Unraveling the Old-yet-new Problem of Common Ground Using Cutting-edge Technology to Enable Computer-to-human Communication

Ryuichiro Higashinaka
Visiting Senior Distinguished Researcher, NTT Human Informatics Laboratories

Abstract

Regarding communication between people, when people share "common ground"—such as common topics, prerequisites, and perceptions—they can understand each other and engage in communication. If we could build common ground between computers and people, communication between them would become smoother. The concept of common ground has been around for a long time; however, many of its facets remain unexplained. By unraveling these facets individually, it will be possible to implement the process of building common ground. Ryuichiro Higashinaka, a visiting senior distinguished researcher at NTT Human Informatics Laboratories, is attempting to elucidate common ground by using generative artificial intelligence and other techniques to enable communication between computers and humans. We spoke with him about his approach to tackling "old-yet-new" problems and his belief that "it wasn't a failure, but rather the effort was simply not successful."

Keywords: common ground, computer-to-human communication, robot-to-robot communication

Explaining the mechanism for constructing common ground to enable communication between people and computers

—Would you tell us about the research you are currently conducting?

Regarding communication between people, mutual understanding becomes possible when the parties agree on "common ground"—an understanding of common topics, prerequisites, perceptions, and so on. Taking this common ground as my research theme, I have been elucidating and modeling the mechanism by which people build common ground, and by applying this mechanism to build common ground between computers and humans, I aim to enable smooth communication between them.

In my previous interview (July 2022 issue), I summarized our joint research with the University of Electro-Communications, Shizuoka University, and

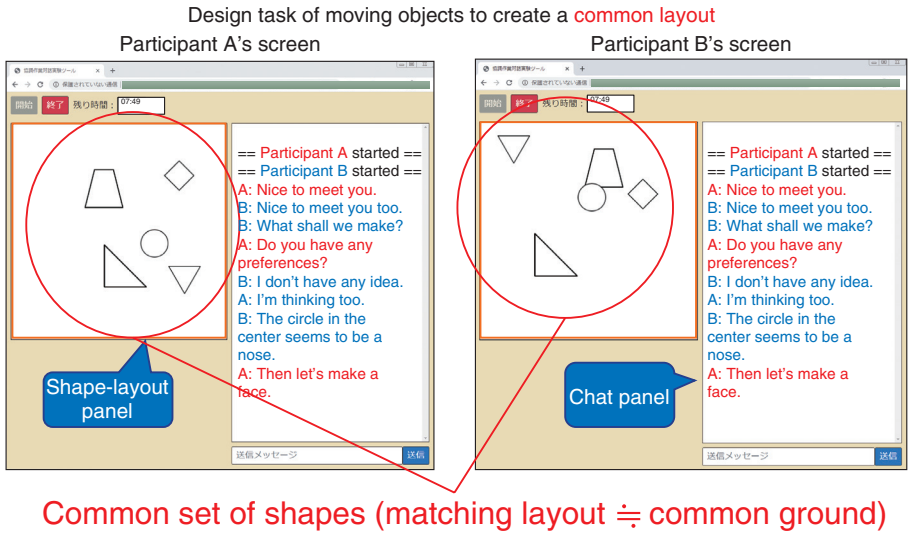


Fig. 1. CommonLayout task.

Keio University as the following four points. First, we set up a task called "CommonLayout (collaborative arrangement of shapes)" (**Fig. 1**). In this task, two people are given a common set of different shapes with different layouts, and they jointly agree on a certain layout through text chat. By measuring the distance between certain identical shapes in each set, we could quantify the extent of construction of common ground at each point of the dialogue. Second, we confirmed the usefulness and importance of "naming" in the CommonLayout task, that is, explaining the way to arrange the shapes by comparing the arranged shapes to general objects (e.g., aligning two circles horizontally like the wheels of a car). Third, we expanded our research on the CommonLayout task through text chat, constructed and analyzed a corpus (a large-scale accumulation of data such as natural language) of conversations using audio and video during the task, and confirmed that it is easier to build common ground when the participants can use video. Finally, we found that it is easier to build common ground under the condition that the participants are acquainted rather than meeting for the first time.

