SPOTTER: A Framework for Investigating Convention Formation in a Visually Grounded Human-Robot Reference Task

Jaap Kruijt, Peggy van Minkelen, Lucia Donatelli, Piek Vossen, Elly Konijn, Thomas Baier
Vrije Universiteit Amsterdam
{j.m.kruijt, p.van.minkelen, l.e.donatelli, p.t.j.m.vossen, elly.konijn, t.baier}@vu.nl

Abstract

Linguistic conventions that arise in dialogue reflect common ground and can increase communicative efficiency. Social robots that can understand these conventions and the process by which they arise have the potential to become efficient communication partners. Nevertheless, it is unclear how robots can engage in convention formation when presented with both familiar and new information. We introduce an adaptable game framework, **SPOTTER**, to study the dynamics of convention formation for visually grounded referring expressions in both human-human and human-robot interaction. Specifically, we seek to elicit convention forming for members of an *inner circle* of well-known individuals in the common ground, as opposed to individuals from an *outer circle*, who are unfamiliar. We release an initial corpus of 5000 utterances from two exploratory pilot experiments in spoken Dutch. Different from previous work focussing on human-human interaction, we find that referring expressions for both familiar and unfamiliar individuals maintain their length throughout human-robot interaction. Stable conventions are formed, although these conventions can be impacted by distracting outer circle individuals. With our distinction between familiar and unfamiliar, we create a contrastive operationalization of common ground, which aids research into convention formation.

Keywords: Referring expressions, convention formation, reference games, human-robot interaction

1. Introduction

When two humans interact with each other for a longer period of time, they build up common ground (Stalnaker, 2002). As common ground increases, they form conventions on how they refer to objects or people that occur frequently in their experiences and conversations. These conventions can be very personal (Hawkins et al., 2023) and difficult to understand for outsiders: a mutual friend being called 'de kale' (*the bald one*) requires knowledge of shared experiences to interpret correctly. Various research has shown that conventions arise in repeated interactions over the same task (Clark and Wilkes-Gibbs, 1986; Hawkins et al., 2017; Haber et al., 2019), leading to a decrease in utterance length while maintaining informative content (Hawkins et al., 2020; Giulianelli et al., 2021). However, new information presented in a conversation may also challenge the established common ground and conventions, when the information proves incongruent or leads to ambiguity.

While much research has focused on the development of conventions in relatively static settings, it is not clear how these conventions influence and are influenced by incoming new information. This is especially relevant for social conversation, where common ground is continually updated to incorporate new information. As social non-human agents start to play larger roles in our daily and social life, this also raises the question of how these agents can reason within this complicated changing common ground, and deal with the potential ambiguity that arises within it.

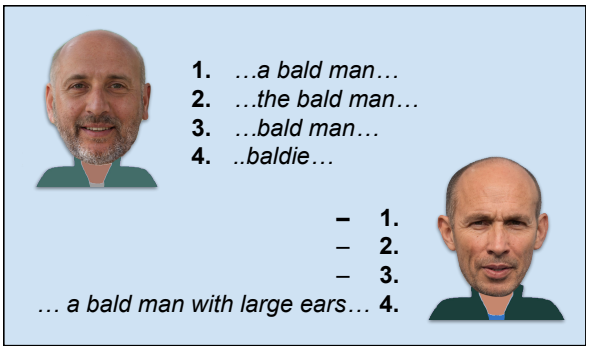

Figure 1: An example of how conventions and ambiguity could influence each other in SPOTTER. The second bald OutC character is only introduced in round 4, when a convention already exists for the first InC bald character.

To investigate these questions, we develop **SPOTTER** (Shared Picture Observation Task for Testing Entity References), a gameplay framework for task-based interaction (Sec. 3). SPOTTER is an interactive task in which two players communicate about a visual scene, like a game of 'spot the difference'. This game is inspired by repeated reference tasks (Krauss and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986) (Sec. 2), where players repeat the same communicative task over a number of rounds. In our task, the goal for the players

LREC-COLING 2024, pages 15202–15215
20-25 May, 2024.

is to identify the location of characters in a group picture through verbal communication. Characters from a known *inner circle* (InC) recur throughout the game; these contrast with unknown *outer circle* (OutC) characters, who occur only once or twice throughout the various rounds of the game. Importantly, all characters have a few salient visual attributes (e.g. glasses or a beard) which may be shared across circles and create ambiguity in creating and interpreting referring expressions. Our overall gameplay design allows us to study (i) how referring expressions to the InC characters develop over time compared to references to the OutC as common ground grows and (ii) how conventions arise (see Figure 1 for an example).

To demonstrate the utility of our gameplay design, we additionally present a multimodal corpus of two experiments in Dutch (Sec 4) using the SPOTTER framework (Sec. 5). Our corpus spans two experiments using a Wizard-of-Oz (WoZ) set-up to simulate human-robot interaction. The corpus consists of about 5000 utterances across 21 participants and is aligned with character visuals. We annotate spans that correspond to mentions of InC and OutC characters, as well as transaction units to facilitate analysis of how referring expressions and convention formation occur through specific dialogue acts. Analysis of our corpus shows that our results differ from previous studies on convention formation in human-human interaction. Specifically, we do not observe a reduction in the length of repeated references; we also observe an important effect of new information as embodied by OutC characters on InC references. We discuss these findings, and their implications for research on convention formation in human-robot interaction, in detail (Sec. 6).

Our main contributions are summarized as follows:

- We develop a new framework for investigating referential expressions which includes a distinction between known 'inner circle' (InC) referents and unknown 'outer circle' (OutC) referents. This allows for a more contrastive analysis of how common ground develops in conversation than has been done in previous work;
- We release a multi-layered annotated multimodal corpus of Dutch task-based interactions between humans and acted robots in real-life settings using Wizard of Oz;
- We provide an in-depth analysis of the interactions and the development of referring expressions during our task, showing how our findings differ from previous work and pointing out the need for additional work in this research space.

