our developed system that recognizes the objects indicated by a user. We developed it by referring to past work that implemented the implicit alignment of reference behaviors for object reference conversations [20]. The system consists of five parts: sensors, an object position detection function, an indicated object recognition function, an object information database, and a robot behavior control function. First, the system detects the positions of the objects arranged in the environment and saves them in the object information database. When a user refers to an object, the indicated object recognition function recognizes the interlocutor's reference behavior and estimates the indicated object. The robot behavior control function chooses a robot behavior that corresponds to the implemented approach and sends a behavior command to the robot. The robot confirms the indicated object and asks an interlocutor to refer to it in the next conversation in a way decided by the robot behavior control function.

The system can also have object reference conversations as a basic function. In its *Ask* and *Confirm* parts, the robot performs a behavior that corresponds to which approach is used by the robot.

Robot

In this study, we used Robovie-R ver.2, which is a humanoid robot developed by the Intelligent Robotics and Communication Labs, ATR, that has a human-like upper body designed for communication with humans. The speaker in its mouth can output recorded sound files from the internally-controlled PC located in its body. We used XIMERA for speech synthesis [21]. It has three DOFs for its neck and four DOFs for each arm. Its body has sufficient expressive ability for object reference conversations. It is 1100 mm tall, 560 mm wide, 500 mm deep, and weighs about 57 kg.

Object Position Detection Function

The object position detection function locates the positions of the objects arranged in the environment through AR-markers and saves its ID and positions in the object information database. The AR-markers are read using an RGB camera in the ceiling. We previously saved the attributes of the objects in the database, and the system can extract the attributes of objects from it by IDs.

Indicated Object Recognition Function

To develop this function, we implemented an algorithm [20] that combines the results of speech recognition, pointing gesture recognition, and face direction recognition. Note that this past work [20] only used the speech and pointing gesture recognition results; in this work we added face direction features to increase the performance of the object reference recognition.

Speech Recognition

The speech recognition module receives human speech that refers to an object and outputs the normalized reference likelihood of each object based on speech recognition. To calculate the likelihood, we used a previously proposed method [20] that uses the number of attributes in the human speech, which is captured by a microphone attached to the human's collar. In this system, we used a speech recognition engine called Julius, which gives good performance in Japanese [22].

Pointing Gesture Recognition

The pointing gesture recognition module obtains the body frame data from a depth image sensor called Kinect for Windows v2 and outputs the normalized reference likelihood of each object based on pointing gesture recognition. We modeled the likelihood as the difference from the pointing vector (between the human head and the tip of the human hand) to a vector between the human head and an object with a normal distribution function $N(0, 1)$.

Face Direction Recognition

The face direction recognition module obtains the face direction vector from the depth image sensor and outputs the reference likelihood based on the face direction recognition. We modeled the likelihood based on an angle parallel to the plane of the floor between the face direction vector and a vector between a human head and an object. If the vector is less than 110 degrees, the person is considered to be viewing the object; its likelihood is 1, otherwise 0. This is because a human's field of view is 110 degrees at most [23]. The likelihoods are finally normalized from 0 to 1.

Integration

The integration module merges the reference likelihoods of the speech and both the pointing gesture and face direction recognitions. These three likelihoods are summed and normalized based on previous work [20]. The object with the highest likelihood is estimated to be the object indicated by the interlocutor.

Robot Behavior Control Function

The robot behavior control function determines how the robot confirms the indicated object (*Confirm* behavior) and how it asks an interlocutor to refer to an object (*Ask* behavior) in subsequent conversations. The conversation contents of the *Confirm* and *Ask* behaviors reflect whether the explicit instruction approach or the implicit alignment approach is used.

When using the explicit instruction approach, this function chooses the *Ask* behavior and adopts the explicit instruction approach, and the robot explicitly provides instructions about how to refer to objects. The *Confirm* behavior does not adopt a particular approach, and the robot confirms the indicated object by pointing and verifying all of the information about the indicated object.

