Using virtual reality to investigate the emergence of gaze conventions in interpersonal coordination

Gregory Mills [0000-0002-1053-8862] and Remko Boschker

Centre for Language and Cognition, University of Groningen,
Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, Netherlands
g.j.mills@rug.nl

Abstract. Gaze plays a central role in regulating turn-taking, but it is currently unclear whether the turn-taking signals of eye gaze are static and fixed, or whether they can be negotiated by participants during interaction. To address this question, participants play a novel collaborative task, in virtual reality. The task is played by 3 participants, and is inspired by games such as Guitar hero, Rock Band, Beat Saber, and Dance-Dance Revolution. Crucially, the participants are not allowed to use natural language – they may only communicate by looking at each other. Solving the task requires that participants bootstrap a communication system, solely through using their gaze patterns. The results show that participants rapidly conventionalise idiosyncratic routines for coordinating the timing and sequencing of their gaze patterns. This suggests that the turn-taking function of eye-gaze can be flexibly negotiated by interlocutors during interaction.

Keywords: Dialogue, Transformed Social Interaction, Eye-gaze, Turn-taking

1 Introduction

When people speak with each other, they dynamically adapt their language to that of their conversational partner (Pickering and Garrod, 2004; Clark, 1996; Gregoromichelaki et al., 2020; Nölle et al., 2018). A central finding in dialogue research is that the meanings of words and phrases used are negotiated ad hoc by participants. Thus, one recurring feature of dialogue is that participants develop novel, idiosyncratic referring expressions. For example, experiments that set participants the task of describing abstract shapes to each other have shown that, when referring repeatedly to a particular novel shape, one pair of participants might conventionalise a referring expression such as "ice-skater", whereas another pair of participants might conventionalise an entirely different referring expression ("the ballerina") to refer to exactly the same shape (Clark and Wilkes-Gibbs, 1986; Clark and Bangerter, 2004).

In addition to natural language expressions, face-to-face conversation is underpinned by myriad non-verbal signals which are used, inter-alia, to regulate procedural coordination in the interaction. For example, speakers tend to look away from their addressee when starting to speak, and then re-establish eye-contact at the end of their turn in order to yield the floor or signal the next speaker (Kendon, 1967; Degutyte and Astell, 2021).

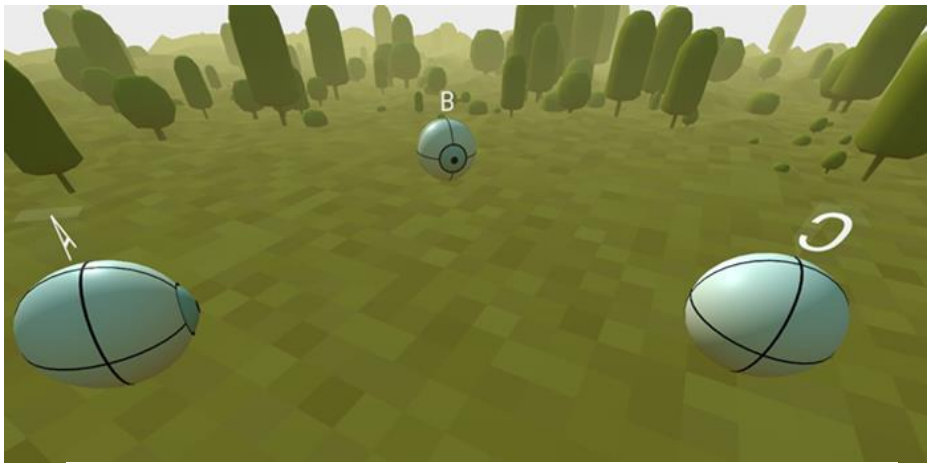

Fig. 1. The virtual environment in which participants play the task

Although research has shown clear cultural differences in such gaze-behaviour (Rossano et al., 2019), it is currently unclear whether the communicative meaning of eye-gaze is static and fixed, or whether, like natural language, it might be dynamically negotiated by participants during interaction.

To address this question, participants play a novel collaborative task within a virtual reality environment which allows for testing whether and how idiosyncratic eye-gaze signals might emerge.

