Performance in a Collaborative Search Task: The Role of Feedback and Alignment

Moreno I. Coco
School of Philosophy, Psychology and Language Sciences, University of Edinburgh
3 Charles Street, Edinburgh EH8 9AD, UK
moreno.cocoi@gmail.com

Rick Dale
Cognitive and Information Sciences, University of California, Merced
5200 North Lake Rd., Merced CA 95343, USA
rdale@ucmerced.edu

Frank Keller
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
keller@inf.ed.ac.uk

Abstract

When people communicate, they coordinate a wide range of linguistic and non-linguistic behaviors. This process of coordination is called alignment, and is assumed to be fundamental to successful communication. In this paper, we question this assumption and investigate whether disalignment is a more successful strategy in some cases. More specifically, we hypothesize that alignment correlates with task success only when communication is interactive. We present results from a spot-the-difference task in which dyads of interlocutors have to decide whether they are viewing the same scene or not. Interactivity was manipulated in three conditions by increasing the amount of information shared between interlocutors (no exchange of feedback, minimal feedback, full dialogue). We use recurrence quantification analysis to measure the alignment between the scan-patterns of the interlocutors. We found that interlocutors who could not exchange feedback aligned their gaze more, and that increased gaze alignment correlated with decreased task success in this case. When feedback was possible, in contrast, interlocutors utilized it to better organize their joint search strategy by diversifying visual attention. This is evidenced by reduced overall alignment in the minimal feedback and full dialogue conditions. However, only the dyads engaged in a full dialogue increased their gaze alignment over time to achieve successful performances. These results suggest that alignment per se does not imply communicative success, as most models of dialogue assume. Rather, the effect of alignment depends on the type of alignment, on the goals of the task, and on the presence of feedback.

Keywords: Interactivity; alignment; task success; dialogue task; eye-tracking.

Introduction

Research in dialogue has shown that effective communication often occurs when the cognitive processes of speakers and listeners align, i.e., become similar. Alignment in linguistic responses manifests itself in a number of ways. Interlocutors use the same syntactic structures (Branigan, Pickering, & Cleland, 2000), the same ways of describing objects or locations (S. E. Brennan & Clark, 1996; Garrod & Anderson, 1987), and converge on the same topic of conversation (Sacks, Schegloff, & Jefferson, 1974). Alignment can also occur in non-linguistic behaviors. For example, two people working together move their bodies in similar ways (Shockley, Santana, & Fowler, 2003), distribute their visual attention similarly (Richardson & Dale, 2005), and exhibit alignment across a wide range of non-verbal responses, such as nodding and smiling (Louwerse, Dale, Bard, & Jeuniaux, 2012). In the present study we will focus on gaze alignment as a measure of shared visual attention.

Even though it seems natural to assume that alignment might underpin successful communication, the literature on the topic shows mixed evidence. When interlocutors use language to help each other identify unfamiliar shapes (in tangram-matching tasks), or follow directions on a path (in maze or map tasks), they are successful if they converge on a common set of referring expressions or re-use similar syntactic structures (Krauss & Glucksberg, 1969; Glucksberg, Krauss, & Higgins, 1975; Garrod & Anderson, 1987; Schober & Clark, 1989; A. H. Anderson et al., 1991; S. E. Brennan & Clark, 1996; Reitter & Moore, 2014). However, not all types of linguistic alignments are predictive of task success. Fusaroli et al. (2012), for example, show that indiscriminate and widespread alignment[1] leads to a *lower* performance than a more moderate level of specific alignment in a joint detection task (see also Wu & Keysar, 2007, who show that excessively entrained dyads are more likely to commit errors in a tangram task). A related result is presented by Ireland and Henderson (2014), who found that dialog partners with higher levels of language style matching were more engaged, but were also less likely to negotiate successfully. Furthermore, interactivity, which can be defined as the possibility of interlocutors to provide feedback, seem to play a key role in alignment. When participants do the tangram task alone, for example, they fail to come up with efficient referring expressions (Hupet & Chantraine, 1992).

Interactivity is certainly crucial for communication, but it is currently not clear how it influences communication outcomes, or what the benefits of coordinating behavior are when dyads interact to solve a task. In tangram tasks, for example, dyads reduce their speech more when they can exchange feedback (Krauss & Weinheimer, 1966), and benefit from physical co-presence, which aids the grounding of shared knowledge, and consequently increases task success (Schober & Clark, 1989; Clark & Krych, 2004). The ability to interact, of which the exchange of feedback is an instance, is crucial for task performance. However, more interaction does not automatically imply stronger alignment of responses. It is conceivable that dyads utilize feedback to *disalign* rather than align responses if the task requires it. In a study similar to the one we present here, S. Brennan, Chen, Dickinson, Neider, and Zelinsky (2008) (and follow-up work by Neider, Chen, Dickinson, Brennan, & Zelinsky, 2010) had dyads work remotely to identify a "sniper target" (a small red circular shape) in a complex scene. The authors manipulated the amount and type of feedback the dyad was allowed to share (no communication, voice, gaze, or both voice and gaze). When the dyads could interact, they had reduced error rates compared to when they could not communicate. Crucially, the presence

[1] In this study, indiscriminate alignment refers to the repetition of arbitrary lexical items, rather than just the repetition of task-relevant lexical items.

of feedback led to disalignment, rather than alignment, of the attentional responses of the dyad (refer to Figure 5 of S. Brennan et al., 2008 for an example). The exchange of feedback helped the dyads improve their performance by diversifying, rather than homogenizing, their joint search space.

This is an intriguing result: most models of dialogue assume that interactivity fosters alignment, rather than disalignment, and that alignment will boost task success. Several interactive models of alignment have been proposed in the literature, explaining alignment using a range of different cognitive mechanisms such as priming (e.g., Pickering & Garrod, 2004), partner-directed adaptation (e.g., Keysar, Barr, Balin, & Brauner, 2000), mutual adaptation (e.g., S. E. Brennan, 2004), or lower-level mechanisms of perceptuomotor coupling (e.g., Shockley, Richardson, & Dale, 2009). The debate is ongoing, and the various mechanisms underlying successful dialogue and joint tasks will likely involve aspects of several accounts, rather than just one of the prevailing theories (Dale, Fusaroli, Duran, & Richardson, 2013; Fusaroli, Raczaszek-Leonardi, & Tylén, 2014). Despite their differences, however, all these theories assume that interactivity plays a fundamental part in the dynamics of dialogue. They share the assumption that interactivity mediates alignment and supports performance in communicative tasks.

In the present study, we investigate gaze alignment. Our aim is to work out the relative contributions of alignment and interactivity to successful task performance, and in particular to elucidate their interaction. This will allow us to distinguish between purely alignment-based theories (such as the Interactive Alignment Model of Pickering & Garrod, 2004) and theories based on low-level coupling mechanisms (such as the coordination model of Shockley et al. (2009)), which do not tie task success directly to alignment. We hypothesize that interactivity plays a crucial role in determining whether alignment is correlated with task success, which leads us to ask under which conditions feedback fosters disalignment rather than alignment.

In previous work, reviewed above, participants were either introduced as new listeners or were mere overhearers, and mostly took part in interactive tasks involving full two-way dialogue. Moreover, alignment and interactivity were typically not experimentally distinguished, and often studied separately (as in the tangram task). The work reported here instead uses a dyadic task, and we experimentally manipulate the amount of information that the interlocutors are allowed to exchange; from no interaction (a listener follows the instructions of a speaker in real time) to full dialogue (both interlocutors communicate to achieve the task), while also including an intermediate scenario, in which communication feedback is limited to backchannels. The comparison of these three setups provides a direct test of the nature of interactive information exchange. Crucially, the non-interactive version of our task is also dyadic – the listeners collaborate with the speakers, they are not mere overhearers (as in previous work). If it is merely this shared experience that is required for task success, then performance should not change as a function of feedback. The alternative hypothesis is that the alignment of a dyad, and possibly also their task performance, is altered by feedback in systematic ways.

Therefore, our experimental setup is designed to answer basic questions, yet unsolved, that are of interest to all theories of interaction: Does interactivity lead to increased alignment? Does more alignment correlate with improved task performance, and is interactivity a mediating factor? Can we observe other successful strategies, such as disalignment?

The Present Study

In this experiment, we measure gaze alignment and task performance in a spot-the-difference task, in which interlocutors have to decide whether they are viewing the same visual scene or not.