Since the previous interview, our research has progressed as explained as follows. We first found that goal naming, that is, the two participants set the goal of a discussion, is a vital factor in successful collaboration. Regarding the tangram naming task (**Fig. 2**), in which the participants name shapes created with a "tangram" (a puzzle made from seven geometric pieces cut from a square) through conversation, we then found that the participants use utterances that name the entire shape (holistic utterances) and utterances that describe the details of the shape (analytic utterances) and experimentally clarified how they use these utterances. We next conducted an experiment on cognitive models of successful and unsuccessful cases of communication when the two participants in the task are looking at the same tangram. Finally, we devised new tasks to measure the extent of construction of common ground, collected data and investigated the extent to which multimodal information affects common grounding.

Regarding the effectiveness of goal naming, we found that naming something that does not yet exist is effective in achieving the goal of the two participants recognizing the same set of shapes and arranging them in the same way (in a successful dialogue). For example, it is effective to achieve that goal by first naming a set of shapes by saying, "It looks like it has a driver's cab and a trailer, so let's call this set a truck," then using that name as the basis for indicating the positions of the shapes (e.g., the position of the trailer) and arranging them (e.g., arrange the trailer face up). We then analyzed the timing at which goal naming appears, and the analysis results indicate that in successful dialogues, goal naming appears early in the flow of the dialogue; in other words, aligning the awareness of the two participants from the outset results in more successful dialogues (**Fig. 3**).

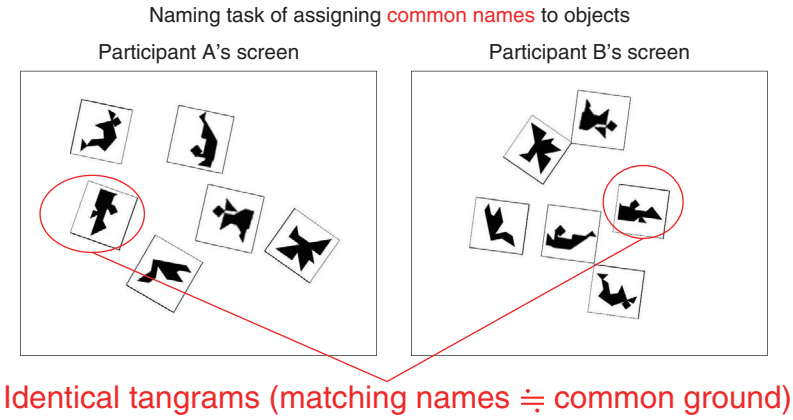


Fig. 2. Tangram-naming task.

