Exploring the Impact of Modalities on Building Common Ground Using the Collaborative Scene Reconstruction Task

Yosuke Ujigwa [1], Asuka Shiotani[1], Masato Takizawa[1], Eisuke Midorikawa[1], Ryuichiro Higashinaka[2], Kazunori Takashio[1]

[1]Keio University, Japan , [2]NTT Corporation, Japan

Correspondence: ujigawa@keio.jp

Abstract

To deepen our understanding of verbal and non-verbal modalities in establishing common ground, this study introduces a novel "collaborative scene reconstruction task." In this task, pairs of participants, each provided with distinct image sets derived from the same video, work together to reconstruct the sequence of the original video. The level of agreement between the participants on the image order—quantified using Kendall's rank correlation coefficient—serves as a measure of common ground construction. This approach enables the analysis of how various modalities contribute to the construction of common ground. A corpus comprising 40 dialogues from 20 participants was collected and analyzed. The findings suggest that specific gestures play a significant role in fostering common ground, offering valuable insights for the development of dialogue systems that leverage multimodal information to enhance the user construction of common ground.

1 Introduction

Understanding the essence of human communication is a crucial challenge in the fields of artificial intelligence and human-computer interaction (HCI). The concept of common ground, proposed by Clark, refers to the shared knowledge and beliefs between participants in a dialogue, forming the foundation for smooth communication (Clark, 1996). Unraveling the process of grounding is not only essential for understanding the mechanisms of deep relational building among humans but also holds significant implications for developing AI agents and robots capable of interacting naturally with humans (Morita et al., 2024).

Recent research has highlighted the influence of multimodal communication channels and social relationships on grounding (Furuya et al., 2022). Visual cues, in particular, have been shown to facilitate common ground construction, though the specific elements of visual information that are most effective remain insufficiently clarified. Additionally, traditional experimental settings often feature tasks with relative ease, making it challenging to conduct a detailed analysis of failures in grounding (Udagawa and Aizawa, 2019).

In this study, we propose a novel collaborative task that enables clearer observation of the influence of physical expression as a visual modality and allows for detailed analysis of both successful and unsuccessful grounding instances. The task emphasizes the role of non-verbal communication, enabling precise analysis of how physical modalities, such as gestures and gaze, influence grounding (Kendon, 1983). By appropriately adjusting task difficulty, the study aims to observe the dynamics of grounding in more realistic scenarios.

2 Construction Process of Common Ground

2.1 Common Ground

Common ground refers to the totality of shared knowledge, beliefs, and assumptions between participants in a dialogue (Clark, 1996). In everyday conversations, it is assumed that a basic common ground concerning general knowledge and language understanding already exists, and through interaction, new common ground is dynamically constructed. This process is critical for enhancing the efficiency and effectiveness of communication (Mitsuda et al., 2021).

Understanding and constructing common ground are essential for smooth dialogues. When dialogue participants do not accurately grasp common ground, misunderstandings and discrepancies may occur, potentially hindering communication. Conversely, when sufficient common ground is established, it allows for the omission of information and reliance on implicit understanding, facilitating efficient communication (Nakano et al., 2015).

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 12–19
May 27–30, 2025.

In recent HCI research, the concept of common ground has been applied to the design of interactions between humans and AI agents (Nakano, 2019). Developing advanced dialogue systems requires the ability to appropriately construct and maintain common ground with users, a capability that significantly influences the naturalness and effectiveness of the system.

2.2 Modalities in Dialogue

In dialogue, modality refers to the various sensory channels and forms of expression used for information transmission. Beyond linguistic modalities (spoken and written language), non-verbal modalities (such as facial expressions, gestures, posture, and gaze) enable rich and multi-layered communication (Ekman and Friesen, 1969).

Research on multimodal communication has demonstrated that, compared to dialogue relying on a single modality, the efficiency of information transmission and comprehension improves (Kipp, 2005). Non-verbal modalities are particularly crucial in conveying linguistically ambiguous content or complex concepts. For instance, gestures and facial expressions contribute to complementing and emphasizing verbal content, as well as communicating the speaker's emotions and attitudes (McNeill, 1992).