2 Methods

2.1 The task

Groups of 3 participants play a collaborative task[1], in virtual reality, using Oculus Go headsets. Participants, who are rendered as "eye-ball" avatars, are placed equidistantly and facing each other in a virtual environment (see Figure 1, above). The task is inspired by games such as Guitar Hero, Rock Band, and Dance-Dance Revolution. The three key differences are:

1. Instead of performing target sequences of musical notes or dance moves, each triad needs to perform, together, sequences of gaze events. The possible gaze events are (a) looking at a specific participant or (b) looking at oneself in a mirror that is positioned on the right of each participant. For example, a typical target sequence might be: "*Person B must look at Person C. Then Person C must look at Person A. Then, while Person C continues looking at Person A, Person A and Person B must look at each other. Then, Person 3 must look at themselves in their mirror*". Crucially, if any participant makes a mistake, the triad needs to restart the sequence. On each

[1] The source-code is available at https://github.com/gjmills/VRLookingGame

round, the target sequences are generated randomly by the server. The difficulty (i.e., length) of the target sequence is set dynamically by the server: Initially, triads are presented with simple target sequences. On successfully completing a target sequence, participants are presented with more complex (i.e., longer) target sequences. Conversely if a triad fails to solve a sequence within 90 seconds (i.e., a "timeout" occurs), the next sequence is less complex.
2. On each trial, only one participant (the Director) sees the target sequence. This means that in order for the group to complete the target sequence, the Director has to instruct the others, while also themselves participating in completing the target sequence (see Figure 2, below).
3. Crucially, the participants are not allowed to use natural language to communicate – they may only communicate by looking at each other.

This task presents triads with the recurrent procedural coordination problem of communicating and then performing sequences of actions (i.e. "look events") in the correct order and with the correct timing. Solving the task, therefore, requires that triads bootstrap an ad hoc communication system (see, e.g., Scott-Phillips, 2009; Nölle and Galantucci, 2022; Stevens and Roberts, 2019) for instructing and taking turns, solely using their gaze patterns (See https://youtu.be/ctXXtFBr6Cc for a video of participants playing the game).

2.2 Manipulation

In order to test whether participants develop idiosyncratic signals for coordinating procedurally, the experiment used a technique similar to that used by Healey (2008) and Mills (2011), namely, using *transformed social interaction* (Bailenson et al., 2004; McVeigh Schultz and Isbister, 2021; Cheng et al., 2017) to artificially manipulate the participants' communicative behaviour.

The experiment was divided into a 25 minute "training phase" followed by a 5 minute "test phase". During the training phase, triads complete the task as described above. At the start of the test-phase, the identities of the participants were swapped in the following manner: Each participant continues to see the other two avatars in the same locations. However, the participants controlling those avatars are swapped: In Participant A's headset, Participant B's physical head movements are mapped onto Participant C's avatar, while Participant C's physical head movements are mapped onto B's avatar. Similarly, in B's headset, B now sees A's head movements animating C's avatar and sees C's head movements animating A's avatar. Also in C's headset, C sees A's physical head movements animating B's avatar, and vice versa.

While the training phase tests whether participants are able to bootstrap a communication system, this later manipulation[2] in the test-phase investigates whether

[2] We originally intended to use 3 groups of triads in order to create triads in the test-phase that comprise participants who were members of different triads in the training phase, similarly to the setup in Healey (2008). However, due to technical difficulties with networking 9 headsets we used the approach of 3 triads.

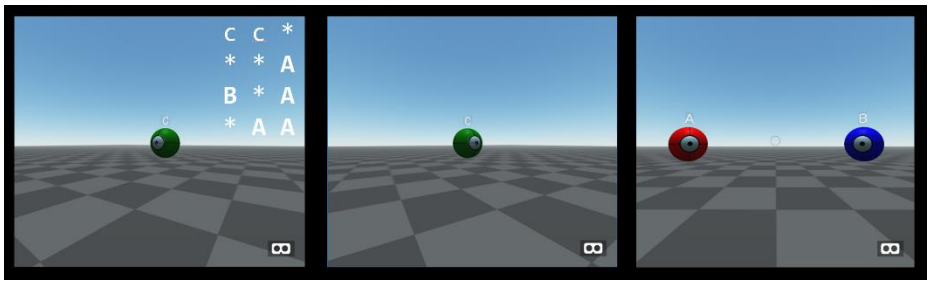

Fig.2. The view from each of the three participants' headsets
(From left to right: Participants A, B, C).

Participants are rendered as virtual eye-balls that are anchored at a particular location in the sky. Each participant's ID (e.g., A, B, C) is displayed above their heads. In this example, Participant A is assigned the role of Director, this why the target sequence is displayed in the first screen, i.e., A's view. The target sequence of "look events" is displayed as a three-column table in the top-right hand corner of A's display. The table is read from top to bottom. The left-most column cells represent the actions to be performed by Participant A. The middle column represents the actions to be performed by Participant B, and the right-most column represents the actions to be performed by Participant C. Each row describes a gaze configuration that must be achieved simultaneously by the triad. An asterisk means that the participant corresponding to that cell does not need to perform any action. For example, the target sequence displayed in Participant A's window represents the following sequence of actions: "*First A and B both need to look at C.* (row 1) *Then C needs to look at A.* (row 2) *Then while C looks at A, B needs to look at A.* (row 3) *Then B needs to look at A* (row 4)". The task of the Director is to get the triad to perform this sequence of look events, which requires that the Director communicates this sequence to the other participants. On successful completion of a look event, the corresponding letter in the Director's window changes to lower-case. Crucially, if any participant produces the wrong look event, the triad needs to restart the sequence, i.e., all letters return to upper-case. Fig 2 then shows the configuration of A,B,C after successfully completing the first row: The left-most screen shows, from A's perspective, A looking at C; the middle-screen shows, from B's perspective, that B is looking at C; the right-most screen shows, from C's perspective, both A and B looking at C.