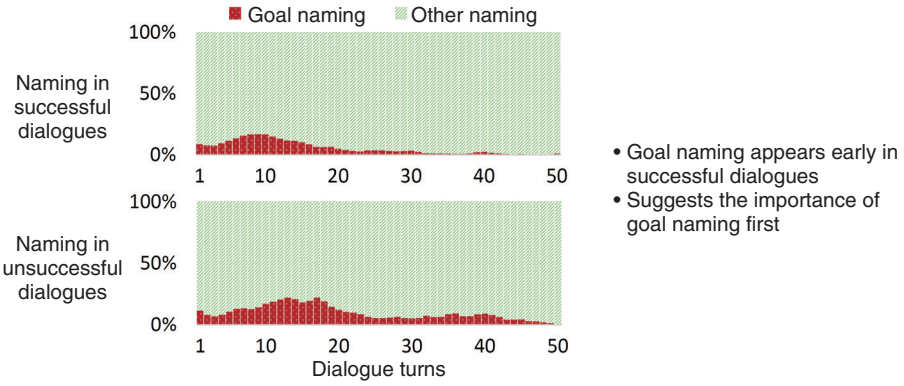


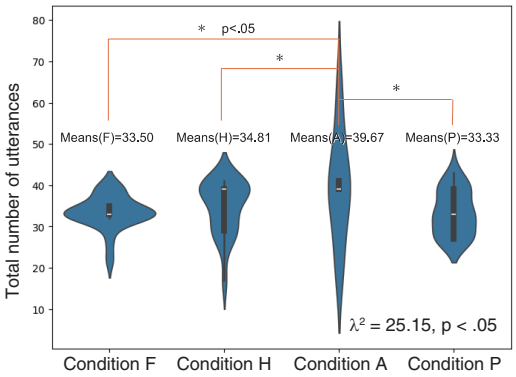
Fig. 3. Goal naming.

Regarding holistic utterances and analytic utterances, when we look at actual human activities, we find that people tend to use holistic utterances more frequently as an initial response, but when a problem arises, they tend to use analytic utterances.

Considering those findings, we prototyped a prompt-based system—using generative artificial intelligence—that first produces holistic utterances then produces analytical utterances according to the response, and we analyzed the system's accuracy (degree of task accomplishment) (**Fig. 4**). Under condition P, in which both holistic utterances and analytic utterances were used appropriately in accordance with the response, the task was completed with a stable number of utterances. However, under condition H, in which only holistic utterances were used, the task was sometimes completed in a small number of utterances, but the number of utterances varied greatly. Under condition A, in which only analytic utterances were used, a huge number of utterances were required to complete the task (which was time-consuming). These results suggest that in dialogue between two people, they can achieve a stable conversation by first using holistic utterances then switching to analytic utterances if the other party does not understand the holistic approach.

—Apart from words, what else have you learned?

Let me explain the processing that people use for the tangram-naming task from the perspective of a cognitive model. First, the sender (one party in the conversation) imagines the image they want to convey to the other party (receiver). They then verbalize that image and utter it to the receiver. Next, the receiver creates an image from the words, matches



Condition P (using both holistic and analytic utterances appropriately) allows the task to be accomplished with a stable number of utterances.

Condition H alone is sometimes successful.
Condition A alone is very time-consuming.

Condition F: Normal conversation without awareness of holistic utterances and analytic utterances
Condition H: Conversation using holistic utterances only
Condition A: Conversation using analytic utterances only
Condition P: Conversation with awareness of using both holistic and analytic utterances appropriately
The width of each blue plot shows the frequency of the total number of utterances until the task is accomplished.

Fig. 4. Total number of utterances for each experimental condition and system accuracy.