Recent HCI research has actively incorporated these insights into the design of multimodal interfaces (Krauss et al., 2000). In human-AI agent interaction, elucidating insights into the process of grounding and adapting linguistic and non-linguistic modality elements that contribute to its construction are expected to enable more natural and effective communication.

2.3 Previous Research on Construction Process of common ground

In the study of grounding processes, a common approach involves setting specific tasks and analyzing the dialogue between participants (Benotti and Blackburn, 2021). Tasks such as the map task (Ichikawa et al., 2000) and the referential communication task (Anderson et al., 1991) have been widely used. These studies have provided valuable insights into the formation of common ground and its impact on dialogue efficiency.

However, many traditional studies have focused on the relationship between the final task outcome and common ground, with limited detailed analysis of the grounding process itself (Nakano, 2019). A pioneering study addressing this issue is the research by Udagawa and Aizawa (2019), which proposed a new corpus for analyzing the grounding process in a continuously and partially observable context. Nevertheless, this study used text chat, thus failing to account for the influence of non-verbal modalities (Carney and Harrigan, 2003) and the social relationships between interlocutors (Taylor, 1968).

The study by Furuya et al. (2022) analyzed the impact of modality and social relationships on grounding using the "CommonLayout". Their research demonstrated that rich modalities and deep social relationships facilitate grounding. However, it did not clarify which elements of visual information are particularly effective, and the low task difficulty made detailed analysis of grounding failures challenging.

Building on these previous studies, the current research aims to develop a new collaborative task that allows for a more refined analysis of the impact of non-verbal modalities, particularly physical modalities, on grounding. This task will also enable the observation of both successful and unsuccessful grounding instances.This enables the analysis of elements of physical modality that contribute to foundational construction, providing deeper insights into the fields of Human-Computer Interaction (HCI) and communication studies.

3 Collaborative Scene Reordering Task

This study proposes a new collaborative task, the "Collaborative Scene Reordering Task," designed to analyze the impact of modality on the construction of common ground. The task aims to examine how physical modalities during dialogue influence grounding and to provide a detailed analysis of this process.

The task is designed to meet the following requirements:

1. Enable two participants to construct common ground through dialogue.

2. Ensure that as the construction of common ground progresses, task performance improves.

3. Enable the analysis of the degree of grounding achieved at the conclusion, including both successful and unsuccessful cases of grounding construction.

4. Encourage the manifestation of non-verbal behaviors during communication, enabling a more detailed analysis of the impact of physical modalities.

This task is expected to offer insights into the role of physical modalities in grounding, enhancing our understanding of their contribution to effective communication.

The Collaborative Scene Reordering Task involves two participants, each possessing separate pieces of information. The task is divided into a transmission phase and a working phase, which alternate as a single set. In this task, participants share their respective information to reorder a set of images according to the narrative flow of a single story.

The transmission phase, where participants exchange information, is clearly separated from the work phase, where they physically reorder the images. This separation allows participants to allocate more cognitive resources to communication, encouraging them to focus on physical expressions and their partner's information during the interaction. The task is specifically designed to encourage nonverbal behaviors during the transmission phase. The use of visually dynamic and motion-rich video material as the basis for the images also supports this objective.

Task Description:

Setup: Each participant receives 10 shuffled images, extracted from a one-minute video, out of a total of 20 images. Neither participant has the full set, requiring them to infer and communicate about the missing parts.

Transmission Phase: Participants discuss their images face-to-face, focusing solely on sharing information. Physical modalities such as gestures and expressions are encouraged to aid communication. The phase is designed to elicit non-verbal behaviors by separating it from the work phase, preventing simultaneous reordering and discussion.

Work phase: Participants independently reorder their images based on the insights gained from the transmission phase. No communication is allowed during this phase, enabling a clear assessment of the understanding and common ground constructed earlier.