When using the implicit alignment approach, this function chooses the *Confirm* behavior and adopts the implicit alignment approach, and the robot confirms the indicated object with minimum information for distinguishing among objects. The *Ask* behavior does not adopt a particular approach, and the robot does not explicitly instruct how to make references.

EXPERIMENT

We experimentally compared the two interactive approaches: explicit instruction and implicit alignment.

Hypotheses and Predictions

If the robot explicitly instructs a particular reference way, the interlocutor knows how to refer to an object and may use it in object reference conversation. This enables the robot to recognize the indicated objects more accurately. However, referencing to an object based on explicit instruction is not common in daily conversations, and so the interlocutor might deem the conversation unnatural. Similarly, the interlocutor might feel the conversation is a load and troublesome because of the unaccustomed conversations.

On the other hand, if there is no explicit instruction about the reference way, the interlocutor might not know how to refer to an object, complicating the robot's ability to recognize indicated objects. However, the interlocutor might feel the conversation is more natural than explicit instructions. Based on these considerations, we made the following two predictions:

Prediction 1: Conversations with the implicit alignment approach will be perceived as a less load, less troublesome, and more natural by the interlocutor than conversations with the explicit instruction approach.

Prediction 2: Object reference recognition performance will outperform the conversations with the explicit instruction approach than conversations with the implicit alignment approach.

Conditions

We controlled an approach that was applied to our developed system (applied approach factor).

The applied approach factor had two levels: explicit instruction and implicit alignment. Both were respectively applied to the *Ask* and *Confirm* parts of the object reference conversation described in the interaction design section. The applied approach factor was a within-participant condition.

Explicit Instruction Condition

In the explicit instruction condition, the robot gives instructions about how to refer to objects in a way that asks interlocutors to make a reference that includes as much information as possible with comments that encourage the interlocutor to use the information missing in the *Ask* part.

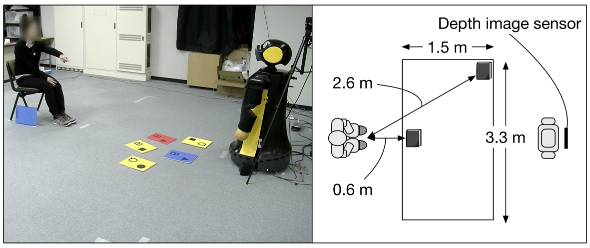

Figure 6. Experimental environment

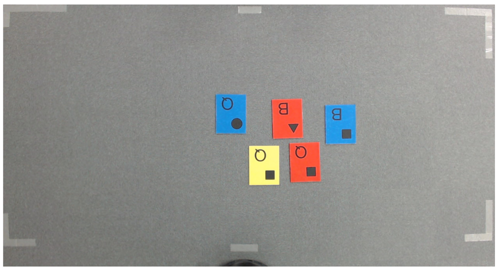

Figure 7. Example of book arrangement

The speech format of the explicit instructions includes two sentences. The first sentence is used every time; the second sentence is only used when a participant did not use all of the information requested by the first sentence in the previous conversation.

For example, the robot says, “Can you refer to a book using its color, a shape on its cover, a letter on its cover as well as by pointing and looking at it? Please refer to a letter and point.”

In the *Confirm* part of this condition, since the robot confirmed the objects with all of the information, it gave every attribute of an object and pointed during the confirmations.

Implicit Alignment Condition

In the implicit alignment condition, unlike the explicit instruction condition, the robot does not explicitly provide instructions about the reference way; it just says, “Please choose a book” in the *Ask* part.

On the other hand, in the *Confirm* part the robot utters a different sentence. For this purpose, we implemented an implicit alignment design for the reference behavior. In this condition, the robot confirmed the object with minimum information for distinguishing among objects; the confirmations were based on the implicit alignment approach of references proposed by Kimoto et al. [20]. This approach determines robot’s object reference behaviors, i.e., pointing behavior and speech, by considering objects’ position relationships and characteristics. A robot used pointing gesture when it would be useful to decrease candidates of referenced objects. For example, if robot’s pointing gesture would become vague to an interlocutor, the robot did not use the gesture.