2. Related Work

Repeated Reference Tasks First designed by Krauss and Weinheimer (1964) and refined by Clark and Wilkes-Gibbs (1986), repeated reference tasks have been used to study convention formation in interaction. In these tasks, players communicate about a set of images which they have to match with the other player. The same images return over a number of rounds, allowing for multiple references to the same image. While references to these images usually start out long and descriptive, they become shorter over time as participants form a convention together in how they refer to this image. Repeated reference tasks exist in various forms, such as in the Photobook task (Haber et al., 2019) and in the work of Mankewitz et al. (2021).

The dynamics of convention formation and the process of utterance length reduction in these various versions of repeated reference tasks have been analysed by Hawkins et al. (2017), Hawkins et al. (2020), Giulianelli et al. (2021) and Boyce and Frank (2023). Hawkins et al. (2020) and Boyce and Frank (2023) provide a syntactic and semantic analysis of convention formation, and Giulianelli et al. (2021) and Hawkins et al. (2017) analyse the pragmatic and information-theoretic underpinnings of convention formation. However, these findings are based on human-human interaction studies only, and to our knowledge, it is not known if convention formation happens in similar ways in HRI.

Visually grounded dialogue Repeated reference tasks fall within a broader category of visually grounded dialogue tasks. One such visually grounded task is GuessWhat!? (de Vries et al., 2017), a guessing game for two players where the goal is to locate an object among other objects with visually similar features acting as distractors. Our task also includes visual distractors which introduce ambiguity. However, in our case the dialogue is not structured around yes/no questions, allowing more free-form dialogue. Furthermore, our human-like characters used as visual targets are more likely to become familiar than the objects used in GuessWhat.

Datasets of Human-Robot Interaction Another aspect in which our game differs from existing repeated reference games is that it is designed to study human-robot interaction as well as human-human interaction. The dataset which we release from our two pilot field experiments will contain human-robot dialogue rather than human-human dialogue, although in this case the robot behavior was achieved using Wizard of Oz (WoZ, (Kelley, 1984; Riek, 2012). In this paradigm, the speech and actions by the robot are provided at the right

moment by a hidden human programmer. In this way, our dataset shares some similarities with the dataset by Traum et al. (2018) who also used WoZ for their multi-party human-robot interaction task (Bonial et al., 2017). We use the transaction unit structure designed by Traum et al. (2018) for our data annotation. The robot in the current study functioned via WoZ to allow us to gather data which we can use to design an autonomous robot. A follow-up study with this autonomous robot is currently being prepared.

Inner and Outer Circle In human-human interaction, the presence and importance of a particular individual within the common ground influences how readily and the manner in which that individual can be introduced into the conversation. Here, we build upon the idea of familiar *inner* and unfamiliar *outer* circles in designing our repeated reference game (Kruijt and Vossen, 2022).

In SPOTTER, the inner/outer circle (InC/OutC) distinction makes more explicit what is and what is not part of the common ground compared to existing work. It allows us to study how references to the InC develop within a common ground which is built up as the surrounding context changes. This changing context is more representative of real-life communication, which is also addressed by Udagawa and Aizawa (2021). Investigating common ground in a changing context is therefore an important step in the development of social robots. In our case, the OutC creates this changing context. Players need to contrast that which they deem shared knowledge with information which is not yet shared. For artificial agents, this adds an extra challenge: the agent needs to infer what is part of the common ground and what is not and develop different strategies for interpreting both cases.

3. Task Description

3.1. Research Goals

Like in related repeated reference games, the primary purpose of our task is to study how references change and conventions form as common ground increases. Our InC/OutC distinction adds an additional layer to this analysis, allowing us to investigate the buildup of common ground within changing contexts. For this, we compare how the average length of references to the InC and OutC develop. Finally, in this paper we implement the repeated reference task in an HRI setting, which to our knowledge has not been done before. With this setting we aim to explore how humans develop conventions with a robot interlocutor. In summary, in this paper we examine two questions:

Q1 How does convention formation in Human-Robot Interaction compare to convention formation in related work in Human-Human Interaction?

Q2 How do references to the InC and OutC change over time as common ground is built up while the surrounding context changes?

We expect that conventions will form for the InC as participants build up common ground with the robot, similar to studies in human-human interaction (see Section 2. Based on these studies, we also expect that the references for the InC decrease in length as conventions are formed. However, we expect that the references to the OutC do not decrease in length as no conventions form for this group. Rather, we expect that their references become *longer* over time as participants need to emphasise the differences with the InC to avoid ambiguity. Though we expect that conventions develop in the same way as in Human-Human interaction, it may be that humans change their strategies due to the presence of the robot interlocutor. The lack of a prior repeated reference study within HRI makes this work exploratory.

3.2. Formalization

Figure 2 shows what our game looks like. For this game, we created 15 images of **characters** $c_1...c_{15} \in C$ and 6 images of **contexts** $x_1...x_6 \in X$. The characters are divided into 3 **main characters** $c_1...c_3$ and 12 **side characters** $c_4...c_{15}$. Main characters correspond to the InC, whereas side characters belong to the OutC. Each character has a set of visual **attributes** $a_1...a_n \in A_i \models c_i$ with which characters can be described. Character attributes may partially overlap. While the formal set of attributes is finite, there are potentially infinite ways to describe the characters, meaning that there are more attributes possible which are not formalized.