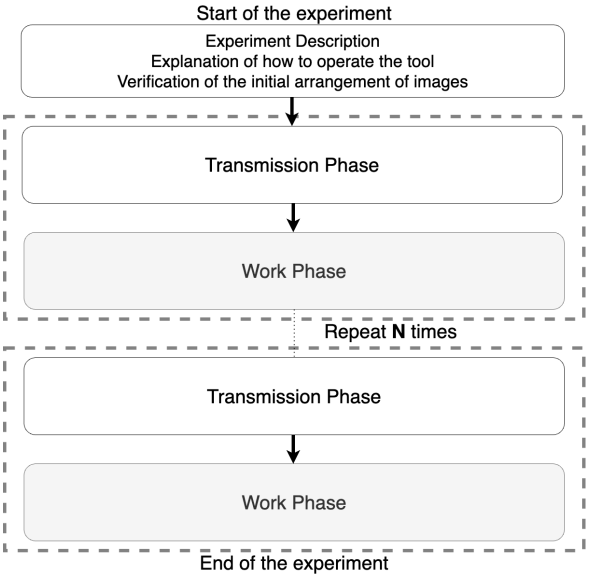

Figure 1: Flow of the Task

The Collaborative Scene Reordering Task builds upon the basic structure of the Collaborative Object Arrangement Task, where two participants are given objects and individually manipulate them based on the communication content with their partner. The final outcomes are compared to analyze the process of grounding. By recording the degree of completion of the image reordering during each working phase, the process of grounding facilitated by each transmission phase can be analyzed in detail. The flow of task implementation is as follows Fig.1.

This task design is based on Clark's theory of common ground (Clark, 1996), intentionally creating asymmetry of knowledge between participants to enable a clearer observation of the shared understanding process of grounding. Additionally, it emphasizes the importance of nonverbal behavior, drawing on Kendon's research on gestures (Kendon, 1983).

4 Experiment

This section describes the experiment conducted using the task proposed in the previous section, aimed at analyzing the influence of modalities in the common ground construction process.

4.1 Participants and Environment

The participants in the experiment were 40 individuals, unrelated to the project, who were gathered via a cloud service. The participant pairs consisted of 10 randomly formed pairs (20 participants in total: 4 males and 16 females), with an average age

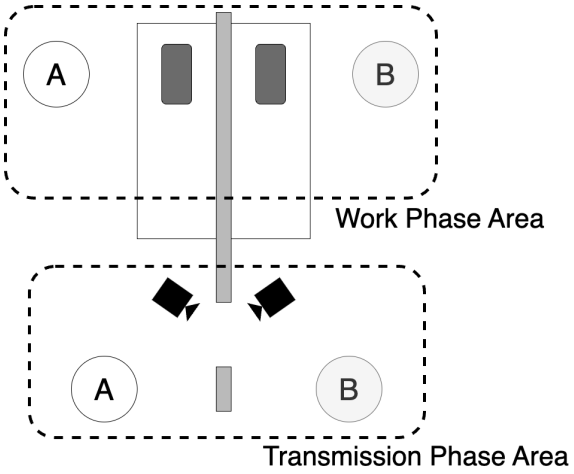

Figure 2: The environment of a Task

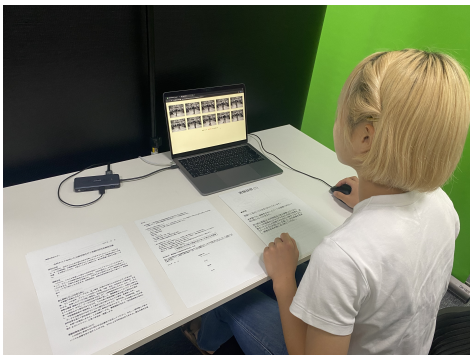

Figure 3: The environment of a work area

Figure 4: The environment of a transmission area

Figure 5: Image of the initial arrangement of images

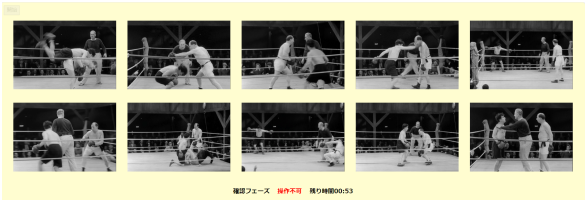

Figure 6: Example of the actual screen of the tool

of 47.15 years (SD = 10.38). All pairs were meeting for the first time.. Each pair completed tasks for 4 videos, collecting a total of 40 data points.

The experiment was conducted in a space divided into a work area and a transmission area (Fig.2).