The speech format of the confirmations is the sequence of the attributes of objects. For example, the robot says, “That blue book with a circle on its cover?” or “That yellow book?”

Environment

Figure 6 shows the environment. The participants were seated in front of the robot. Five objects were placed in a 1.5 m by 3.3 m rectangular area between the robot and the participant. The books were grouped close together without overlapping. These objects were approximately 0.6-2.6 m from the participants.

We controlled the attributes of the books, all of which were 21 cm by 27.5 cm. Their attributes were color and the shape and the letter on the cover. There were three colors: red, blue, or yellow. Three shapes were placed on the book covers: a circle, a triangle, or a square. There were two letters: Q and B. We prepared 18 books to satisfy all combinations of the attributes.

Procedure

We conducted our experiment as follows. First, we explained it to participants who signed consent forms. Next, we gave them the following oral instructions: “The robot can recognize human speech, pointing gestures, and face direction. The robot will ask you to indicate a book. Indicate it as if you were dealing with a person.”

After the instructions, the participants selected five books among the 18 and arranged them based on the experimenter’s instruction: “Please arrange the books in one place.” An example of the arrangement is shown in Figure 7. After that the participant repeated the object reference conversations ten times. We call this set of ten object references the *conversation sessions*, which were conducted in both applied approach conditions: explicit instruction and implicit alignment. The participants answered questionnaires about their impressions of the conversations after each conversation session. We counterbalanced the order of the interactive approach conditions.

Participants

Twenty (ten females and ten males) native Japanese speakers whose average age was 35.5 (*SD* = 9.9) participated in our experiment.

Measurement

Impression of Conversations

To investigate the participant’s impressions of the conversations, we prepared the following three questionnaire items and evaluated them on a 7-point scale:

1. The conversation with the robot was a load (load feeling).
2. The conversation with the robot was troublesome (troublesome feeling).

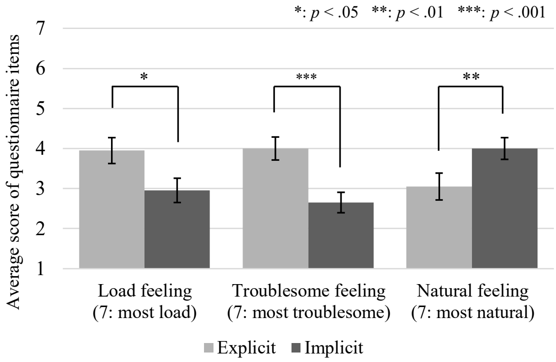

Figure 8. Impression of conversation (mean ± *SE*)

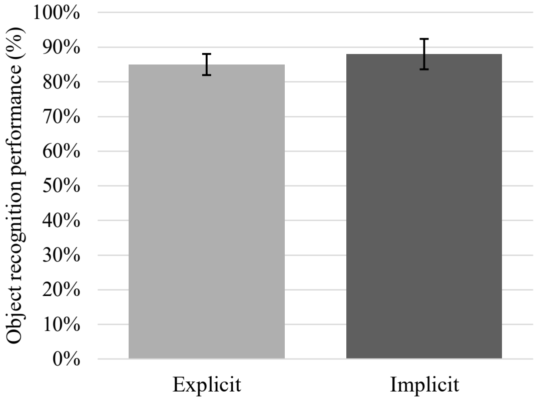

Figure 9. Object recognition performance (mean ± *SE*)

3. The conversation with the robot was natural (natural feeling).

Recognition Performance

The recognition performance is the success rate of the object reference recognition. We calculated it from the number of object references correctly recognized by the robot per conversation session, which was a set of ten object reference recognitions.

RESULTS

Verification of Prediction 1

Figure 8 shows the results of the questionnaire items. To verify the effect of each condition, we conducted a paired t-test for each of the questionnaire items. For the load feeling, we found a significant difference among the conditions ($t(19) = 2.1$, $p = .049$, $d = .72$). For the troublesome feeling, we found a significant difference among the conditions ($t(19) = 4.2$, $p < .001$, $d = 1.1$). For the natural feeling, we also found a significant difference among the conditions ($t(19) = -3.0$, $p = .008$, $d = .70$). These results supported Prediction 1.