We manipulate the amount of feedback that the interlocutors can exchange. This manipulation is implemented using a between-participant design involving three different groups of dyads. The design compares the following conditions: (1) *no-feedback*, where one participant (i.e., the speaker) describes the scene, while the other one (i.e., the listener) is not allowed to communicate and has to decide whether the scene is the same or not; (2) *minimal-feedback*, where the listener is allowed to provide backchannel responses to the speaker to signal understanding (e.g., "uh huh", "mhm", or "yeah", as well as "yes" and "no"); (3) *full-dialogue*, where the interlocutors can discuss freely to reach a joint consensus before taking a decision.

The central prediction of alignment-based models of dialogue (such as Pickering and Garrod's (2004) Interactive Alignment Model) is that more alignment should lead to more task success. However, in the literature, this prediction is mostly, if not uniquely, based on linguistic responses. Alignment on visual responses can in fact be detrimental in a search task, whereas diversifying visual attention can increase the likelihood of the dyad spotting a difference S. Brennan et al. (2008).[2] Moreover, in such a task, alignment can only be predictive of task success if the dyad can build a common ground by fully interacting with each other. Common ground refers to what the interlocutors know about each other's knowledge (e.g., which objects they have detected in a visual scene). Such common ground is necessary if interlocutors are to develop effective visual search strategies through dialogue.

We expect dyads to display stronger gaze alignment in the *no-feedback* condition. The inability of the listener to signal his/her understanding to the speaker presumably forces them to follow the speaker's instructions more closely, leading to more gaze alignment. If there is too much alignment, however, then we expect this to be detrimental to task performance: if the listener merely follows the speaker's gaze, rather than utilizing the information provided, then they are more likely to miss key differences in the scene. In contrast, the more feedback is possible, the better the dyads can diversify their search strategy. This should result in decreased alignment (as the attentional responses of the interlocutors diverge), but increased task performance, especially when the dyad can fully interact, i.e., in the *full-dialogue* condition. Finally, we predict that only when interlocutors can fully interact, they can incrementally construct and maintain gaze alignment over the course of a trial. We expect this to be a critical signature of task success. However, when interlocutors are not free to exchange feedback, or when feedback is only minimal, they cannot construct a common ground. As a consequence, they are unable to coordinate gaze over time and to use this to successfully accomplish on the task.

Method

Participants. Forty-eight dyads (16 per sub-experiment) were recruited through the Student Careers Service of the University of Edinburgh. Each participant gave informed consent and was paid £7 for participating. Only two dyads knew each other before participating in the study.

The sample size of 16 participants per sub-experiment was determined before running the experiment, based on the prior literature on eye-tracking studies of dialogue behavior (e.g., Dale, Kirkham, & Richardson, 2011). No stopping criterion was used for the data collection, i.e., all participants of a given sub-experiment were run before the data was analyzed.

Ethical approval for this study was granted by the Ethics Committee of the School of Philosophy, Psychology and Language Sciences of the University of Edinburgh, in accordance with the

[2] Note that this study did not explicitly examine alignment.

Figure 1. An example photo-realistic scene used in our experiment. The red box indicates the target object (flag), which was placed differently for the speaker (bottom left) and the listener (top right). The red box is used for illustration only and was not shown to participants.

University's Ethics Code of Practice and the British Psychological Society guidelines on ethics.

Materials. One hundred photo-realistic scenes were used, representing a mix of indoor and outdoor settings taken from the internet, as well as from existing image databases (e.g., LabelMe; Russell, Torralba, Murphy, & Freeman, 2008). A target object, and the distractors, were inserted in the scene using Photoshop. Distractors were used to avoid the development of a scanning strategy, and to make the identification of the targets more challenging (refer to Figure 1 for an example image). Each scene was fully annotated with polygons using the LabelMe toolbox (Russell et al., 2008). The polygons were then used to map eye-movement fixations onto objects using their screen coordinates. On average there were 21.48 ± 10.74 annotated objects per scene.

Procedure. Participants performed a spot-the-difference task on 100 visual scenes (four practice trials and 96 experimental trials) while their eye-movements and speech was recorded. Each participant used a separate screen; participants were not able to see each other or each other's screens while performing the task. For the *no-feedback* and *minimal-feedback* conditions, one of the participants was the speaker, who had to describe the visual scenes to the other participant, the listener, who had to decide whether they were viewing the same scene or not. For the *full-dialogue* condition, the dyads could fully interact, so there were no such roles. However, also in this scenario, one member of the dyad, chosen at random at the beginning of the experiment, provided the task response. However, we highlighted in the instructions that the decision should have been reached jointly.

Half of the scenes were identical, while a difference was present in the other half of the scenes. To make the task more challenging, in the Different trials, we changed either the position of the target object in the scene (i.e., both scenes contained exactly the same objects, but one was displaced to the left or right), or its visual presence (i.e., in one of the two scenes, the target object was missing). For the Same trials, both scenes were identical.

In a subset of the scenes (38), we manipulated the visual saliency of the target object, as well as its contextual congruency within the scene (similar to M. Coco, Malcolm, & Keller, 2014). This manipulation was introduced following Underwood, Templeman, Lamming, and Foulsham (2008) to examine whether low- and high-level informational properties of the target could mediate

performance on the spot-the-difference task in a dialogue setting. However, this manipulation is not the focus of the current study, which examines a broader experimental hypothesis about gaze alignment, task success, and interactivity. Therefore, all experimental trials were analyzed together.

As noted above, interactivity was manipulated by running the three experimental conditions (No-Feedback, Minimal-Feedback, and Full-Dialogue) as three sub-experiments with different sets of participants. The no-feedback situation was created by not allowing the listener to give any information to the speaker. The only information about the listener that the speaker receives is that they know when a decision has been made by the listener, as this is when the trial ends. The speaker is not given any information on which decision was made by the listener, or whether it was correct or not. The minimal-feedback condition was created by allowing the listener to provide the speaker with yes/no responses and backchannel utterances. In order to ensure that feedback is constrained in this way, the experimenter monitored the speech of the listeners remotely through a Motorola baby monitor, and checked a sample of the recorded speech of the listeners to ascertain that the instructions were correctly followed. The listeners were made aware of this in the written instructions. Finally, the full-dialogue condition was created by allowing the dyads to communicate as they wished. The constraint was that only one of the interlocutors could provide the response, after a joint decision was taken. We investigate this manipulation as a Feedback variable with three levels.

Two SR EyeLink II head-mounted eye-trackers were used to monitor participants' eye-movements with a sampling rate of 500 Hz. Images were presented on a 21" Multiscan monitor at a resolution of 1024×768 pixels. Participants sat 60–70 cm from the computer screen, which subtended a region of approximately 20 degrees of visual angle. Eye-movements of participants were co-registered, i.e., the onset of the scene and the timestamps of trackers were synchronized. A test of eye-dominance was performed at the beginning of each session for both participants, and only the dominant eye was tracked. For the No-Feedback and Minimal-Feedback conditions, participants in the dyad were invited to decide themselves whether they wanted to play the role of the speaker or the listener, after reading written instructions which explained both roles. In the instructions, the speaker was asked to describe the scene to the listener, such that they would be able to decide whether the scene was the same or not. For the Full-Dialogue condition, the member of the dyad providing the spot-the-difference response was randomly chosen by the experimenter. In the instructions of this sub-experiment, participants were told to have a discussion in order to reach a joint decision, and to reach agreement before providing the response.

The participants were not informed of the types of differences that could be present in the scenes. They were just told that there was either one difference or no difference. Both participants were recorded using lapel microphones. The trial ended when the decision ("different" or "not different") about the scene was made by pressing the "l" or "s" key on the keyboard. No time limit was set to take a decision. At the end of every trial, drift correction was performed on both participants, after which the next trial started. A nine-point calibration was performed at the beginning of the session and repeated approximately halfway through the session. Some participants required more than two calibrations. At the beginning of every session, participants were given four practice trials to familiarize them with the experiment. The duration of the experiment was between 45 and 60 minutes.

Analysis. We examine gaze alignment in the dyad in order to determine whether it is a necessary precondition for task success, and whether feedback enhances or reduces alignment, especially for correct responses.