The experiment was conducted in an environment designed to meet the task requirements outlined in the previous section. Participants sat at a central table while receiving instructions on using the tool and performing the work phase (Fig.3). To prevent communication during the work phase, a partition was placed between the participants.

During the transmission phase, participants stood at marked positions on the floor to engage in communication with each other (Fig.4). A single camera was positioned to capture each participant frontally, while wide-angle cameras were placed diagonally in front of each participant to capture a broader view, including facial expressions and gestures.

The camera placement was carefully designed to ensure clear visibility of non-verbal communication without hindering the participants' interaction.

4.2 Experimental Procedure

Participants first received an explanation of the experiment and provided their consent. The experiment was conducted using a tool running on a workstation PC, which displayed 10 images in a web browser interface (Fig.5 and 6). Participants could rearrange the images by dragging and dropping them with the mouse. Each action was transmitted to a server for recording. The workstation display continuously showed the current phase, whether operations were permitted, and the remaining time at the bottom of the screen.

During the transmission phase, the images were concealed to prevent viewing, while in the work phase, the images were displayed, allowing participants to reorder them. Following the task design described earlier, each set of images involved five repetitions of a 2-minute transmission phase and a 1-minute work phase.

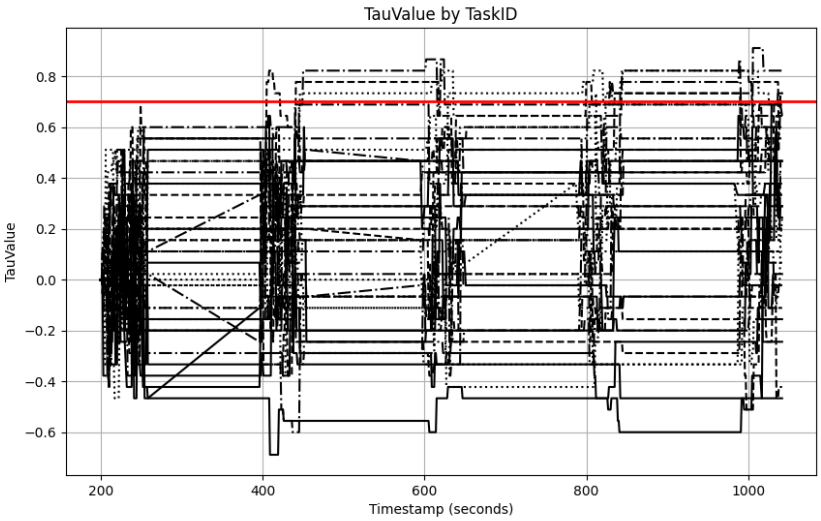

Figure 7: The changes in the correlation coefficient

The images used were derived from approximately 1-minute scenes from films like “City Lights," with 20 images obtained from each video, resulting in four distinct image sets. All videos used in the experiment were in the public domain, ensuring no copyright issues under Japanese and U.S. regulations.

5 Analysis of Collected Corpus

This section will describe how the dialogue data collected from the experiment was analyzed to examine the process of building common ground. Evaluation was performed based on the common ground construction process using information recorded during each phase.

The experimental tool recorded the movement of images and the order of images at each time point. Numbers were assigned to the images on the tool according to their chronological order. Kendall's rank correlation coefficient is calculated based on whether the image arrangements between participant pairs are consistent or inconsistent with each other. This coefficient ranges from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation. In this study, the objective is not to evaluate whether the sequence of images follows a chronological order but rather to assess the extent to which a shared foundation is accurately constructed through participant interactions. Therefore, the coefficient, which indicates the degree of agreement in image arrangement among participants, is treated as an index of the shared understanding construction process.

work phases	1st	2nd	3rd	4th	5th
M	0.12	0.15	0.19	0.26	0.28
SD	0.31	0.37	0.36	0.35	0.36

Table 1: The change of the coefficients

This number was calculated each time participants rearranged images, and the change in value at the end of the work phase was considered as the change in the ground constructed through transmission.

During transmission, cameras set between participants and behind them recorded facial expressions, gestures, and dialogue content during the experiment.