A game consists of 6 **rounds** $r_1...r_6 \in R$. Each round depicts a different context from X displayed as a background image against which 5 characters from C appear. The background context allows for an additional **context-attribute relation** (a_i, x_i), which is an attribute induced by the context. The three main characters are present in every round, as well as two characters from the set of side characters. These side characters are selected in such a way that they appear only once or twice during the game, and never in subsequent rounds. The characters are lined up next to each other in the image in a certain order. A round contains two **scenes** (S_1, S_2). One player sees S_1, while the other player sees S_2. The only difference between the scenes is the order in which the characters appear. For instance, for one player the order is

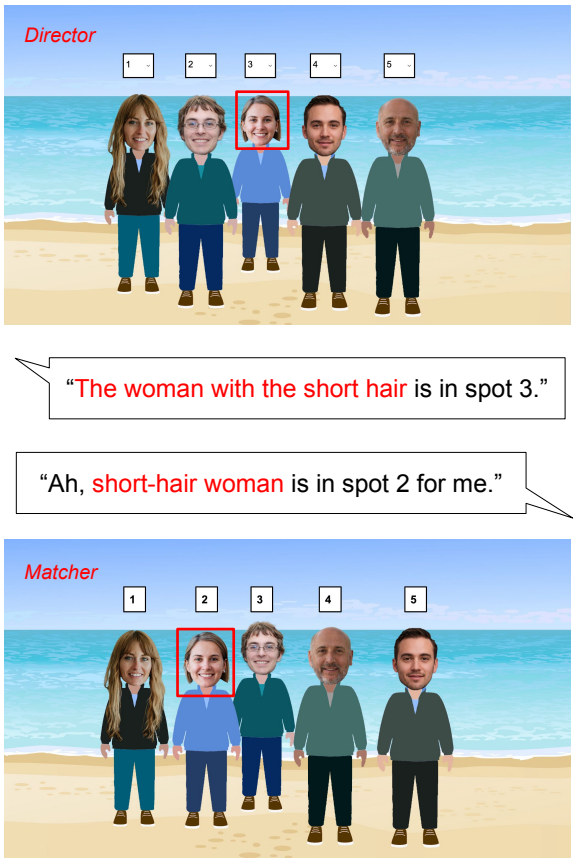

Figure 2: An example of the game view (red boxes and player roles added for clarity and not shown during experiments). The director gives a description (in red) of one of the characters (in the red box) and the matcher finds the position of this character and responds accordingly.

$(c_2, c_5, c_3, c_1, c_{14})$, while for the other player the order is $(c_2, c_3, c_{14}, c_5, c_1)$.

The goal for each round is to figure out the differences in the order of characters in their scenes through verbal communication. One player takes on the role of **director**, describing the characters and their position, while the other player takes on the role of **matcher**, matching the characters and their position in her own scene (cf. Clark and Wilkes-Gibbs (1986)). In case of a human-human game, players can switch the roles of matcher and director between rounds. In case of our human-robot game, the human is the director, and the robot is the matcher. Players are encouraged to mention the characters in the order that they appear from left to right by way of numbers above the heads.

Each round will contain one or more **mentions** $m_1 ... m_i \in M$ (cf. Fokkens et al. (2013)) for each character by the director. The matcher tries to identify the corresponding character in their own scene based on the mention, which we call a character-mention **mapping** (c_i, m_j). The interactive process of describing and mapping one character forms a **transaction unit (TU)** (cf. Traum et al. (2018)). Transaction units consist of at least two **turns**, one turn in which the first player provides a mention and a position for a character, and one turn in which the second player provides a response with that character's position in their scene. However, longer transaction units are possible, for instance when a correction or repair is required. A round finishes when players have filled in the mappings for all the characters. Players are scored on the amount of correct mappings. At the beginning of the game, the main characters are introduced as part of the introduction to the game. In the human-robot version, the robot mentions that these characters are its 'friends'. As part of the introduction, the players play a practice round of the game (which we discarded from analyses).

3.3. Design Motivation

Because of the overlap in attributes between InC and OutC characters, we can investigate how ambiguity influences mentions and their interpretation. For instance, if one InC character has been established as 'the bald guy', another OutC character who is bald should be described in a different, more elaborate way to avoid confusion. This should then be reflected in the length of their respective mentions.

The context-attribute relation additionally creates the potential for using mentions based on contextual associations outside of the direct visual information. For instance, one scene is called 'at the family reunion', and it contains a small child and an elderly woman as side characters. The idea behind this is to evoke associations between family relations, which could influence mentions and provide a new way of describing a character. This new reference could potentially impact the convention.

3.4. Material

All the visual material for this game was designed using Adobe Photoshop. The background images for the contexts were obtained from Freepik[1]. We selected 6 contexts which are clearly identifiable, such as a beach, a fairground and a park. The backgrounds have a cartoon-like appeareance. For our first pilot field study, the characters also had a cartoon-like face. For our second pilot, we generated 'real' faces using This Person Does Not Exist[2], an online AI random face generator which is based on Nvidia's StyleGAN (Karras et al., 2019). This was done to make the faces appear more unique

[1] www.freepik.com

[2] https://this-person-does-not-exist.com/en

than they had been in the cartoon-like variant. Between the two studies, we kept the sets of character attributes as stable as the random face generator would allow us. The size of the character's faces was large compared to their body and the rest of their body was made to be hard to distinguish. This was done so that players would be encouraged to focus on facial features. The game is designed as a web application in HTML and Javascript to work on a computer or tablet. It can be hosted on an internal server.

4. Experiments

We conducted two Dutch-language pilot studies to test the game design and to obtain first results on how people play the game, interact with the robot and make reference to the characters we designed. Through the use of a WoZ set-up, we simulated the language and behavior of a robot that is aware of the common ground but not an active participant in the convention formation process. The pilot studies are described in this section and the next. They are referred to as Ex1 for Experiment 1 and Ex2 for Experiment 2.