Cross-Recurrence Quantification Analysis. In order to obtain empirical measures of gaze alignment, we utilize Recurrence Quantification Analysis (RQA, Zbilut, Giuliani, & Webber, 1998; Marwan & Kurths, 2002; Marwan, Carmen Romano, Thiel, & Kurths, 2007). This technique makes it possible to quantify how, and to what extent, a signal is revisiting a similar state over time. When RQA is applied on two different streams of the same type of information, such as the eye-movement trajectories of two interlocutors, it is called Cross-Recurrence Quantification Analysis (CRQA).

In this study, we focus on the following CRQ measures: (a) the recurrence rate (RR), which measures the density of recurrence points in the whole Cross-Recurrence Plot (CRP). This measure summarizes the amount of recurrence occurring overall. A high gaze RR indicates that the interlocutors look at the same objects, including recurrence of the interlocutors with themselves and regardless of directionality. However, we are not just interested at such "indiscriminate" recurrence. Rather, we want to focus on the recurrence properties observed when the two time series align. This can be done by looking at the properties of the diagonal lines of a CRP. In particular, we consider: (b) the average length of the diagonal (L), which reflects the regularity of the system, whereby high values of L indicate that the dyad consistently align on the same set of objects; (c) the percentage of recurrence points forming diagonal lines (DET), which reflects the predictability of the system: high DET values indicate that when the alignment of the gaze of the dyad on the same objects, it does so for a long period of time; and finally (d) the entropy of the line distribution (ENTR), whereby high ENTR values indicate that the time segments in which the dyad gaze alignment varies widely.[3]. As scan-patterns are categorical sequences, we have used a delay of 1, an embedding of 1, and a radius of 0.0001 to run the CRQA analysis (see Dale, Warlaumont, & Richardson, 2011 for more details).

Moreover, in order to track how gaze alignment develops as the interaction progresses, we compute window cross-recurrence (refer to Boker, Xu, Rotondo, and King (2002), for a similar approach based on correlation). For this, a cross-recurrent plot of the two series is computed in overlapping windows of a specified size for a number of delays smaller than the size of the window over the two series. The window is moved with a fixed step. As our series are normalized to be 101 bins, we have chosen a window of size 10, we use a delay of 5, and move the window by a step of 2. On each CRP, the same measures described above (e.g., RR) can be extracted. In the main paper, we report RR for correct trials only, as we are interested in examining how overall gaze alignment is established over time under the different feedback conditions during successful interactions. In the Supplementary Material, however, we provide the reader with the time-course results for the other measures of L, DET and ENT.

Please refer to Marwan et al. (2007) for a more detailed description of these measures, to M. Coco and Dale (2014) for an explanation of the CRQA method in the context of behavioral data, to N. C. Anderson, Bischof, Laidlaw, Risko, and Kingstone (2013) for an explanation of the method in the context of eye-movement, and to M. I. Coco and Dale (2014) for the R `crqa` package,[4] which was used to compute the recurrence measures reported in this study.

(In of the Appendix, we also report results corroborating those ones presented in the main text, using diagonal-wise cross-recurrence, which is an approach previously applied to measures gaze alignment during dialogue, e.g., Richardson & Dale, 2005; Richardson, Dale, & Kirkham, 2007.)

[3]Note that we could observe a high DET or high L (the dyad aligns gaze for long period of time), while at the same time having high ENTR (the duration of such alignment varies substantially).

[4]The `crqa` package has been shown to yield exactly the same results as the widely used MATLAB package `crptoolbox` by Norbert Marwan.

Gaze Alignment and Co-variates. CRQA is computed on eye-movement responses of the interlocutors in each dyad, represented in the form of scan patterns (SPs), i.e., temporal sequences of fixated objects (e.g., Noton & Stark, 1971; M. Coco & Keller, 2012), for windows of 25 ms each [5]. Each trial is self-terminated by the listener, thus SPs differ in length, especially across the three Feedback conditions. (The mean durations are: No-Feedback: 14.22 ± 9.41 seconds, Minimal-Feedback: 17.31 ± 11.18 seconds; Full-Dialogue: 22.97 ± 13.72 seconds.) We therefore normalize each scan pattern (SP_{old}) by mapping it onto a normalized time-course of fixed length SP_{new} (101 bins). For each SP_{old}, we slide a time-window w with the number of old time-points k^i corresponding to $k^i = \text{length}(SP_{old})/\text{length}(SP_{new})$.

In each w, we calculate the proportion of fixations for each unique object looked at, and subsequently select the object with the highest proportion of fixation to be mapped into the corresponding unit of the normalized time-course.[6] In practice, for each SP we select the sequence of objects attended to most of the time. The technical advantage of normalizing the SPs is that we can construct summary heat-maps of the CRP for the experimental factors of interest (Feedback, Accuracy), as all CRPs have the same 101×101 dimension, rather than having to pick just a couple of illustrative examples, as it is done by most of the literature using this method. Moreover, the theoretical advantage is that the measures of gaze alignment are now comparable between the three Feedback conditions, and any difference observed can be genuinely attributed to the presence of Feedback, rather than to incidental differences in trial duration.

We fit linear-mixed effect models with measures of gaze recurrence as our dependent variables (DVs) and two independent variables (IVs): Feedback, a between-participants variable (No-Feedback, Minimal-Feedback, or Full-Dialogue), and the Accuracy of the listener in detecting whether they were viewing the same scenes as the speaker (i.e., a binomial variable with 1 corresponding to correct and 0 to incorrect responses).

Moreover, we consider the Response Time (accounting for the duration of the trial) and the Order of trials (accounting for learning strategies) as co-variates, and control for their effects on all DVs reported in this study. In particular, we residualize them against the DV under analysis in a simple linear regression model (`depM ~ RT + Order`, using R syntax), and we take the residuals obtained as the DV for further inferential analysis. This ensures that the effects of the IVs (Feedback and Accuracy) on each DV analyzed are not influenced by these incidental co-variates. We report and visualize the cross-recurrence measures RR, L, DET, and ENTR.

Statistical Analysis. All statistical inferences were drawn using the framework of linear mixed effects (LME) models as implemented by the `lme4` (Bates, Martin, Bolker, & Walker, 2014) package in the R programming language. Simply put, LME is a form of hierarchical regression that can account for the variability of random variables, which usually relate to sampling, e.g., Participant and Item (Baayen, Davidson, & Bates, 2008).

Specifically, in LME models, the dependent variable is a linear function of different predictors (fixed effects), and the variance implicit in the multilevel structure of the data is accounted for by grouping based on the random variables of the design. Our fixed effects are Feedback (No-Feedback, Minimal-Feedback, Full-Dialogue) and Accuracy (Correct, Incorrect). Our random effects are Dyad (48 levels), entered as a between-participant variable and Scene (384 levels, i.e., the overall number

[5] We have extracted the fixation events using the Data Viewer parsing algorithm developed by SR Research, and kept its default parameter settings. For each data sample, the SR parser computes velocity and acceleration in degrees of visual angle. If a sample is faster than 30 degree per second, it assumes that a saccade is taking place

[6] Note that 101 bins is smaller than the minimum length of 133 observed in the dataset.

of individual scenes).[7] We attempted to fit mixed-effects models with full fixed effects structure (i.e., all main effects and their interaction, `depM ~ Feedback * Accuracy`, using `lme4` syntax), while also including a maximum random effects structure, in which random variables are included both as random intercepts and as uncorrelated random slopes (e.g., `(0 + Feedback | Dyad)`).[8] This approach is known to result in the lowest rate of Type 1 error (Barr, Levy, Scheepers, & Tily, 2013). However, none of such maximal LMEs converged on any of the DVs extracted from our data. Thus, in order to have a principled way of selecting the final model, which is also justified by the data, we utilized the R package `lmerTest` (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2014), and performed a backward selection only on the random structure of the model removing those terms (evaluated one at time, and starting from the largest model including all random effects) which, when included, did not improve the model fit at $p < 0.1$ (see Kuznetsova, Christensen, Bavay, & Brockhoff, 2015 for greater details on the selection procedure).

Finally,in order to analyse windowed cross-gaze recurrence, beside the predictors of Accuracy and Feedback, we include a Time predictor in the LME model, represented as an orthogonal polynomial of order two (Time^1 and Time^2), to capture how gaze recurrence evolves during the course of a trial.