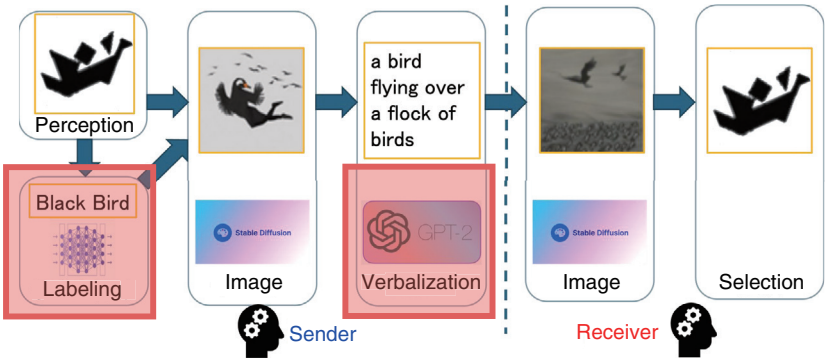


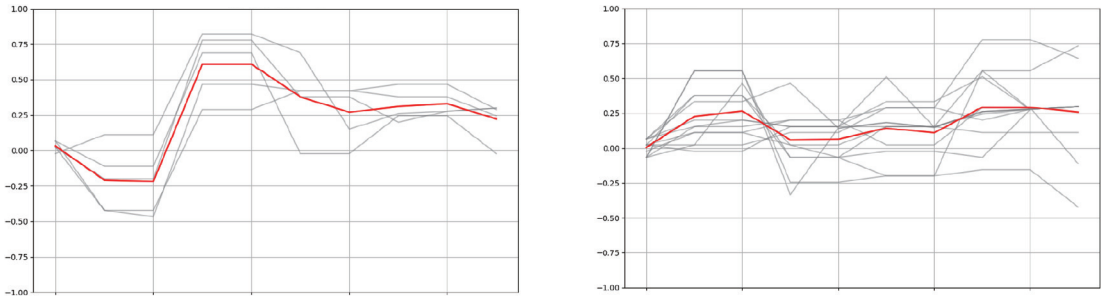
Fig. 5. Modeling based on cognitive models.

that image with a tangram in front of them, and selects that tangram. With that process in mind, we experimentally simulated this process by cognitive modeling on a computer to see how the process could be adjusted to enable the receiver to understand the image sent by the sender when the receiver could not understand it. Specifically, we focused on adjusting both visualization and verbalization and investigated which aspects of these processes needed to be adjusted to improve the accuracy of determining the same tangram (**Fig. 5**).

The results of this simulation confirmed the following points. While accuracy improved just by adjusting verbalization, adjusting visualization as well produced a variety of utterances, approaching human-like interactions. This result means that it is important to adjust both visualization and verbalization. We presented these results at the 47th Annual Meeting of the Cognitive Science Society (CogSci 2025), held from July 28 to August 3, 2025.

Regarding new tasks and data collection to measure common ground, we proposed a collaborative scene reconstruction task and used it to collect data. In this task, two participants (speakers) are given a set of images extracted from the same video, which was divided into ten parts, and one image was cut from each part. Each speaker received a different set of images, which were presented to them in a random order, and were tasked with predicting the correct order for each image and rearranging them accordingly using body language while engaging in a verbal dialogue with the other speaker.

Considering the degree to which the order of arranged images of each speaker match is as a match of common ground, we analyzed the relationship between the speakers' facial expressions and gestures and common ground. Compared with the above-described CommonLayout task, the proposed collaborative scene reconstruction task has an added time axis, and time is expressed through physical movements, etc.,



Example of good progress toward common ground

Example of poor progress toward common ground

Fig. 6. Visualizing the construction of common ground.

so it can be considered a multimodal version of the CommonLayout task. By separating the time spent sorting the video images from the time spent explaining the images, we could measure only the impact of facial expressions and gestures on building common ground. In this cognitive-modeling experiment, we confirmed that the concordance rate between the two participants increased with each phase, and that common ground was being built with each phase (**Fig. 6**, left graph). By investigating what facial expressions and gestures emerged in relation to the building of common ground, we can analyze the relationship between modality and common ground.

Research on robot-to-robot dialogue and turn-taking in human conversation

—What type of research are you conducting other than that on common ground?

In collaboration with Osaka University and the Advanced Telecommunications Research Institute International (ATR), I have been researching conversational robots. During this collaboration, we have been investigating the effect of interaction with multiple robots on the opinion of a speaker.

Two conversational robots, each with a different opinion, were set up to engage in a debate. In this situation, several conversational patterns were set; for example, in one pattern, both robots agreed, and in another, one robot continued to disagree. During the debate, one robot asked the person also in attendance, "What do you think?", and we investigated how the person's opinion changed in accordance with whether the robots agreed or disagreed with that opinion.

The pattern that changed the person's behavior the most was when, for example, the robots were debating whether to take a trip to Hokkaido or Okinawa, and when the person was asked, "What do you think?" and replied "Hokkaido sounds good," the robot that had suggested Okinawa suddenly changed its opinion and replied "That's true (Hokkaido sounds good)," and the person felt more convinced and confident in their opinion. This investigation was demonstrated at Miraikan, the National Museum of Emerging Science and Innovation, from March to August 2022 as part of the "You and Robots—What is it to be Human?" exhibition. The results of the investigation were published in the International Journal of Social Robotics.