4.1. Experiment 1

4.1.1. Implementation

In our first experiment, the game design differed in a few ways from the final design described in section 3. The most important difference is that we did not encourage participants to mention the characters from left to right using numbers. Instead, they were free to mention the characters in any order they liked. Because there were no numbered positions, the goal for each round was also slightly different: instead of locating the exact position, the robot matcher only decided whether the character the participant mentioned was in the *same* or a *different* position in its scene, which the participant could then fill in. The left-to-right order was added in the final design to make the conversation more streamlined. As mentioned in section 3.4, the characters had a cartoon-like appearance which contained less detail and were less unique than the real faces in the final design.

4.1.2. Material & Procedure

The robot used for the experiment is the NAO V6 robot built by Softbank robotics. The web application was integrated into Baier et al. (2022)'s modular architecture and used the EMISSOR platform (Baez Santamaria et al., 2021) to capture and store the interaction. This architecture was also used as a basis for the WoZ software which controlled the robot's verbal behavior. The WoZ software and the web application were run on the same internal server.

The interactions took place in a closed cubicle to ensure privacy and concentration for the participant. The WoZ input was provided from an adjacent closed cubicle by a researcher. The web application running the game was presented to participants on a tablet at the. The robot was standing in front of the participant and the tablet was presented to the side of the robot. Both the robot's head and the tablet were at the height of the participant's eyesight to facilitate smooth transition from the robot to the tablet and back. After having provided active informed consent, video and audio of the participant and the interaction were captured for analysis and for monitoring of the interaction by the researcher providing the WoZ input.

4.1.3. Participants

The experiment was conducted at a Dutch children's science fair. Adults and children participated in pairs, in separate cubicles. The children played an easier version of the game which did not include side characters. In total, 12 adult-child pairs participated. The study was approved by the Institutional Ethical Review Board. Participants signed an informed consent form including options agreeing to their video and audio being captured. After the first and last round of the game, participants filled out a questionnaire which is used for a parallel Media Psychological study examining human perception of and relationship-building with their robot interlocutor and during this game. Data from both experiments were combined for analyses. Therefore, the data are analyzed together in section 5.2.

4.2. Experiment 2

Based on the experience and findings of Ex1, we modified the game design and conducted a second pilot field experiment to test our modifications. The adjusted implementation for this experiment is as described in section 3. The material and procedure is the same as in Ex1. This experiment was also performed at a Dutch children's fair. 18 adult-child pairs participated in this experiment.

5. Dataset

5.1. Dataset Creation

We used the video and audio captured during both experiments to create a dataset of human-robot interactions while playing this game. We only used the data from the adults for this dataset. We excluded 9 participants (5 from Ex1, 4 from Ex2) since they failed to complete all six rounds properly, resulting in 21 participants which we used for analyses.

	Ex1	Ex2
	# in utterances	
# Utterances	1646	3332
# Mentions	304	567
# Turns	846	1666
Mdn TU length	2	3
# Repair TU's	29	51
	# in TU's	
Mdn # TU's	34	31
Top 2 DA's	statement, pos_answer	statement, back-channeling

Table 1: Details of the SPOTTER dataset of our 2 pilot studies. '#' means 'amount'. TU = Transaction unit, Mdn = Median, DA = Dialog Act.

Parts of the experiment where the participant filled out the questionnaires were cut from the video to obtain just the human-robot interaction of the game.

The audio from the videos was transcribed in Dutch using OpenAI's Whisper (Radford et al., 2023). We combined the transcription with the utterances produced by the robot which were stored in EMISSOR to obtain the speaker and timestamps for each utterance. Errors in speaker assignment were corrected by hand. The data were hand-annotated for mention spans and the character that was mapped to the mention. The mention can be a part of the utterance (such as '*the woman with the short hair* is in spot 3'), or the entire utterance can be the mention. We also annotated round numbers, transaction units and the relations between TU's by hand. Transaction unit relations were obtained from Traum et al. (2018), but we added four relations to better match our interaction data: 'response-description' and 'continue-description', respectively for replying to a question with a mention and continuing that mention in the next utterance, and 'req-confirm', 'clar-confirm' and 'ack-confirm' for requesting, providing or acknowledging the confirmation of a character position. The utterances were also automatically annotated with dialog acts using XLM-RoBERTa (Conneau et al., 2021) fine-tuned on the MIDAS dataset of dialog acts (Yu and Yu, 2021). Details of our dataset can be found in Table 1 and an example section in Appendix A. We release the data in a Github repository together with the code for the framework free for research [3].

5.2. Results

Convention formation To test if conventions were formed for the InC over time as we expected, we measure the stability of mentions to the InC in terms of semantic content. Similar to the analysis in Boyce and Frank (2023), we use the multilingual SentenceBERT (Reimers and Gurevych, 2020) to calculate cosine similarity for inner circle character mentions between each round and its preceding round. In Figure 3, we show the highest similarity score per inner circle character for every two subsequent rounds. Characters are numbered per experiment (1.{1,2,3} for Ex1 and 2.{1,2,3} for Ex2). On average, the mentions become and stay semantically similar after one or two rounds, with scores up to 0.9 for experiment one and up to 0.8 for experiment two. This reflects a relatively stable mention pattern, which suggests that conventions did indeed form. However, the results show considerable differences between characters in how their mentions develop. For instance, character 2.3 () in Figure 3 seems to have a very stable mention for the first 4 rounds, but in round 5, the similarity score suddenly drops, after which it recovers slightly. Character 1.3 seems to have a very stable mention across the whole game. These individual differences seem to suggest that the specific visual features of InC characters and the presence of OutC characters can have an impact on mentions. We will examine this further in section 6. We also note that there are differences across participants in how quickly they converge on a convention, as shown by the large error bars in the first three rounds. These error bars are considerably smaller in the latter rounds, suggesting that by then most participants reached a stable convention.