In the results tables, we report the coefficients, standard errors, and t-values. We derive p-values for the fixed effects in the LME models from F-tests based on the Satterthwaite approximation of the effective degrees of freedom (Satterthwaite, 1946).[9]

Results and Discussion

From a total of 4,608 trials (16 dyads per 3 feedback condition over 96 experimental items), we had to remove 686 trials (i.e., 15% of the data; 187 in the No-Feedback condition, 311 in the Minimal-Feedback condition) and 188 in the Full-Dialogue condition, due to poor calibration (the threshold for excluding trials was set at $> 10\%$ of out-of-range fixation for either partner in the dyad), reaction-times smaller than 250 ms (responses taken involuntarily), failed synchronization between the eye-trackers, or machine error. Therefore, the results reported will be based on the analysis of 3,922 unique trials.

Gaze Alignment. In Figure 2, we show heatmaps that visualize the recurrence rate for the alignment of gaze across the conditions of Feedback (No-Feedback, Minimal-Feedback and Full-Dialogue) and Accuracy (Correct, Incorrect). We observe that the amount of alignment decreases with increased levels of feedback. Crucially, the overall amount of RR in the heatmaps changes as a function of both Feedback and Accuracy. In particular, Incorrect responses are associated with a higher RR of gaze for the No-Feedback condition, while the opposite effect is observed for the Full-Dialogue condition.

To further analyze the patterns underlying gaze alignment, we focus on summary measures extracted from the CRP, graphed in Figure 3, with LME model coefficients reported in Table 1. Starting with recurrence rate (RR), which represents how likely it is that the dyads look at the same objects (irrespective of directionality), we find that RR is marginally higher in the No-Feedback

[7] There are 384 individual scenes rather than 100, because the position of the target was counterbalanced, and visual saliency and contextual congruency was manipulated in a subset of 38 scenes, as described above.

[8] Note that we did not introduce interactions as random slopes (e.g., `(0 + Feedback:Accuracy | Dyad)`), as the resulting models did not converge.

[9] Identical results are obtained fitting Generalized Linear Mixed Models using Markov Chain Monte Carlo techniques, with package `MCMCglmm`, (Hadfield, 2010).

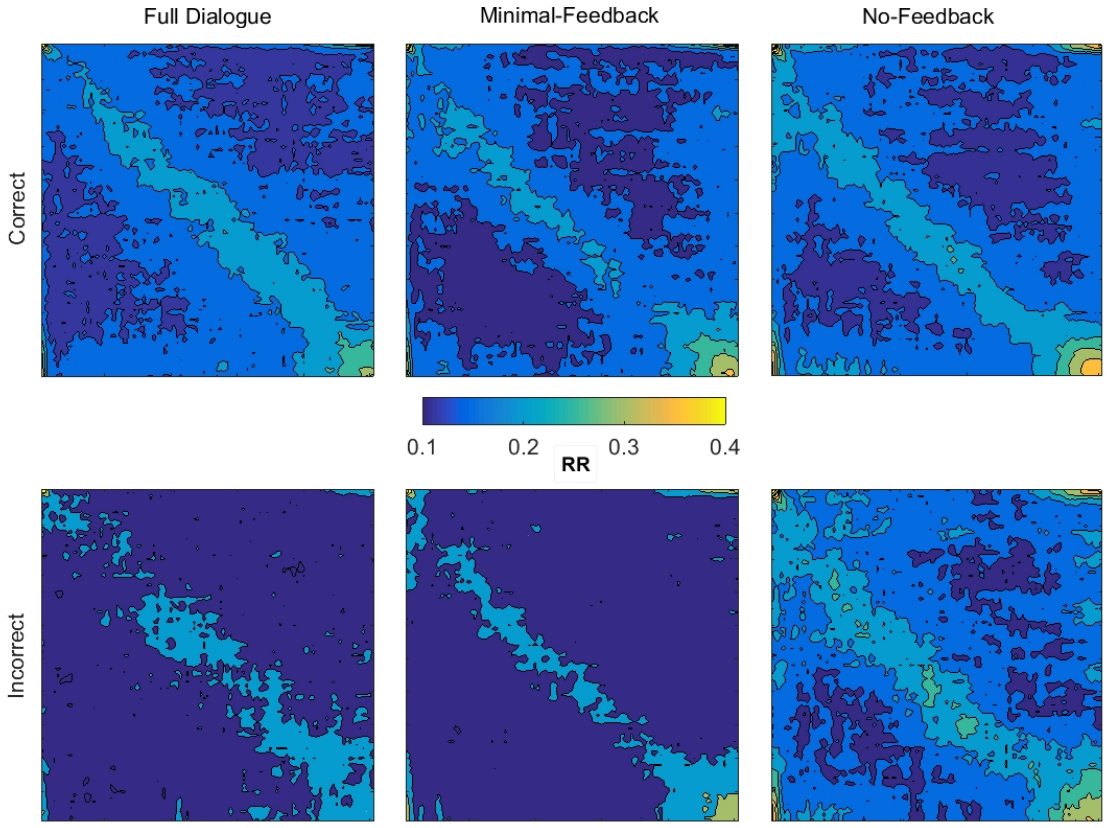

Figure 2. Heat map of recurrence rate for the cross-recurrence plot of gaze alignment crossing Feedback (No-Feedback, Minimal-Feedback, Full-Dialogue), and Accuracy (Correct, Incorrect). Recurrence values range from 0 to 0.4, as each heatmap was normalized to sum to 1. The color map used is jet, which goes from blue (low recurrence) to red (high recurrence).

condition than in the Full-Dialogue condition (marginal effect of No-Feedback in Table 1). Furthermore, RR is significantly higher for Incorrect trials than for Correct ones, but only in the No-Feedback condition (significant interaction of Accuracy:No-Feedback in Table 1). The same pattern is observed for L (length of the diagonal), which represents the average number of time-points along which the dyad aligns gaze. We find that L is significantly higher in the No-Feedback condition compared to Full-Dialogue, and that L is higher for Incorrect trials than for Correct trials, but only in the No-Feedback condition.

This indicates that the impossibility of exchanging feedback induces the listener to rely more strongly on the information delivered by the speaker; and hence they tend to look at the locations that the speaker has examined. This manifests itself as increased alignment. However, as our task is inherently a visual search task, too much alignment can mean that the listener is missing out visual information that is not directly referred to by the speaker, leading to the association between alignment and incorrect trials that we observe. If the listener can feed back information to the speaker,