The Kendall rank correlation coefficient was calculated for each participant pair's work, and changes in the correlation coefficient over time were recorded to analyze the building common ground process (Fig. 7).

To analyze the construction process of the common ground through repeated work phases, we summarize the statistical information on the coefficients at the end of the five work phases (Table 1). From the results showing an increase in similarity with each phase, we can see that the process of building a Common Ground was successfully recorded.

By separating the work and transmission phases in the task design, we were able to record the com-

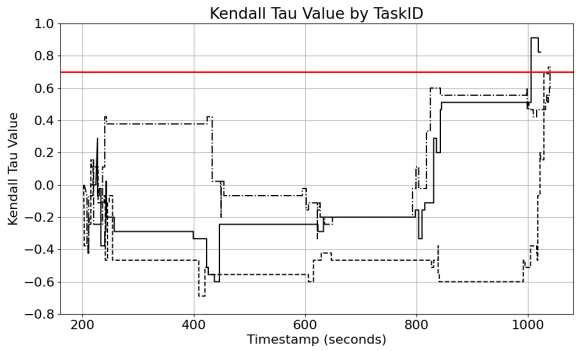

Figure 8: The result of successful sessions

mon ground construction process during communication.

We analyze gestures that contribute to the construction based on the final values for each session. When sessions with a final similarity of the arrangement order between the pairs of participants exceeding 0.7 were considered successful, there were 3 successful sessions (Fig. 8).

6 Clustering the grounding process

Clustering techniques are used to clarify the typical process of grounding in the collected data. The results of clustering using hierarchical clustering, a method for clustering time-series data, are shown. Hierarchical clustering is suitable for certain types of time-series data, particularly when the data is represented as fixed-length vectors with fully aligned time steps across all samples.

The results of the clustering are illustrated in Figure 9. Given the consistency of similar clusters when increasing the number of clusters, we classified the data into four clusters. The vertical axis represents the Kendall rank correlation coefficient calculated for each pair of participants' tasks. The horizontal axis corresponds to the step numbers associated with the beginning and end of the five work phases.

To delve deeper into the content of dialogues within each cluster, we sampled several conversations from each classified group and analyzed them in relation to their dialogue content.

Cluster 1: This pattern shows significant progress in grounding common ground early in the task. In these cases, participants tended to share the overall flow and key features using physical expressions, facilitating the grounding process early on.

Figure 9: The result of successful sessions

Cluster 2: This pattern exhibits steady, average progress in grounding common ground throughout the task. In these instances, participants often identified distinctive characters or motifs from the scenes, progressively building the common ground. Some pairs demonstrated a dynamic where one participant led the direction of the flow, while the other followed, contributing to the grounding process.

Cluster 3: This pattern indicates a general difficulty in establishing common ground. Here, participants struggled to share information and find common elements, leading to unsuccessful grounding attempts.

Cluster 4: In this pattern, significant progress in grounding common ground occurred towards the end of the task. Participants successfully shared information over time, leading to successful grounding by the task's conclusion. In some cases, pairs initially built smaller subgroups by sharing parts of the scene flow, then combined these to establish the overall flow.

By classifying the data into clusters based on the grounding process, we identified distinct patterns in how common ground develops, offering insights into the dynamics of successful and unsuccessful grounding scenarios.

From the video recordings of phases where significant progress in grounding common ground was observed, the following gestures were noted. These gestures are considered to be strong contributing factors to common ground construction:

Video Imitation Transmission A method of mimicking specific people or situations shown in a video using hands or body movements. This approach expresses the state or condition of objects in the video through bodily movements, making it easier to visually understand the other person's state or emotions.

Structure Expression Transmission A method of using hands or arms to show the spatial structure of the scene or the flow of time in a video. For example, it can be used to convey the position of objects in a spatial arrangement or to express chronological order. This method is particularly effective when conveying spatial or temporal information.

Imitation Agreement Transmission A method of showing agreement or understanding by mimicking the other person's actions or gestures. By repeating the other person's movements, physical expressions are used to convey understanding or agreement with a statement. This approach may emphasize empathy or cooperation within communication.

Other Cultural Gestures A method of expressing emotions or states through body movements or gestures used in specific cultures. These can include signs of hesitation, agreement, or requests for clarification during communication.