Since we expected the convention formation to lead to a decrease in utterance length for the InC, we measure the average utterance length of a mention per round. We do this both for the inner circle and outer circle. Figure 4 shows the mean utterance length in words for inner and outer circle characters per round per experiment. The results show no clear decrease in utterance length, neither for the InC where we expected it, nor for the OutC where we did not expect such a decrease.

Inner vs. Outer Circle As shown in Figure 4, utterance lengths are generally longer for OutC mentions than for InC mentions, with Ex1 rounds 1 and 3 and Ex2 rounds 2 and 3 being the exceptions to this rule. According to our hypothesis for Q2, outer circle characters that are introduced later in the game would require longer references than InC characters, because the more straightforward and more efficient description has already been taken up by the convention formed to describe an InC character with similar features. Our results suggest that this is indeed the case. During the first round, the differences in utterance lengths between inner and outer circle mentions are not so

[3] https://cltl.github.io/docs/projects/spotter

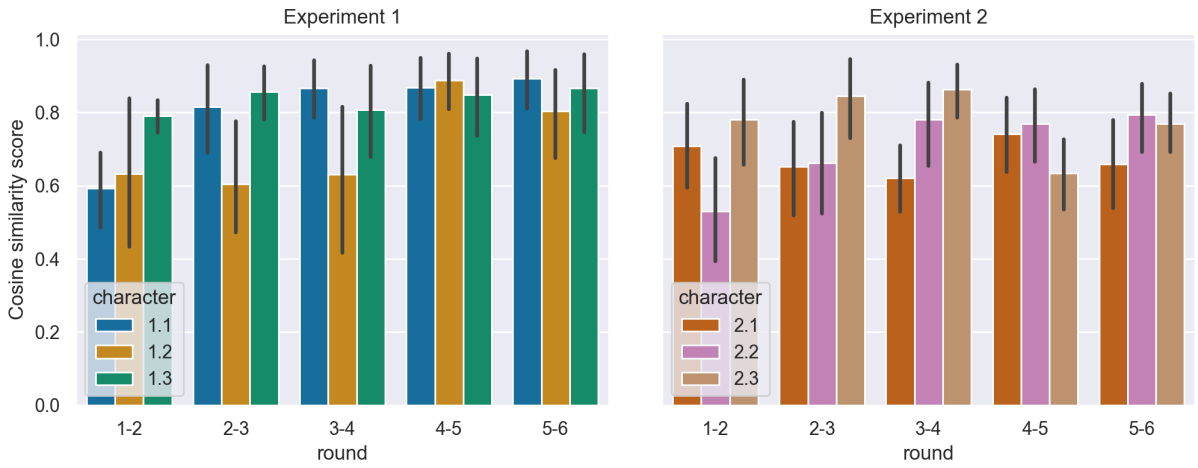

Figure 3: Cosine similarity scores between rounds for inner circle mentions. Similarity scores are calculated between each two rounds shown on the x-axis.

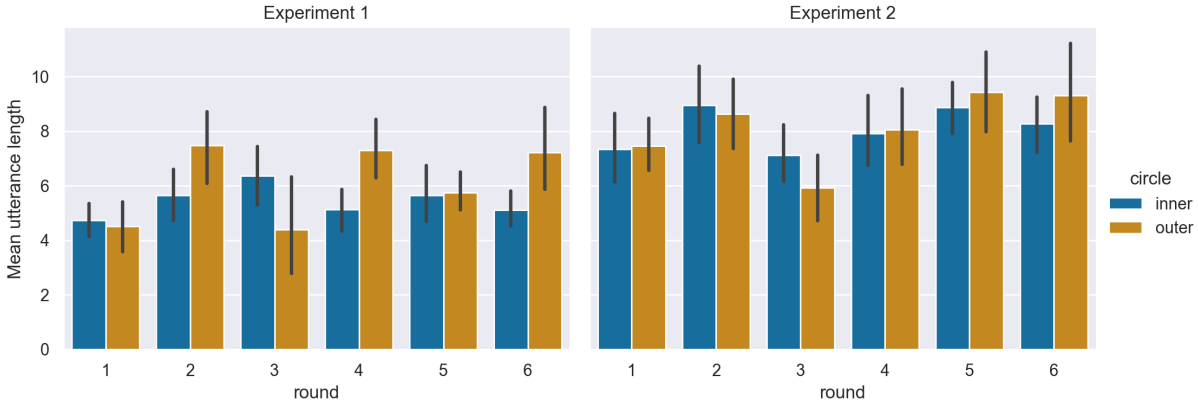

Figure 4: Average utterance lengths in words per round for mentions to inner and outer circle characters.

pronounced, but in this round no conventions can exist for the InC characters yet. For an analysis of the notable exception to the trend in round 3, where OutC references are much shorter, we refer to section 6.

Experiment 1 vs. Experiment 2 Figure 4 further shows that mentions are on average longer for Ex2 than for Ex1. One explanation for this could be that the cartoon faces used in Ex1 were visually less complex than the real-life faces used in Ex2. As a result, there would be more attributes which could be used to describe characters in Ex2. To examine this, we looked at references to the InC character who could be described as 'bald man'. In the first experiment, many mentions contained only this information, i.e. 'the/a bald man' (sometimes with an additional reference to his glasses). In the second experiment however, though some people also used 'the bald man', we also found many more elaborate descriptions such as 'a middle-aged man with a bald head' and 'man with the bald head, but with a small moustache and beard'.