participants within the triads develop a different communication system with each partner: participants are unaware that the identities of their partners are swapped, so if they have indeed established different systems, then, on entering the test phase, they will attempt to reuse a convention with the same partner (who is actually the other partner). Under the effect of the manipulation of identity swapping, this should lead to more errors and less efficient communication.

2.3 Hypotheses

The experiment tested two hypotheses:

1. During the training phase, participants will establish a communication system with each other that will allow them to collaboratively solve the target sequences.

2. In the test phase, the manipulation will cause participants to inadvertently use the wrong signals with each other, causing disruption to task performance.

3 Results

69 triads took part in the experiment.

3.1 Training phase

During the 25-minute training phase, triads completed a mean of 20.5 sets (S.D. = 3.45). The most successful triad completed 27 sets. By the end of the training phase, triads were solving sets with a mean of 5.5 target items (S.D.=1.2). The most successful triad completed sets containing 8 targets (see, e.g., Figure 2 which shows a target set containing 7 "look events").

3.2 Test phase

To test the effect of the intervention, we compared participants' performance in the 5 minutes preceding the swap with their performance during the 5-minute test phase . We used two measures of disruption to task performance.

The first measure, task success, was modelled with a mixed binary logistic regression, using the lme4 package (Bates et al., 2014), which showed that triads solved significantly fewer games in the test phase ($b = -0.49$, S.E. = 0.193, $z = -2.54$), $p = 0.0111$). The model predicts that triads successfully solve 66% [95%CI: 0.60, 0.72] of target sets in the training phase and 54% [95%CI: 0.48, 0.61] of target sets in the test phase.

The second measure recorded the number of "look events" per game, i.e., the number of times a participant selected a target. All things being equal, if participants are encountering more difficulties coordinating with each other, this will lead to them having to make more selections, i.e., expend more effort, to solve a set. A linear mixed model using the lme4 package showed that triads produced significantly more look events in the 5-minute test phase than in the last 5 minutes of the training phase ($b = 10.4$, S.E. = 2.98, $t = 3.5$, $p < 0.001$). The model predicts 40 [95%CI: 36.2, 43.8] look events per game in the training phase, and 50.4 [95% CI: 45.5, 55.4] look events in the test phase.

4 Discussion

The results provide support for both hypotheses. The average sequence length at the end of the training phase suggests that the participants were solving the sets by communicating with each other, as opposed to solving via individual trial and error. During piloting, we observed participants attempting to solve the sequences without attempting to establish a communication system with each other – these triads almost never managed to solve sequences longer than length 2.

Moreover, the increased number of timeouts and look events in the test phase suggest that the manipulation disrupted participants' coordination. A plausible explanation for this pattern is that many participants communicated differently with each partner. This was confirmed by the participants themselves. On debriefing, we asked participants about the communication system they had developed. Some participants explicitly stated that they noticed that their partners communicated differently (e.g., using different signals for the same actions, or communicated faster/slower), which they had attempted to accommodate.

Given that participants develop idiosyncratic signalling systems with each of their co-players simultaneously, it is clear that they demonstrate ability to discriminate and adapt dynamically to different participants at the same time during a single task. It is an open question how this form of audience design compares with how participants take each other's perspective into account when they adapt their language to the interlocutor, e.g., when producing referring expressions (Fischer, 2016; Yoon and Brown-Schmidt, 2019; Healey and Mills, 2006) or when associating expressions' meanings with particular sequential positions in the unfolding interaction (Mills and Gregoromichelaki, 2010; Gregoromichelaki et al., 2011; Mills, 2014).

These findings are subject to a couple of important caveats concerning the ecological validity of the experimental setup: First, the participants' movements are severely constrained. The Oculus Go headsets only capture rotations around the x,y,z axes, but do not capture any change in location: throughout the experiment, the avatars are anchored at a fixed location. Second, the setup conflates "head gaze" and "eye gaze", as participants' head-movements are mapped onto their virtual eye-ball (see, e.g., Špakov et al., 2019).

Nonetheless, these findings suggest that the interactive signals that participants use to attract and direct another's visual attention can be flexibly negotiated during an interaction. In addition, the restriction of movement to rotations around the x, y, z axes makes the findings all the more surprising, as they show that participants are still able to bootstrap a communication system within these quite severe constraints.