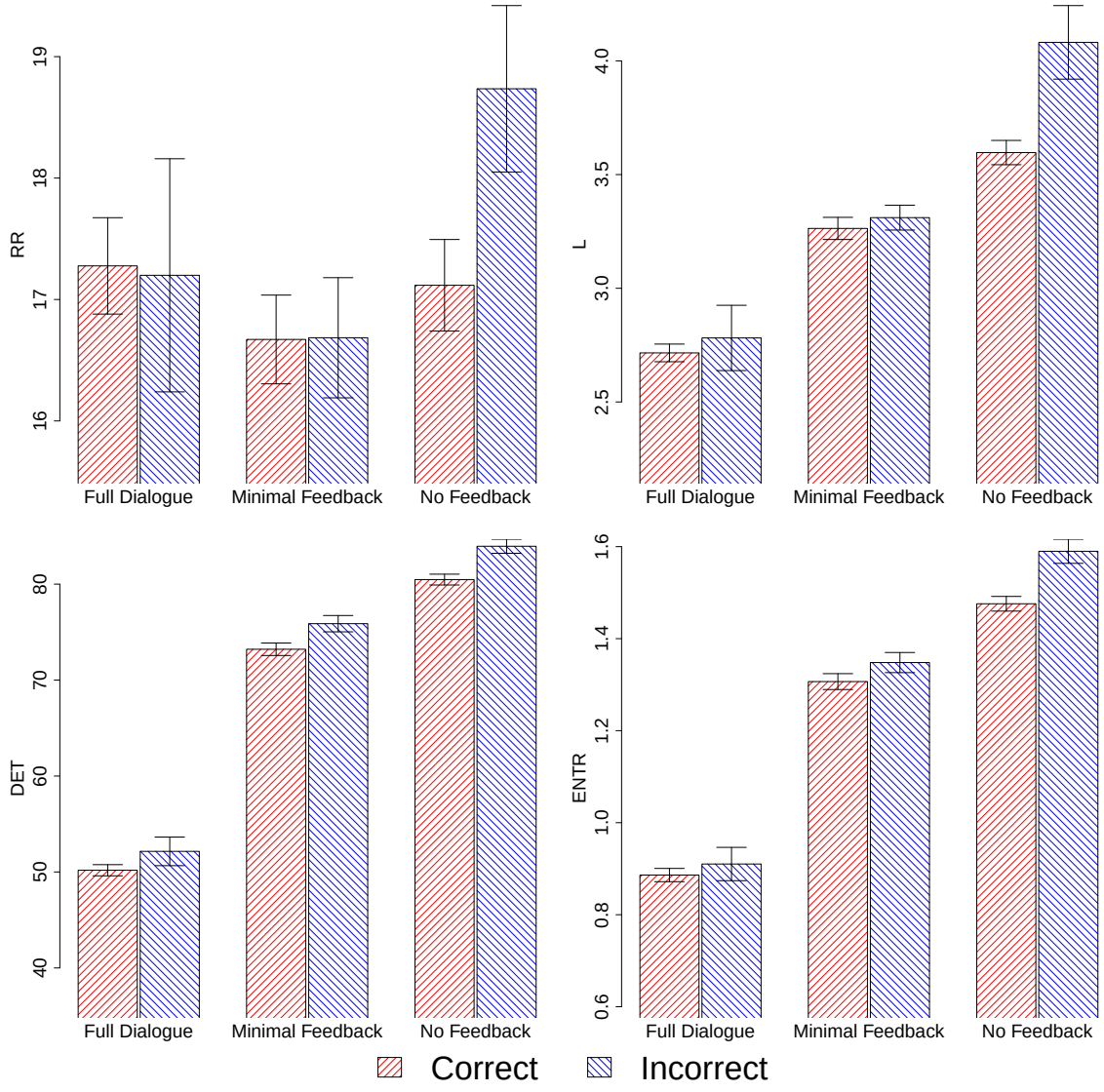

Figure 3. Bar plots for the gaze alignment recurrence measures of RR (recurrence rate), L (length of the diagonal line), DET (percentage determinism) and ENTR (entropy), mean and 95% CI, for the variables Feedback (No-Feedback, Minimal-Feedback, Full-Dialogue) and Accuracy, represented as oriented and colored lines (Correct: 45 degrees, red; Incorrect: −45 degrees, blue).

a better strategy for the dyad is to diversify their visual search strategy to increase the probability of finding out whether the scenes differ. This is presumably what happens in the Full-Dialogue and Minimal-Feedback conditions, where we observe decreased alignment, and no association between Accuracy and alignment.

When looking at determinism (DET), we find that both No-Feedback and Minimal-Feedback show significantly higher DET values than Full-Dialogue, meaning that the gaze alignment in these

	RR			L			DET			ENTR		
Fixed Effect	β	SE	t	β	SE	t	β	SE	t	β	SE	t
Intercept	0.21	0.45	0.46	0.03	0.07	0.48	0.29	1.11	0.26	0	0.02	0.94
Accuracy	0.04	0.46	0.1	-0.1	0.06	-1.61	-0.65	0.6	-1.04	-0.01	0.01	0.55
No-Feedback	0.95	0.51	1.86°	0.49	0.1	5.17***	9.04	1.52	5.92***	0.2	0.03	6.31***
Minimal-Feedback	-0.66	0.51	-1.3	-0.02	0.1	-0.25	5.07	1.52	3.31**	0.06	0.03	2.01*
Accuracy:No-Feedback	-1.30	0.56	-2.31*	-0.23	0.08	-2.71**	0.96	0.76	1.26	-0.01	0.02	0.42
Accuracy:Minimal-Feedback	0.45	0.57	0.79	0.08	0.08	0.97	-1.21	0.77	-1.56	0	0.02	0.73

$^{\circ}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 1

Coefficients of mixed-effects models for the dependent variables of RR, L, DET, ENTR, organized across columns, and modeled as a function of the predictors Feedback (sum-coded, with Full-Dialogue as the reference level for No-Feedback and Minimal-Feedback) and Accuracy (contrast-coded, Correct = 0.72, Incorrect = −0.22). We report coefficient β*, standard error, t-value and associated p-value. Random effects included are Dyad and Scene.*

conditions is more predictable. By exchanging information, the dyads are able to better divide their search space, which implies a less predictable pattern of alignment. Finally, when looking at the entropy of the gaze (ENTR), which represents how regular the phase of the alignment is, we find that both the No-Feedback and Minimal-Feedback condition have significantly higher entropy than the Full-Dialogue condition. Being able to exchange feedback helps the dyad to establish a more regular pattern of alignment, reducing entropy [10].

This analysis examined the visual attentional strategies of a dyad, and clearly indicated that in a spot-the-difference task, gaze alignment per se does not increase task success. Moreover, we found that feedback decreases rather than increases the attentional alignment of the dyad. This result is in line with previous literature on collaborative search tasks (S. Brennan et al., 2008; Neider et al., 2010), where the presence of feedback (especially a gaze cue) was shown to diversify the dyad's search strategies and improve their response accuracy [11].

Gaze Recurrence over Time. In Figure 4, we visualize how gaze recurrence evolves as a function of the trial for the three Feedback conditions, and in Table 2, we report the coefficients of the mixed model. Gaze recurrence increases linearly as a function of Time, and it has an upward bowing trend, i.e., a decrease followed by an increase (main effects of Time^1 and Time^2). Crucially, for the dyads in the No-Feedback and Minimal-Feedback condition, gaze recurrence is significantly lower than in the Full-Dialogue condition. This result suggests that only when interlocutors can

[10] As RR varies between time series, we have re-run the analysis on L and ENTR but after having residualized them against RR. The results hold with the only noticeable differences being: a reduction in the t-value for the interaction between Accuracy:No-Feedback (from -2.71 to -2.01) for L, and an increase in t-value for the main effect of Minimal-Feedback (from 2.01 to 2.91) for ENTR.

[11] In order to make sure that these results are not a consequence of the normalisation procedure, we have computed CRQA on non-normalised sequences finding nearly identical results. The only noticeable difference was on RR with: a significant main effect of Accuracy, whereby RR was found higher for correct versus incorrect responses ($p = 0.04$), the No-Feedback effect was now found significant ($p < 0.01$) and the interaction between Accuracy:No-Feedback became stronger ($p < 0.001$).

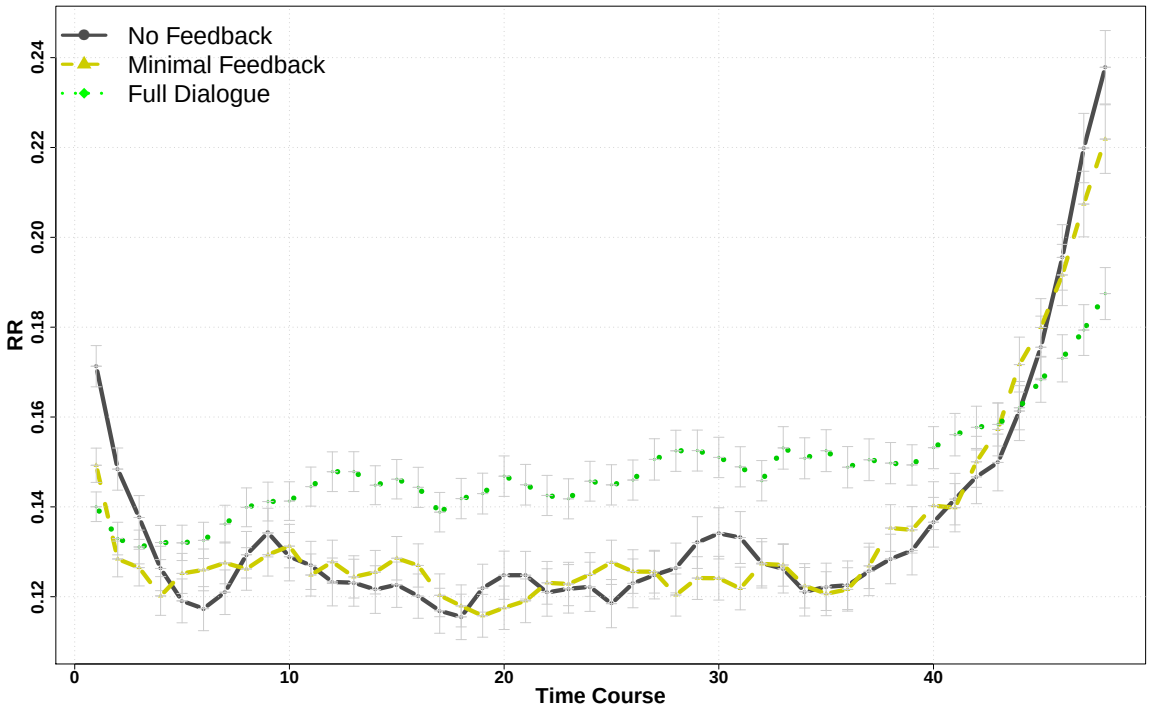

Figure 4. Windowed cross-recurrence of gaze (recurrence rate, RR) for correct trials only in the three feedback conditions (No-Feedback: black, tick line; Minimal-Feedback: yellow, dashed line; Full-Dialogue: green, dotted line), over time (50 points, which results from moving the window by a step of two over 101 points of the normalized scan-patterns).

fully interact, they manage to form and maintain aligned gaze, while also being successful at the task. When looking at the interactions between Feedback and Time, we observe that for the No-Feedback and Minimal-Feedback condition, gaze recurrence increases more over the trial than for Full-Dialogue, especially at the end of trial.