Figure 5 shows the average number of mentions per character per round. The results show that characters are mentioned only once per round on average. However, the InC characters in round 1 and 2 are mentioned more frequently in Ex1. This could be an effect of the free order in which participants could mention the characters in Ex1 (rather than the 1-5 order we used in Ex2). We observed that the InC characters were sometimes used as 'anchor points' for relative descriptions of the other characters' locations (such as '*to the left of* the bald man'). In this case, they would occur in the transaction unit of the mapping process for another character in addition to their own mapping TU. To analyse this, we calculated the average number of transaction units in which characters were mentioned. We found that InC characters were mentioned in 1.25 TU's on average in rounds 1 and 2 in experiment 1. After round 2, this anchoring effect disappears. This is probably because participants resorted to a left-to-right order (as in experiment 2), eliminating the need for these anchor points. Why the OutC characters were mentioned more often

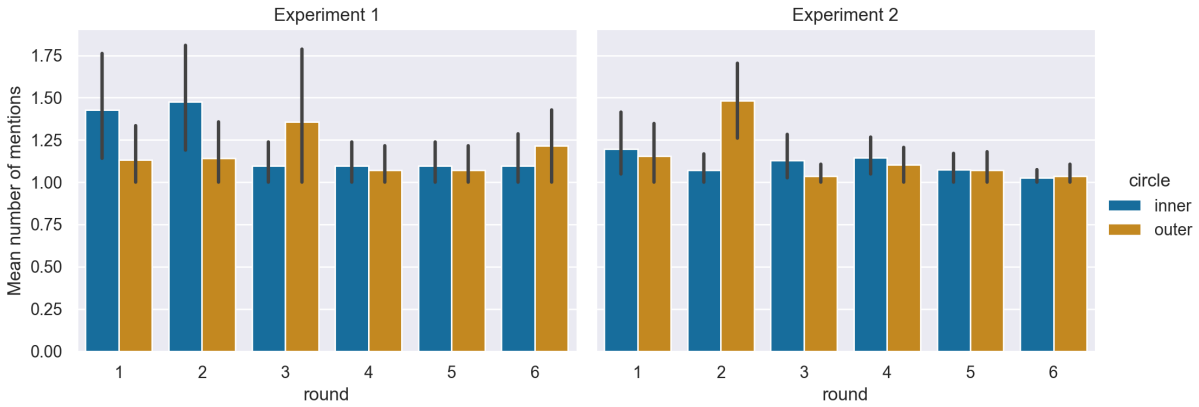

Figure 5: Average number of mentions to inner and outer circle characters per round.

in Ex1 round 3 and Ex2 round 2 does not become clear.

6. Discussion

Differences with human-human studies From our results and analyses, it becomes clear that the process of convention formation in our HRI experiments did not progress as should be expected based on findings in existing work on convention formation in human-human reference games (Q3) (Clark and Wilkes-Gibbs, 1986; Hawkins et al., 2020; Haber et al., 2019; Boyce and Frank, 2023). Specifically, we did not see a pattern of reduction in utterance length over time. Furthermore, the conventions seemed to be relatively unstable, being easily disrupted or changed by the presence of a distractor mention to an OutC character. It might be that the robot's affordances and perceived understanding played a role in the mentions participants used, for instance because participants thought that the robot was less intelligent and required clear and elaborate descriptions to understand the participant. This is being investigated in the parallel Media Psychological study.

The difference could also be explained by the explicit presence of distractors in our task, which impacts the convention formation process. Below, we analyze some specific features of our findings in more detail, to see how our game design influenced convention formation and referring expressions in general.

Inner vs. Outer Circle One of the overarching questions of this study is how the mentions to the inner and outer circle differ from each other. A closer look shows that participants used certain strategies to distinguish between the known InC characters and unfamiliar OutC characters. Some participants opted for 'our/your friend' versus 'not our/your friend', relying on the way in which the main characters were introduced at the start of the game. Another strategy often employed is that participants used 'another' for the OutC characters, e.g. 'the bald man' versus 'another bald man'. Lastly, the InC characters were often used as a comparison in mentions to the outer circle. This led to references such as 'a bald man *without* a beard', compared to 'the/a bald man' (*with* the beard). Sometimes, participants would very clearly signal the familiarity of InC characters and the novelty of OutC characters, for instance by using 'the one who we have been seeing all the time' versus 'we haven't seen this one yet'.

Nonetheless, the OutC characters seemed to have quite an impact on the way conventions were formed. Usually after the appearance of an OutC character with similar attributes, participants would be more specific in their description of the InC character with the same attributes. For example, one character was described as 'a boy with glasses' at first. However, after the appearance of another male character with glasses, the reference changed to include his hairstyle as well, making the reference longer rather than shorter. This shows an interesting effect of the changing context in which the common ground was built up (Q2).

Convention length and indefinite conventions In the previous section, we showed that mentions to the InC did not decrease in length during the game. The mean utterance length is also relatively high, at around 8 words for experiment 2. It seems that participants very often were as specific as possible, referring to most if not all salient attributes. However, when we observe the mentions within one participant, we find that participants very often used a specific, codified way of referring to that character, which could be considered a convention. For instance, they would use 'a woman with short blond hair and earrings who is smiling'. This structure remained the same throughout the

game. This example also shows another interesting observation, which is that many participants used indefinite articles for InC characters instead of definite articles, which typically suggests novelty or underspecification. The fact that participants continued to use indefinite articles seems to suggest that the article was part of their convention. Another explanation could be that the rounds in the game and the game itself were too short to grant the use of definite articles.