In a joint research project between Nagoya University, where I teach, and NTT Communication Science Laboratories, we are researching turn-taking, a mechanism by which the initiative is passed to other speakers in accordance with the flow of the conversation. In turn-taking, it is important to know when the other party has finished speaking, so we are constructing a model (in Japanese) called the voice activity projection (VAP) model, which can predict when the other party will finish speaking. We conducted basic evaluation experiments to train the VAP model—based on the Transformer deep-learning model—using Japanese, English, and English-plus-Japanese data and to see how the performance of the VAP model changed when the pattern (i.e., Japanese, English, or English plus Japanese) changed. We also verified the impact of incorporating the VAP model into an actual dialogue system, such as whether the performance would improve.

I published two textbooks (in Japanese)—based on our activities to date—for the entire field, "How to

Create a Dialogue System" (published by Kindai Kagaku Sha Co., Ltd.) in February 2023 and "Creating a Real-time Multimodal Dialogue System with Python and Large Language Models" (co-authored, published by Kagakujyoho shuppan Co., Ltd.) in June 2024.

Tackling challenging, old-yet-new problems with cutting-edge technology

—What do you keep in mind as a researcher?

Although it is true in any field of research, we face many old-yet-new problems, and it is important to tackle these problems thoroughly. My theme, "common ground," has existed for a long time as a concept and classical theory; however, we still do not understand many of its aspects, and new challenges continue to emerge as the research progresses, so it is truly an old-yet-new problem.

Taking on these problems involves mundane, time-consuming tasks such as collecting data, running simulations, and creating the environment for those tasks. You also won't know whether a task will lead to a solution until you try it. By repeating these tasks, you will discover something, but often that something is only a small part of a certain phenomenon. Although these are old problems, they are still "new" problems because they have not been addressed for a long time.

Many new and cutting-edge technologies, such as large language models, have emerged, and by using them, hitherto impossible simulations are now possible. If we keep abreast of cutting-edge technology and take new steps in research, I believe we can advance our research to a practical level. The impact of our research is great because it is a problem that has remained untouched for a long time. That is why it is important to tackle old-yet-new problems thoroughly.

Even if the result is a failure, if you can learn something, failure just means you didn't succeed. That's why it's important to pursue research that captures the essence of the matter

—Do you have a message for younger researchers?

I hope that you will grasp the essence of a problem or theme when proceeding with your research. To do that, it is important to have a firm grasp of the basic aspects of the problem or theme. As I mentioned previously, when tackling an old-yet-new problem, you can take a new step by keeping abreast of cutting-edge technology and using it to advance your research; however, if you just pursue cutting-edge technology for the sake of it, you may look good but may end up being drawn to that pursuit and lose sight of the essence. This risk does not only apply to old-yet-new problems; if you can get a firm grasp of the basics, you will be able to understand the relationship of your research to cutting-edge technology and, conversely, come up with ideas for effective ways to use such technology, which will help move your research forward.

After grasping the essence of the problem or theme, I urge you to take on the problem without fear of failure. If you can grasp the essence, even if the result seems to be a failure, you will be able to gain something useful by understanding the situation and the cause of failure. In other words, it is not a failure but rather the effort was simply not successful, and you can use what you have gained to move forward.

■ **Interviewee profile**

Ryuichiro Higashinaka received a B.A. in environmental information, Master of Media and Governance, and Ph.D. from Keio University, Kanagawa, in 1999, 2001, and 2008. He joined NTT in 2001. Since 2020, he has been a professor at the Graduate School of Informatics, Nagoya University and visiting senior distinguished researcher at NTT. His research interests include building question-answering systems and spoken dialogue systems. From November 2004 to March 2006, he was a visiting researcher at the University of Sheffield in the UK. He received the Maejima Hisoka Award from the Tsushinbunka Association in 2014 and the Prize for Science and Technology of the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2016. He is a member of the Institute of Electronics, Information and Communication Engineers, the Japanese Society for Artificial Intelligence, the Information Processing Society of Japan, and the Association for Natural Language Processing.