Interestingly, we also observe a sharp increase of recurrence rate both at the beginning and at the end of the trial, especially when dyads cannot exchange any feedback. This can be seen in the two-way interaction between No-Feedback and Time^2, which indicates an upward bowing trend of recurrence. In section *Additional measures influencing attention allocation and gaze alignment* of the Supplementary Material, we isolate additional bottom-up (visual saliency) and top-down (e.g., number of fixated objects) mechanisms that may underlie this trend. Moreover, in section *Time-course windowed analysis for the C/RQA measures of L, DET and ENTR*, we report also results for the time-course of the measures L, DET and ENTR, which entirely corroborate the results reported above, where the same measures are computed on the trial as a whole.

General Discussion

Research in dialogue has often assumed that interlocutors align their cognitive processes to maximize mutual understanding (Pickering & Garrod, 2004). Alignment emerges both in linguistic terms, such as converging on a common lexicon (S. E. Brennan & Clark, 1996), and in non-linguistic

Predictor	β	*SE*	*t*	*p*
Intercept	14.39	0.34	41.79	.0001
$Time^1$	8.38	1.37	6.10	.0001
$Time^2$	9.70	1.41	6.86	.0001
No-Feedback	-0.29	0.06	-4.65	.0001
Minimal-Feedback	-0.63	0.06	-10.08	.0001
$Time^1$:No-Feedback	0.82	0.43	1.90	.06
$Time^1$:Minimal-Feedback	0.82	0.43	1.92	.05
$Time^2$:No-Feedback	4.78	0.43	11.09	.0001
$Time^2$:Minimal-Feedback	1.77	0.43	4.16	.0001

Table 2

Windowed cross-recurrence. Coefficients of mixed-effects model of recurrence rate, modeled as a function of the predictors Feedback (sum-coded, with Full-Dialogue as the reference level for No-Feedback and Minimal Feedback), and Time represented as an orthogonal polynomial of order two ($Time^1$ and $Time^2$). We report coefficient β*, standard error, t-value and associated p-value. Random effects included are Dyad and Scene.*

responses, such as in postural sway (Shockley et al., 2003), or in the distribution of visual attention (Richardson & Dale, 2005). Most of the current models of dialogue agree, moreover, that the exchange of interactive cues between interlocutors plays a key role in the formation of a common ground (e.g., S. E. Brennan, 2004), the management of individual cognitive effort (e.g., Shintel & Keysar, 2009), the micro-dynamics of perceptuomotor coordination (e.g., Shockley et al., 2009), as well as the development of abstract communicative systems (e.g., Garrod, Fay, Lee, Oberlander, & MacLeod, 2007).

An important implication of this literature is that more alignment (indexing better mutual understanding of interlocutors) should result in more effective joint actions. Thus, in a task in which dyads need to share information to order to take a joint decision, increased alignment should predict higher task success. This hypothesis has been tested mostly using speech data collected during referential communicative tasks such as the map/maze task or the tangram task (e.g., Krauss & Glucksberg, 1969; Garrod & Anderson, 1987; A. H. Anderson et al., 1991); but the results have been mixed. On one hand, dyads who are lexically entrained, syntactically or lexically primed, are faster and less error-prone (e.g., Clark, 1996; Nenkova, Gravano, & Hirschberg, 2008; Foltz et al., 2015; Reitter & Moore, 2014). On the other hand, a large degree of knowledge overlap in the dyad, obtained through excessive entrainment, induces the dyad to commit more errors (Wu & Keysar, 2007). Other studies using alternative experimental paradigms have uncovered similar contradictions, whereby dyads collaborating on a perceptual task are more efficient than single participants (Bahrami et al., 2010), but this is not directly reflected in language use, where indiscriminate lexical alignment leads to lower performance (Fusaroli et al., 2012).

Interactivity between interlocutors is also a crucial component of task performance. Studies

that have examined this issue by looking at, for example, the role of overhearers, the physical co-presence of interlocutors, and the types of feedback used. The results have mostly demonstrated a positive correlation between interactivity and task performance (Krauss & Weinheimer, 1966; Schober & Clark, 1989; Clark & Krych, 2004).

However, the assumption that more interactivity and higher alignment would automatically imply more accurate task performance is not supported in all studies. The interplay between these three factors could in fact depend on the goals of the task, as well as on type of response observed. S. Brennan et al. (2008), and later Neider et al. (2010), for example, demonstrate that interacting with the partner improves target detection in a collaborative search task. However, feedback seems to foster the disalignment of the eye-movement responses of the dyad, rather than encouraging alignment (even though these studies did not explicitly test this claim). In fact, through feedback, the dyad diversifies the individual search strategy of the interlocutors, so as to increase the joint likelihood of finding the target.

In this paper, we set out to disentangle the relationship between alignment and task performance by focusing on the role of *interactivity*. In an eye-tracking dialogue experiment, we used a spot-the-difference task, in which dyads of participants had to guess whether they were looking at the same visual scene or not. We manipulated interactivity as the amount of information that could be shared between the participants in each dyad: the interlocutors could either exchange no feedback or minimal feedback (backchannels only), or they were allowed to engage in full dialogue.

We analyzed the experimental data using cross-recurrence quantification analysis on the eye-movements of speakers and listeners (Marwan & Kurths, 2002; Richardson & Dale, 2005; N. C. Anderson et al., 2013). The results show that the recurrence rate and determinism of gaze alignment is higher, and the mean diagonal longer, when the listener cannot exchange feedback with the speaker. Crucially, increased gaze alignment in the no-feedback condition was associated with significantly worse task performance. This result is consistent with the previous literature on collaborative search tasks (S. Brennan et al., 2008; Neider et al., 2010), in which the presence of feedback was shown to diversify the dyad's search strategies. Moreover, when looking at how gaze alignment is established during those trials that were answered correctly, we find that dyads are best able to form and maintain aligned gaze in the full-dialogue condition.

Performance in a visual search task is optimal when the members of the dyad diversify their strategies, i.e., the listener disengages to some extent from the precise visual implications of what the speaker is saying. This strategy is particularly successful when the dyad cannot exchange information, and therefore cannot form a shared common ground for the scene. The presence of feedback makes it possible to better divide the search space in the scene, and obtain a more organized attentional alignment (which manifests itself as lower CRQA entropy).

Overall, our results have three implications for current models of dialogue: (1) interactivity (feedback, in our case) directly mediates cognitive alignment and ultimately also task success; however (2) cognitive alignment is not directly associated with task success, as most models of dialogue have claimed; rather (3) alignment of gaze is negatively correlated with performance in a collaborative search task when feedback cannot be exchanged.

This study therefore poses important challenges to models of dialogue which uniquely center around alignment (e.g., the Interactive Alignment Model, Pickering & Garrod, 2004). We find that alignment per se cannot be taken as a proxy for effective communication. In fact, dyads align their gaze to compensate for the lack of feedback, rather than the other way around (i.e., aligning their gaze because of feedback). The importance for such compensatory mechanisms are recognized by

recent models of dialogue, which give a prominent role to interpersonal synergy, and envision communicative dialogue as a fluid experience, in which alignment and disalignment can both be strategies to reach shared understanding and task success (e.g., Dale et al., 2013; Mills, 2014; Fusaroli et al., 2014).