Salient visual features In section 5.2 we also remarked that there are individual differences between the characters in how their mentions develop. Here, we take a closer look at these individual characteristics. Some characters seemed to be more easily distinguishable than others. In Ex2, a bald man with a small beard seemed to be the most easily recognizable character. This is shown by the high similarity across his mentions in the first four rounds of the game (character 2.3 in the right graph in figure 3). Interestingly though, this similarity score drops in round 5. This coincides with the introduction of an OutC character who also has a bald head as his most defining feature. This seems to have impacted the reference to 'the bald man', requiring participants to give a more detailed or different description to distinguish the two. Thus, while we expected conventions for inner circle characters to impact the reference to outer circle characters, this effect also seemed to go the other way, with certain OutC characters affecting or disrupting the convention that was formed for InC characters.

Another effect that a very salient visual feature has on mentions can be seen in figure 4, where the mean utterance length for mentions to the OutC suddenly drops in round 3. This is likely due to the presence of a child as one of the OutC characters in this round. This child was very clearly distinguishable from everyone around it, and almost everyone used just 'a child' as a sufficient mention for this character. This is of course a very short and efficient description, and this likely brought down the mean utterance length for OutC mentions. In general, participants thus adjusted the length of an OutC mention depending on the ambiguity with InC characters.

7. Conclusion

SPOTTER is a novel task in the field of repeated reference tasks which allows for studying convention formation over time in the presence of distractor outer circle mentions. Its design is flexible, allowing it to be used for human-human as well as human-robot interaction, and both for online and in-person experiments. Our first implementation of SPOTTER in human-robot pilot field experiments gave unexpected and insightful results: conventions formed, but they remained long and descriptive rather than shortening in length as we expected. Whether this finding is due to our novel addition of the outer circle, and thus part of the game design, or due to the robot interlocutor remains to be examined in future work. Therefore, we plan to use SPOTTER online in a human-human experiment to analyse how humans communicate around the inner versus outer circle characters parallel to examining the effect of the robot's affordances in the current study. We will also test more human-robot interaction, for which we will use the data gathered in these experiments to design a language model for a robot that can play the game autonomously, without using WoZ. We will also test the effect of functional contexts (e.g. restaurant, school, family dinner) and longer interaction history on the conventions by making the context more explicit.

Limitations

Due to the WoZ-approach, we decided that the robot should not engage in the mention process as much. The robot simply incorporated a deictic reference to the previous mention into their response rather than using a more descriptive mention to refer to the character. This was done to avoid influencing the end result by the researchers: we wanted the participants themselves to shape the references. However, this possibly made the robot interlocutor less cooperative and reciprocal, perhaps adding to participants' impressions that the robot was not intelligent. The lack of a verbal acknowledgement of the mention or convention used could mean that it does not evolve in the same way as was observed in previous work in human-human interaction, where the convention formation was more collaborative.

Another possible limitation is the length of the game. The game has only six rounds and is played only once between a human-robot pair. Convention formation may be stronger if the game is played multiple times between the same players with the same inner circle.

Acknowledgements

We would like to thank the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) for providing this project with a research grant under the ID 406.DI.19.005. We also thank the NWO, the Nationale Wetenschapsagenda (NWA) and The Network Institute for additional funding, Bianca Pander at BKB, ExpeditieNext Middelburg and Wesly Struik at Alles Kids in Drenthe for providing locations, and Mara Polak, Nina van Gulik and Lieke Hoorn for assistance.

Bibliographical References

Selene Baez Santamaria, Thomas Baier, Taewoon Kim, Lea Krause, Jaap Kruijt, and Piek Vossen. 2021. EMISSOR: A platform for capturing multimodal interactions as episodic memories and interpretations with situated scenario-based ontological references. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 56–77, Groningen, Netherlands (Online). Association for Computational Linguistics.

Thomas Baier, Selene Baez Santamaria, and Piek Vossen. 2022. A modular architecture for creating multimodal agents. *arXiv preprint arXiv:2206.00636*.

Claire Bonial, Matthew Marge, Ashley Foots, Felix Gervits, Cory J Hayes, Cassidy Henry, Susan G Hill, Anton Leuski, Stephanie M Lukin, Pooja Moolchandani, et al. 2017. Laying down the yellow brick road: Development of a wizard-of-oz interface for collecting human-robot dialogue. *AAAI Fall Symposium*.

Veronica Boyce and Michael C Frank. 2023. Communicative reduction in referring expressions within a multi-player negotiation game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition*, pages 127–149.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Interspeech 2021*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Antske Fokkens, Marieke Van Erp, Piek Vossen, Sara Tonelli, Willem Robert van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. 2013. Gaf: A grounded annotation framework for events. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 11–20.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Robert D. Hawkins, Michael C. Frank, and Noah D. Goodman. 2020. Characterizing the dynamics of learning in repeated reference games. *Cognitive Science*, 44(6):e12845.

Robert D Hawkins, Michael Franke, Michael C Frank, Adele E Goldberg, Kenny Smith, Thomas L Griffiths, and Noah D Goodman. 2023. From partners to populations: A hierarchical bayesian account of coordination and convention. *Psychological Review*, 130(4):977.

Robert XD Hawkins, Mike Frank, and Noah D Goodman. 2017. Convention-formation in iterated reference games. In *CogSci*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Robert M Krauss and Sidney Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1:113–114.