Verification of Prediction 2

Figure 9 shows the recognition performance results. To verify the effect of each condition, we conducted a paired t-test and found no significant difference between the two interactive approach conditions ($t(19) = -.58$, $p = .57$, $d = .18$). This result indicated that Prediction 2 was not supported.

DISCUSSION

Interpretation of Results

From the experiment results, the interlocutor impressions of the conversations with the implicit alignment approach were perceived to be significantly a less load, less troublesome, and more natural than the explicit instruction approach. Accordingly, explicit instruction is not natural for people and created feelings of uneasy interaction in the interlocutors.

We identified no significant difference in the recognition performances between the two approaches. This result can be interpreted in two ways. First, the implicit alignment approach improved the recognition performance at the same level as the explicit instruction approach. Although we cannot verify the effects of the implicit alignment approach in our experimental settings, a counter condition of an implicit alignment approach is not a no-implicit alignment approach but an explicit instruction approach, and the implicit alignment approach probably improves the recognition performance, as argued by past work [20].

Second, the explicit instruction approach did not improve the recognition performance very much. We predicted that the interlocutors would refer to an object as the robot instructed; but in the experiment 28% of references in the explicit instruction condition did not follow the robot's instructions. Some instructed with information that was dropped out, for example, a pointing gesture and a color. Explicit instruction is probably not an effective way to induce interlocutors to encourage users to adopt clear but recognizable references.

For these reasons, we conclude that the implicit alignment approach is better than the explicit instruction approach for object recognition contexts in conversations with people. Our findings are useful for designing interactions for social robots because good impressions of conversations are important for them because they often interact with people. These findings can be integrated not only for object recognition contexts but also for many other contexts since determining whether to use an explicit or implicit approach is conceivable in other contexts.

Limitation

We conducted this experiment's study in a limited situation. The participants referred to objects with only three features: color, a geometric shape, and a letter. In real environments, the features of objects are not limited and obviously influence the reference ways. But since the interaction way between a robot and an interlocutor does not depend on features, our findings are general for other objects.

CONCLUSION

In this study, we focused on two interactive approaches for object recognition contexts in conversations with people: explicit instruction and implicit alignment. We developed a system that recognized the indicated objects by integrating the speech, pointing, and face direction recognition results and experimentally compared the performance and feeling perspectives between the two interactive approaches.

Experimental results indicated that the participants perceived the conversations with the implicit alignment approach to be a less load, less troublesome, and more natural than the explicit instruction approach. The overall impression of the conversations with the implicit alignment approach exceeded the explicit instruction approach. The object reference recognition performance did not differ between the two approaches, indicating that the implicit alignment approach is better than the explicit instruction approach for object recognition contexts in conversations with people. We believe that our findings are useful for the design interaction of social robots that frequently interact with people.

ACKNOWLEDGMENTS

Omitted for anonymized review.