What is emerging from this literature, and from our study, is therefore a new model of dialogue in which the type of interaction between the interlocutors is crucial, as it determines whether they are able to develop an optimal strategy for the task they are trying to solve. In our case, full-dialogue interaction (and to a lesser degree minimal-feedback interaction) makes it possible for the dialogue partners to deploy a strategy that relies on division of labor to efficiently search a visual scene, resulting in increased task success. As a consequence of this strategy, alignment is reduced in full dialogue compared to less interactive conditions. In this scenario, alignment is a consequence of interaction type and task. This differs markedly from the assumption that a cascade of alignment underpins successful dialogue per se (as in the Interactive Alignment Model).

Note that the claims we can make based on the present study are limited to gaze alignment during linguistic interaction. However, alignment processes occur across a range of other domains, affecting a variety of coordinative behaviors people engage in. This includes simple joint tasks to large collective activities such as war (McNeill, 1997). Also, the growing literature on joint action is seeking a mechanistic understanding of how we coordinate with each other both in laboratory settings and in more natural tasks (Sebanz, Bekkering, & Knoblich, 2006). All these domains invoke different levels of analysis. For example, the capacity to coordinate musically, such as in a duet, is not merely a matter of "getting the notes right," but involves using various multimodal signals to guide and structure each other's musical behavior (Kawase, 2014). But in these non-linguistic domains, the same principles that we observed in this study may hold. Different interactive tasks demand a balance of alignment and disalignment, suggesting that task success depends on a mixture of behavioral and cognitive strategies. In the music case, for example, aligning too much may sound odd. During jazz improvisation, for example, it is common to align on particular motifs, but success in such improvisation also involves moving away from these aligned motifs in new and different ways (e.g., Walton, Richardson, Langland-Hassan, & Chemero, 2015).

Our study raises new questions, which we aim to address in future research. The most important limitation is that we only measured gaze alignment. In the full-dialogue condition, it is also possible to observe linguistic alignment between the two dialogue partners. This can include re-use of lexical items, syntactic categories, grammar rules, or whole constructions. It is perfectly possible that linguistic alignment behaves differently from gaze alignment, for example in that more alignment increased task success (as in Reitter and Moore's (2014) study using the Maptask corpus). It is important to note that an analysis of linguistic priming between interlocutors in not possible in the no-feedback and minimal-feedback condition, as only one of the dialogue partners is allowed to speak in these conditions. This means that a direct investigation of the interaction between the amount of feedback and the amount of linguistic alignment (and their effect on task success) is not possible. However, we could instead measure self-alignment (the degree to which a speaker repeats their own linguistic items) in the no-feedback and minimal-feedback conditions.[12]

Another related item of future work is the issue communicative efficiency, in particular the question how the structure of utterances changes as the experimental session develops (e.g., do

[12] A recent paper by Fusaroli and Tylén (2016) suggests using all the language production within a trial as the unit of CRQA; this effectively offers a way of analyzing self-alignment in the no-feedback and minimal-feedback conditions, and should be explored in future work.

utterances become shorter as the task progresses or if more alignment occurs?). Another possible avenue for future research regards the issue of decision making per se. The task used in this study was simple, and decision-making performance was evaluated on the basis of a single goal. More complex decisions involving multiple goals are likely to show more interesting dynamics of alignment, where dyads couple and decouple their cognitive processes according to the necessity of the goal currently being attempted.

Overall, our study contributes novel insights into the dynamics of alignment across different modalities, relates alignment to interactivity, and elucidates how alignment and interactivity conspire to influence task performance. This study also poses new challenges for models of dialogue that aspire to explain alignment phenomena beyond the domain of language processing.

Acknowledgments

European Research Council under award number 203427 "Synchronous Linguistic and Visual Processing" to FK, and the Leverhulme Trust under award number ECF-2014-205 to MIC, are gratefully acknowledged.

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., ... others (1991). The hcrc map task corpus. *Language and speech*, *34*(4), 351–366.

Anderson, N. C., Bischof, W. F., Laidlaw, K. E., Risko, E. F., & Kingstone, A. (2013). Recurrence quantification analysis of eye movements. *Behavior research methods*, *45*(3), 842–856.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010, August). Optimally interacting minds. *Science*, *329*, 1081–5.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bates, D., Martin, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed effects models using eigen and s4.(r package v. 1.0-7).*

Boker, S., Xu, M., Rotondo, J., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, *7*, 338–355.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–B25.

Brennan, S., Chen, X., Dickinson, C. A., Neider, M. B., & Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, *106*, 1465–1477.

Brennan, S. E. (2004). How conversation is shaped by visual and spoken evidence. In *J. Trueswell & M. Tanenhaus (Eds.), Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (p. 95-129). MIT Press.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *22*, 1482–1493.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, *50*(1), 62–81.

Coco, M., & Dale, R. (2014). Cross-recurrence quantification analysis of behavioral streams: An r package. *Quantitative Psychology and Measurement*, *5*, 510.

Coco, M., & Keller, F. (2012). Scan patterns predict sentence production in the cross-modal processing of visual scenes. *Cognitive science*, *36*(7), 1204–1223.

Coco, M., Malcolm, G., & Keller, F. (2014). The interplay of bottom-up and top-down mechanisms in visual guidance during object naming. *Quarterly Journal of Experimental Psychology*, *67*(6), 1096–1120.

Coco, M. I., & Dale, R. (2014). crqa: Cross-recurrence quantification analysis for categorical and continuous time-series [Computer software manual]. (R package version 1.0.5)

Dale, R., Fusaroli, R., Duran, N., & Richardson, D. (2013). The self-organization of human interaction. *Psychology of Learning and Motivation: Advances in Research and Theory*, *59*, 43–95.

Dale, R., Kirkham, N. Z., & Richardson, D. C. (2011). The dynamics of reference and shared visual attention. *Frontiers in Psychology*, *2*.

Dale, R., Warlaumont, A. S., & Richardson, D. C. (2011). Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *International Journal of Bifurcation and Chaos*, *21*, 1153–1161.

Foltz, A., Gaspers, J., Meyer, C., Thiele, K., Cimiano, P., & Stenneken, P. (2015). Temporal effects of alignment in text-based, task-oriented discourse. *Discourse Processes*.

Fusaroli, R., Bahrami, B., Olsen, K., Rees, G., Frith, C. D., Roepstorff, A., & Tylén, K. (2012). Coming to terms: an experimental quantification of the coordinative benefits of linguistic interaction. *Psychological Science*, *23*. (8)

Fusaroli, R., Raczaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147–157.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, *40*(1), 145–71.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987.

Glucksberg, S., Krauss, R., & Higgins, E. T. (1975). The development of referential communication skills. In F. Horowitz (Ed.), *Review of child development research* (Vol. 4, p. 305-345). University of Chicago Press.

Hadfield, J. D. (2010). Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22. Retrieved from `http://www.jstatsoft.org/v33/i02/`

Hupet, M., & Chantraine, Y. (1992). Changes in repeated references: Collaboration or repetition effects? *Journal of psycholinguistic research*, *21*(6), 485–496.

Ireland, M. E., & Henderson, M. D. (2014). Language style matching, engagement, and impasse in negotiations. *Negotiation and Conflict Management Research*, *7*(1), 1–16.

Kawase, S. (2014). Gazing behavior and coordination during piano duo performance. *Attention, Perception, & Psychophysics*, *76*(2), 527–540.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*, 32–38.

Krauss, R. M., & Glucksberg, S. (1969). The development of communication: Competence as a function of age. *Child Development*, 255–266.

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*, 343–346.

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2014). lmertest: Tests in linear mixed effects models [Computer software manual]. (R package version 2.0-20)

Kuznetsova, A., Christensen, R. H., Bavay, C., & Brockhoff, P. B. (2015). Automated mixed anova modeling of sensory and consumer data. *Food Quality and Preference*, *40*, 31–38.

Louwerse, M. M., Dale, R., Bard, E. G., & Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. *Cognitive science*, *36*(8), 1404–1426.

Marwan, N., Carmen Romano, M., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, *438*(5), 237–329.

Marwan, N., & Kurths, J. (2002). Nonlinear analysis of bivariate data with cross recurrence plots. *Physics Letters A*, *302*, 299-307.

McNeill, W. H. (1997). *Keeping together in time*. Harvard University Press.

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

Neider, M. B., Chen, X., Dickinson, C. A., Brennan, S. E., & Zelinsky, G. J. (2010). Coordinating spatial referencing using shared gaze. *Psychonomic bulletin & review*, *17*(5), 718–724.

Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers* (pp. 169–172).