Jaap Kruijt and Piek Vossen. 2022. The role of common ground for referential expressions in social dialogues. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 99–110, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Jess Mankewitz, Veronica Boyce, Brandon Waldon, Georgia Loukatou, Dhara Yu, Jesse Mu, Noah D Goodman, and Michael C Frank. 2021. Multi-party referential communication in complex

strategic games. In *ICLR 2021 Meaning in Context Workshop*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Laurel D. Riek. 2012. Wizard of oz studies in hri: A systematic review and new reporting guidelines. *J. Hum.-Robot Interact.*, 1(1):119–136.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2022. Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–42, Dublin, Ireland. Association for Computational Linguistics.

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory Hayes, and Susan Hill. 2018. Dialogue structure annotation for multi-floor interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Takuma Udagawa and Akiko Aizawa. 2021. Maintaining Common Ground in Dynamic Environments. *Transactions of the Association for Computational Linguistics*, 9:995–1011.

Dian Yu and Zhou Yu. 2021. MIDAS: A dialog act annotation scheme for open domain HumanMachine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1103–1120, Online. Association for Computational Linguistics.

A. Dataset Example

#	Utterance	Speaker	Mention	CharID	Round	TU	TU-Rel	Dialog Act
263	Wie staat er nu op 1?	robot			5	22		open_ question _factual
264	Een meneer met een bril en een beetje langer haar bovenop.	human	Een meneer met een bril en een beetje langer haar bovenop.	2	5	22	answer-description	statement
265	Bij mij staat die op plek 5.	robot			5	22	ack-done	statement
266	En daarnaast staat een vrouw met lang haar.	human	een vrouw met lang haar.	11	5	23		statement
267	Die staat bij mij ook op die plek.	robot			5	23	ack-done	statement

Table 2: A section of our dataset. UttID = Utterance ID, CharID = character ID, TU = transaction unit, TU-Rel = transaction unit relation. Timestamps were omitted for readability.

B. List of Characters and Attributes

Char	Type	Sex	Age	HairColour	HairType	HairStyle	FacialHair	Accessory
1	Main	F	Young	Blonde	Straight	Bob	-	Earrings
2	Main	M	Young	Blonde	Curls	Short	-	Glasses
3	Main	M	-	Grey	-	Bald	Beard	-
4	Side	M	-	Brown	Curls	Short	-	Glasses
5	Side	M	-	Brown	-	Bald	Beard	-
6	Side	F	Young	Black	Straight	Bob	-	-
7	Side	F	-	Blonde	Straight	Long	-	-
8	Side	M	Child	Blonde	Straight	Short	-	-
9	Side	F	Old	Blonde	Straight	Short	-	Glasses
10	Side	F	Young	Brown	Curls	Long	-	-
11	Side	F	Young	Brown	Straight	Long	-	Glasses
12	Side	F	-	Brown	Curls	Bob	-	
13	Side	M	Old	-	-	Bald	-	-
14	Side	M	Young	Brown	Straight	Quiff	Beard	-
15	Side	M	-	Black	Spikes	Short	-	-

Table 3: Attributes for characters used in Ex2. Main characters appear every round, side characters only once.

Char	Type	Sex	Age	HairColour	HairType	HairStyle	FacialHair	Accessory
1	Main	M	-	Brown	Straight	Long	Beard	-
2	Main	M	-	-	-	Bald	-	Glasses
3	Main	F	-	Black	Curls	Long	-	-
4	Side	M	-	Brown	Curls	Short	Beard	-
5	Side	M	-	Brown	Straight	Short	Moustache	Glasses
6	Side	& M	& -	& Brown	& Straight	& Long	& Beard	& -
7	Side	& M	& -	& -	& -	& Bald	& -	& Glasses
8	Side	& F	& -	& Black	& Curls	& Long	& -	& -
9	Side	& M	& -	& Brown	& Curls	& Short	& Beard	& -
10	Side	& M	& -	& Brown	& Straight	& Short	& Moustache	& Glasses
11	Side	& M	& -	& Red	& -	& Quiff	& -	& Glasses
12	Side	& M	& -	& Brown	& -	& Bun	& Beard	& Apron

Table 4: Attributes for characters used in Ex1. Main characters appear every round, side characters only once or twice.

C. Explanation of Transaction Unit Relations

Relation	Label	Explanation
Answer	answer	General relation for answers to a question
Answer description	answer-description	Providing an answer to a question by the robot with a description of a character
Continue	continue	Continuation of a turn within the same speaker
Continue description	continue-description	Adding information to a description of a character within the same turn and the same speaker
Acknowledge	ack	General acknowledgement of the previous utterance by the other speaker
Acknowledge understand	ack-understand	Acknowledgement of the previous utterance with expression of understanding
Acknowledge doing	ack-doing	Acknowledgement of the description of a character and signal that a response will be provided
Acknowledge done	ack-done	Acknowledgement of the description of a character and conclude the transaction unit for a character
Acknowledge confirm	ack-confirm	Acknowledgement and explicit confirmation of an utterance by the other speaker
Acknowlege unsure	ack-unsure	Acknowledgement of an utterance by the other speaker with expression of uncertainty
Acknowledge can't	ack-cant	Acknowledgement of a request by the other speaker with expression of inability to fulfill the request
Request confirmation	req-confirm	Request for confirmation of a description or position
Request done	req-done	Request for completion of a transaction unit
Request clarifation	req-clar	Request clarification of a description or position
Clarification confirm	clar-confirm	Confirm a description or position after a request for confirmation
Clarification repair	clar-repair	Provide repair after a request for clarification
Negative acknowledgement	nack	Negative response to the utterance by the other speaker
Offer	offer	Offer to perform an action or let the other player perform an action
Offer accept	offer-accept	Accept the offer to perform an action
Correction	correction	Correction of a previous utterance by the same speaker
Other	other	Any other utterance, usually falling outside of the transaction unit goal

Table 5: Names of all the transaction units found in the dataset with the label used in the dataset and an explanation of their use