To conclude, these findings are of central importance for theories of Human-Computer Interaction. Research on dialogue has shown that in order for systems to converse naturalistically with humans, they must be able to dynamically adapt their vocabularies, ontologies, and emotional signals to their conversational partner during the interaction (Healey, 2021; Mills et al., 2021; Larsson, 2007; Cooper, forthcoming). The findings from the current experiment suggest that, in addition, technologies such as avatars, dialogue systems, as well as self-driving cars when communicating with pedestrians (Habibovic et al., 2018), need to be able to flexibly adapt their non-verbal and turn-taking signals to those of the user.

References

Argyle, M (1988). Bodily Communication. London, England, Methuen, 2nd edition.

Bailenson, J. N., Beall, A. C., Loomis, J., Blascovich, J., & Turk, M. (2004). Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments. *Presence: Teleoperators & Virtual Environments*, 13(4), 428-441.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Cheng, L. P., Marwecki, S., & Baudisch, P. (2017). Mutual human actuation. *In Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (pp. 797-805)*

Clark, H. (1996). Using language. Cambridge University press.

Clark, H. & Bangerter, A. (2004). Changing ideas about reference. In *Experimental pragmatics* (pp. 25-49). Palgrave Macmillan, London.

Clark, H., & Wilkes-Gibbs, D. (1986).Referring as a collaborative process.*Cognition*, 22(1),1-39

Cooper, R., (forthcoming). From perception to communication: An analysis of meaning and action using a theory of types with records (TTR). CUP

Degutyte, Z., & Astell, A. (2021). The role of eye gaze in regulating turn taking in conversations: a systematized review of methods and findings. Frontiers in Psychology, 12.

Fischer, K. (2016). Designing speech for a recipient. *Designing Speech for a Recipient*, 1-337.

Gregoromichelaki, E., Kempson, R., Purver, M., Mills, G. J., Cann, R., Meyer-Viol, W., & Healey, P. G. (2011). Incrementality and intention-recognition in utterance processing. *Dialogue & Discourse*, 2(1), 199-233.

Gregoromichelaki, E., Mills, G. J., Howes, C., Eshghi, A., Chatzikyriakidis, S., Purver, M., ... & Healey, P. G. (2020). Completability vs (In) completeness. *Acta Linguistica Hafniensia*, 52(2), 260-284.

Habibovic, A., Lundgren, V. M., Andersson, J., Klingegård, M., Lagström, T., Sirkka, A., ... & Larsson, P. (2018). Communicating intent of automated vehicles to pedestrians. *Frontiers in psychology*, 1336.

Healey, P. , & Mills, G. (2006). Participation, precedence and co-ordination in dialogue. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (Vol. 320). Vancouver: Cognitive Science Society.

Healey, P. (2008). Interactive misalignment: The role of repair in the development of group sublanguages. *Language in Flux*. College Publications, 212.

Healey, P. (2021). Human-Like Communication. *Oxford University Press, Oxford, England*.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, 26, 22-63.

Larsson, S. (2007). A general framework for semantic plasticity and negotiation. In *Proceedings of the Seventh International Workshop on Computational Semantics* (IWCS-7).

McVeigh-Schultz, J., & Isbister, K. (2021). The case for "weird social" in VR/XR: a vision of social superpowers beyond meatspace. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-10).

Mills, G. (2011). The emergence of procedural conventions in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Mills, G. J. (2014). Dialogue in joint activity: Complementarity, convergence and conventionalization. *New Ideas in Psychology*, 32, 158-173.

Mills, G., & Gregoromichelaki, E. (2010). Establishing coherence in dialogue: sequentiality, intentions and negotiation. *Proceedings of SemDial* (*PozDial*).

Mills, G., Gregoromichelaki, E., Howes, C., & Maraev, V. (2021). Influencing laughter with AI-mediated communication. *Interaction Studies*, 22(3), 416-463.

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, 181, 93-104.

Nölle, J. & Galantucci, B. (to appear). Experimental Semiotics: past, present and future. In Garcia & Ibanez (to appear) *Routledge Handbook of Neurosemiotics*

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169-190.

Rossano F., Brown P., Levinson S. C. (2009). Gaze, questioning and culture. *Convers. Anal*. 27, 187–249. 10.1017/CBO9780511635670.008

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113(2), 226-233.

Špakov, O., Istance, H., Räihä, K. J., Viitanen, T., & Siirtola, H. (2019, June). Eye gaze and head gaze in collaborative games. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (pp. 1-9).

Stevens, J. S., & Roberts, G. (2019). Noise, economy, and the emergence of information structure in a laboratory language. *Cognitive Science*, 43(2), e12717.

Yoon, S. O., & Brown-Schmidt, S. (2019). Audience design in multiparty conversation. *Cognitive science*, *43*(8), e12774.