7 Summary and Future Directions

In this study, we proposed a novel experimental task, the "Collaborative Scene Reordering Task," to analyze the process of grounding common ground in human communication, with a particular focus on the impact of non-verbal modalities.

By separating the transmission and work phases within the task, we were able to observe the effects of physicality more clearly and analyze both the successes and failures in the grounding process. We established a method for quantitatively evaluating the grounding process over time using the Kendall rank correlation coefficient. Clustering was performed based on the grounding process, allowing us to analyze the tendencies in how grounding progresses. Furthermore, the study suggested that specific gestures might strongly contribute to the grounding of common ground, affirming the importance of non-verbal communication.

These findings not only deepen our understanding of human communication but also suggest potential applications in designing more natural interactions between humans and agents.

However, this study has the following limitations.

The sample size was small, with a gender imbalance among participants. The task was designed to observe grounding in specific contexts, and caution is needed when generalizing the findings. Long-term effects and cultural factors were not considered. Due to technical constraints, some non-verbal behaviors may not have been fully captured. Future research should include larger and more diverse samples, cross-cultural validation, and investigation of long-term effects.

A more detailed analysis of gestures and dialogue content during the interaction will be conducted. By observing the frequency and timing of gestures during the dialogue and performing a quantitative analysis, the aim is to clarify the factors that influence the construction of a shared foundation. Furthermore, the current study focuses solely on the physical expressions in the dialogue, without analyzing the content of the dialogue itself. Future analysis will include the relationship between dialogue content and gestures, their impact on the construction process, and the effects of different progression strategies during the task.

This study provides new insights into the role of non-verbal communication in grounding and makes significant contributions to the fields of HCI, the implementation of smoother dialogue systems, and communication research.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Luciana Benotti and Patrick Rowan Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 515–531. Association for Computational Linguistics.

Dana R Carney and Jinni A Harrigan. 2003. It takes one to know one: interpersonal sensitivity is related to accurate assessments of others' interpersonal sensitivity. *Emotion*, 3(2):194.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Paul Ekman and Wallace V. Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98.

Yuki Furuya, Koki Saito, Kosuke Ogura, Koh Mitsuda, Ryuichiro Higashinaka, and Kazunori Takashio. 2022. Dialogue corpus construction considering modality and social relationships in building common ground. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4088–4095.

Akira Ichikawa, Yasuo Horiuchi, and Shun Tsuchiya. 2000. The Japanese map task dialogue corpus. *Journal of the Phonetic Society of Japan*, 4(2):4–15. (In Japanese).

Adam Kendon. 1983. Gesture and speech: How they interact. *Nonverbal interaction*, 11:13–45.

Michael Kipp. 2005. *Gesture generation by imitation: from human behavior to computer character animation*. Universal-Publishers.

Robert M Krauss, Yihsiu Chen, and Rebecca F Gotfexnum. 2000. 13 lexical gestures and lexical access: a process model. *Language and gesture*, 2:261.

David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Koh Mitsuda, Ryuichiro Higashinaka, Yuhei Ohga, and Tetsuya Kinebuchi. 2021. Analysis of common ground building process in dialogue for collaborative figure placement task. In *Proceedings of the 27th Annual Meeting of the Association for Natural Language Processing*, pages 1698–1701.

Junya Morita, Tatsuya Yui, Takeru Amaya, Ryuichiro Higashinaka, and Yugo Takeuchi. 2024. Cognitive architecture toward common ground sharing among humans and generative AIs: Trial modeling on model-model interaction in tangram naming task. *Proceedings of the AAAI Symposium Series*, 2(1):349–355.

Mikio Nakano. 2019. Grounding process in dialogue systems. In *Proceedings of the 86th Special Interest Group on Spoken Language Understanding and Dialogue Processing*, pages 1–4. Japanese Society for Artificial Intelligence.

Mikio Nakano, Kazunori Komatani, and Kotaro Funakoshi. 2015. *Dialogue Systems*, volume 7 of *Natural Language Processing Series*. Corona Publishing. (In Japanese).

Dalmas A Taylor. 1968. The development of interpersonal relationships: Social penetration processes. *The Journal of Social Psychology*, 75(1):79–90.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.