REFERENCES

1. Kai Nickel and Rainer Stiefelhagen, 2007. Visual recognition of pointing gestures for human–robot interaction. *Image and Vision Computing 25*, 12: 1875-1884.
2. Boris Schauerte and Gernot A. Fink, 2010. Focusing computational visual attention in multi-modal human-robot interaction. In *Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, 1-8. http://dx.doi.org/10.1145/1891903.1891912
3. J. Richarz, T. Plotz, and G. A. Fink, 2008. Real-time detection and interpretation of 3D deictic gestures for interactionwith an intelligent environment. In *Proceedings of the Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 1-4. http://dx.doi.org/10.1109/ICPR.2008.4761609
4. Osamu Sugiyama, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, Norihiro Hagita, and Yuichiro Anzai, 2006. Humanlike conversation with gestures and verbal cues based on a three-layer attention-drawing model. *Connection Science 18*, 4: 379-402.
5. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, 1987. The vocabulary problem in human-system communication. *Commun. ACM 30*, 11: 964-971.
6. Shinozawa Kazuhiko, Miyashita Takahiro, Kakio Masayuki, and N. Hagita, 2007. User specification method and humanoid confirmation behavior. In *Proceedings of the 2007 7th IEEE-RAS International Conference on Humanoid Robots*, 366-370. http://dx.doi.org/10.1109/ICHR.2007.4813895
7. S. E. Brennan and H. H. Clark, 1996. Conceptual pacts and lexical choice in conversation. *J Exp Psychol Learn Mem Cogn 22*, 6: 1482-1493.
8. H. P. Branigan, M. J. Pickering, and A. A. Cleland, 2000. Syntactic co-ordination in dialogue. *Cognition 75*, 2: B13-25.
9. Simon Garrod and Anthony Anderson, 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. Cognition 27, 2: 181-218.
10. Albert E. Scheflen, 1964. The Significance of Posture in Communication Systems. *Psychiatry 27*, 4: 316-331.
11. Adam Kendon, 1970. Movement coordination in social interaction: Some examples described. *Acta Psychologica 32*: 101-125.
12. T. L. Chartrand and J. A. Bargh, 1999. The chameleon effect: the perception-behavior link and social interaction. *J Pers Soc Psychol 76*, 6: 893-910.
13. Holly P Branigan, Martin J Pickering, Jamie Pearson, Janet F McLean, and Clifford Nass, 2003. Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society*, 186-191.
14. Susan E Brennan, 1996. Lexical entrainment in spontaneous dialog. In *Proceedings of the 1996 International Symposium on Spoken Dialogue*, 41-44.
15. Joakim Gustafson, Anette Larsson, Rolf Carlson, and K Hellman, 1997. How do system questions influence lexical choices in user answers? In *Proceedings of the Eurospeech*.
16. Jamie Pearson, Jiang Hu, Holly P. Branigan, Martin J. Pickering, and Clifford I. Nass, 2006. Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1177-1180. http://dx.doi.org/10.1145/1124772.1124948
17. Stefanie Tomko and Roni Rosenfeld, 2004. Shaping spoken input in user-initiative systems. In *Proceedings of the 8th International Conference on Spoken Language Processing*.
18. Takamasa Iio, Masahiro Shiomi, Kazuhiko Shinozawa, Katsunori Shimohara, Mitsunori Miki, and Norihiro Hagita, 2015. Lexical Entrainment in Human Robot Interaction. *International Journal of Social Robotics 7*, 2: 253-263.
19. Takamasa Iio, Masahiro Shiomi, Kazuhiko Shinozawa, Takaaki Akimoto, Katsunori Shimohara, and Norihiro Hagita, 2011. Investigating Entrainment of People's Pointing Gestures by Robot's Gestures Using a WOZ

Method. *International Journal of Social Robotics 3*, 4: 405-414.

20. Mitsuhiko Kimoto, Takamasa Iio, Masahiro Shiomi, Ivan Tanev, Katsunori Shimohara, and Norihiro Hagita, 2015. Improvement of object reference recognition through human robot alignment. In *Proceedings of the Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*, 337-342. http://dx.doi.org/10.1109/ROMAN.2015.7333672
21. Hisashi Kawai, Tomoki Toda, Jinfu Ni, Minoru Tsuzaki, and Keiichi Tokuda, 2004. XIMERA: A new TTS from ATR based on corpus-based technologies. In *Proceedings of the Fifth ISCA Workshop on Speech Synthesis*.
22. Akinobu Lee, Tatsuya Kawahara, and Shuji Doshita, 1998. An efficient two-pass search algorithm using word trellis index. In *Proceedings of the ICSLP'98*, 1831-1834.
23. S.J. Ryan, A.P. Schachat, C.P. Wilkinson, D.R. Hinton, S.V.R. Sadda, and P. Wiedemann, 2012. *Retina*. Elsevier Health Sciences.