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, *171*(3968), 308–311.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–189.

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29–46.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, *29*, 39–54.

Richardson, D. C., Dale, R., & Kirkham, N. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science*, *18*, 407–413.

Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1-3), 151-173.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *language*, 696–735.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, 110–114.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in cognitive sciences*, *10*(2), 70–76.

Shintel, H., & Keysar, B. (2009). Less is more: A minimalist account of joint action in communication. *Topics in Cognitive Science*, *1*(2), 260–273.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*(2), 305–319.

Shockley, K., Santana, M., & Fowler, C. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology Human Perception and Performance*, *29*, 326-332.

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? evidence from eye movements during the inspection of real-world scenes. *Consciousness and cognition*, *17*(1), 159–170.

Walton, A. E., Richardson, M. J., Langland-Hassan, P., & Chemero, A. (2015). Improvisation and the self-organization of multiple musical bodies. *Frontiers in psychology*, *6*.

Wu, S., & Keysar, B. (2007). The effect of information overlap on communication effectiveness. *Cognitive Science*, *31*(1), 169–181.

Zbilut, J., Giuliani, A., & Webber, C. (1998). Recurrence quantification analysis and principal components in the detection of short complex signals. *Physics Letters A*, *237*(3), 131–135.

Appendix: Diagonal-Recurrence of Gaze

In this appendix we report results from the diagonal-wise cross-recurrence profile, which is where gaze alignment is expected to occur (around ± 50 normalized lags from the diagonal). These

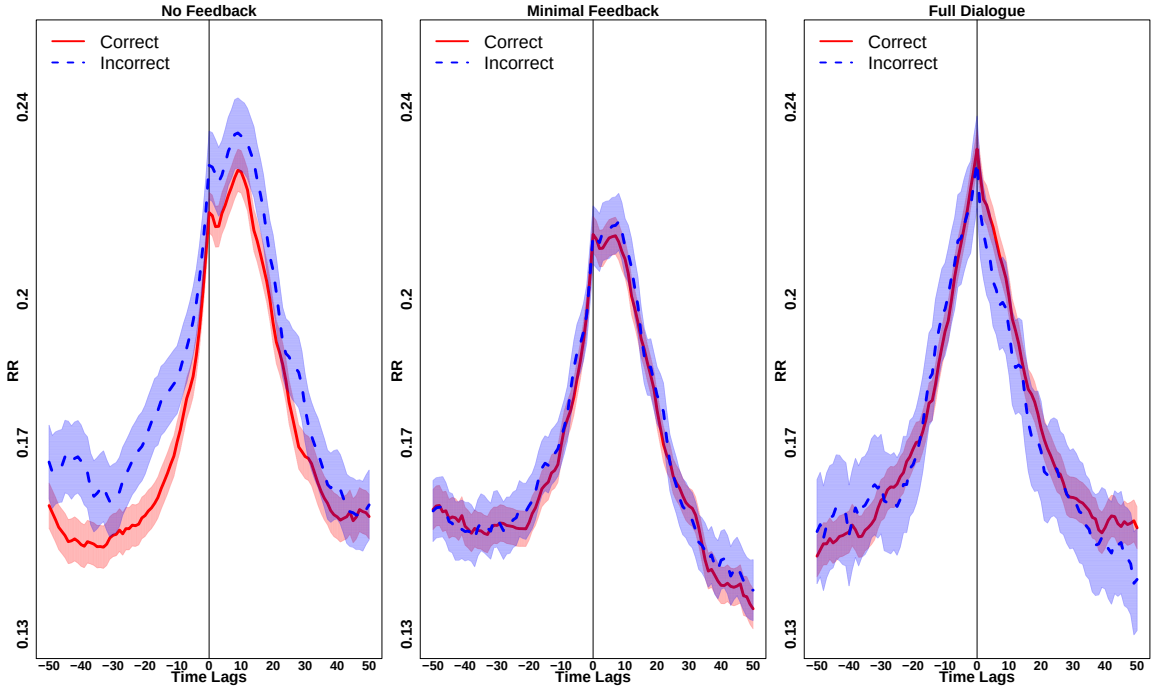

Figure 5. Diagonal-wise gaze alignment (recurrence) of the dyads' eye-movements as a function of the lag (±50), for Incorrect (blue), and Correct (red) responses in the three feedback conditions (No-Feedback, Minimal-Feedback, Full-Dialogue). Recurrence ranges from 0 to 1, with 1 being perfect alignment between speakers and listeners. Lines represent means, and the shaded bands the standard errors around the means. In the Full-Dialogue condition, we observe a peak at zero probably because there is no distinction between speaker and listener.

analyses are meant to corroborate the results presented in the main text. Diagonal-wise recurrence is mostly used in work on dialogue to show how within a certain time frame, e.g., three seconds, dyads of interlocutors align their gaze, and especially, if there is a leading-follower pattern (e.g., Dale, Kirkham, & Richardson, 2011).

We adopt the convention of previous studies, in which positive lags indicate a speaker-leading cross-recurrence pattern, i.e., the eye-movements of the speaker are ahead of those of the listener, and negative lags indicate a listener-leading pattern. We compute separate recurrence profiles for correct and incorrect responses, and for the three feedback conditions. From the recurrence profile, we extract six measures characterizing its distribution: mean recurrence, maximum recurrence, kurtosis, dispersion, central tendency, and maximum lag (refer to Dale, Warlaumont, & Richardson, 2011, where this approach was first proposed). We model each of these dependent variables as a function of Feedback and Accuracy using LMEs (refer to Section for details about the analysis).

In Figure 5, we visualize how attentional alignment is mediated by Feedback and Accuracy. We find a higher mean and maximum recurrence for the No-Feedback condition, especially when incorrect responses are made, as compared to the Full-Dialogue condition. Moreover, for the No-Feedback condition, we also observe a higher kurtosis and dispersion, which indicate the presence of coordination within a small lag window. When there is no feedback and the responses are incorrect,

	mean RR		max RR		kt		sd		ct		lagmax	
Fixed Effect	β	t	β	t	β	t	β	t	β	t	β	t
Intercept	0.002	0.33	0.002	0.334	0.001	0.21	0.001	0.26	0.1	0.3	0.001	0.26
Accuracy	0	0.07	0	0.087	-0.003	-0.73	0.001	0.14	-0.07	-0.14	0.001	0.14
No-Feedback	0.014	2.69 **	0.014	2.694 **	0.013	2.45 **	0.014	2.68 **	-0.78	-1.77 ◦	0.014	2.68 **
Minimal-Feedback	-0.003	-0.56	-0.003	-0.565	-0.002	-0.35	-0.003	-0.56 **	-0.38	-0.86	-0.003	-0.56
Accuracy:No-Feedback	-0.013	-2.24 *	-0.013	-2.246 *	-0.01	-1.71 ◦	-0.013	-2.23	0.43	0.61	-0.013	-2.23 *
Accuracy:Minimal-Feedback	0.005	0.89	0.005	0.891	0.003	0.53	0.005	0.89	0.79	1.12	0.005	0.89

◦ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3

Coefficients of mixed-effects models for the dependent variables extracted from the diagonal-profile: mean recurrence (mean RR), maximum recurrence (max RR), kurtosis (kt), dispersion (sd), central tendency (ct), and maximum lag (lagmax), organized across columns. We model these measures as a function of the predictors Feedback (sum-coded, with Full-Dialogue as the reference level for No-Feedback and Minimal-Feedback) and Accuracy (contrast-coded, Correct = 0.72, Incorrect = −0.22). We report β, t-value and associated p-value. Random effects included are Dyad and Scene.

in contrast, such coordination is found within broader windows. When looking at the maximum lag, we find that in the No-Feedback condition, it is more likely to happen at positive lags (i.e., a speaker-leading pattern), in line with existing literature (e.g., Richardson & Dale, 2005). In full dialogue, there is no such a dominance. The full interactive nature of the dialogue removes any directionality due to leader-follower roles; and it might be that half of the time one interlocutors acts as a 'speaker', while the other half of the time is the other that does it.

Confirming what we found with the summary measures of CRP reported in the main text, these results suggest that the best strategy for the search task, in the absence of the ability to exchange feedback, is for the listener to utilize the information provided by the speaker in a complementary way, i.e., to diversify their allocation of visual attention. In this way, the dyad can jointly maximize the portion of the scenes attended to, thus increasing the likelihood of establishing correctly whether the scenes are different or not.