



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Feedback quality and divided attention

Exploring commentaries on alignment in task-oriented dialogue

Citation for published version:

Crible, L, Gandolfi, G & Pickering, MJ 2024, 'Feedback quality and divided attention: Exploring commentaries on alignment in task-oriented dialogue', *Language and Cognition*.
<https://doi.org/10.1017/langcog.2023.65>

Digital Object Identifier (DOI):

[10.1017/langcog.2023.65](https://doi.org/10.1017/langcog.2023.65)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Language and Cognition

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.



Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



ARTICLE

Feedback quality and divided attention: exploring commentaries on alignment in task-oriented dialogue

Ludivine Crible¹ , Greta Gandolfi²  and Martin J. Pickering²

¹Ghent University, Linguistics Department, Gent, Belgium; ²University of Edinburgh, Psychology Department, Edinburgh, UK

Corresponding author: Ludivine Crible; Email: ludivine.crible@ugent.be

(Received 31 January 2023; Revised 24 October 2023; Accepted 27 November 2023)

Abstract

While studies have shown the importance of listener feedback in dialogue, we still know little about the factors that impact its quality. Feedback can indicate either that the addressee is aligning with the speaker (i.e. ‘positive’ feedback) or that there is some communicative trouble (i.e. ‘negative’ feedback). This study provides an in-depth account of listener feedback in task-oriented dialogue (a director–matcher game), where positive and negative feedback is produced, thus expressing both alignment and misalignment. By manipulating the listener’s cognitive load through a secondary mental task, we measure the effect of divided attention on the quantity and quality of feedback. Our quantitative analysis shows that performance and feedback quantity remain stable across cognitive load conditions, but that the timing and novelty of feedback vary: turns are produced after longer pauses when attention is divided between two speech-focused tasks, and they are more economical (i.e. include more other-repetitions) when unrelated words need to be retained in memory. These findings confirm that cognitive load impacts the quality of listener feedback. Finally, we found that positive feedback is more often generic and shorter than negative feedback and that its proportion increases over time.

Keywords: cognitive load; director–matcher game; feedback; interactive alignment; repetitions

1. Introduction

Dialogue implies the active participation of two or more speakers. This participation is, however, not necessarily balanced in quantity and quality between speakers. Everyone has experienced situations in which one interlocutor speaks much more than the other, such as a friend telling a joke or a realtor showing a house to potential buyers. In such asymmetric contexts, where speakers typically perform some joint task, the addressee’s speech has been termed ‘listener responses’ (Bavelas et al., 2000),



'feedback' (Tree & Jean, 1999), or 'commentaries' (Pickering & Garrod, 2021). These terms stress the metalinguistic nature of these contributions, which help in managing the interaction but, for the most part, provide little new information content themselves. While these commentaries or feedback also frequently occur in more symmetric interactions (e.g. a free conversation), they tend to be less frequent than in asymmetric task-oriented contexts (Bangerter & Herbert, 2003; Fusaroli et al., 2017).

While it differs in form and function from the main speaker's contribution, listener feedback is nevertheless crucial for the success of an interaction, as several studies have shown. Bavelas et al. (2000) go so far as calling listeners 'co-narrators', thus rejecting sender–receiver accounts of dialogue. Insights on the role of the listener go back to early experimental research in the 1960s (e.g. Krauss & Weinheimer, 1964), where the effect of feedback on the listener's comprehension and on the speaker's behaviour was already demonstrated. Results converge in suggesting that a listener who is able to provide feedback better integrates information (Schober & Clark, 1989) and that, in turn, the speaker improves the quality of her speech (Bavelas et al., 2000). In this paper, we shift the focus towards the quality of the feedback itself, by measuring the effect of cognitive load on the listener's behaviour.

One important notion to understand the role of feedback is that of interactive alignment (Pickering & Garrod, 2004, 2021). Alignment is defined as "the extent to which individuals represent things in the same way as each other" (Pickering & Garrod, 2021: 1) and is interactive when it is the result of interaction. These shared representations cover not only words and content but also interactive roles, pragmatic intents, and the evolving situation (discourse) model. Situation model alignment is therefore a high-level mental process (related to what Clark, 1996 refers to as *common ground*) that is not always observable, unlike lexical or syntactic alignment, which is a lower-level process reflected in the repetition of words and structures produced by the other speaker. Misalignment, in turn, emerges in case of a mismatch between speakers' beliefs and knowledge and is typical in task-oriented settings where speakers have a common goal but different information (Dideriksen et al., 2022). It is this type of interaction that we focus on here.

Feedback and interactive alignment can take many forms. In settings of free conversation, where speakers are not performing a specific joint task, much of the feedback is what Pickering and Garrod (2021) call *positive commentaries* – that is, feedback that expresses successful alignment ('I understand you, I am on board with this interaction'). These commentaries can be short and generic (e.g. *okay, alright*) or long and specific (e.g. repetitions of the speaker's words). Bavelas et al. (2000) found out that specific and generic commentaries were distributed differently in the course of an interaction and that both were affected by whether or not the listener was performing a simultaneous task, thus experiencing high cognitive load. However, the speakers in their study were telling a story, a context in which the listener has (at least quantitatively) a minor role. By contrast, it remains unknown whether similar effects of cognitive load would be found in more collaborative contexts, where negative commentaries (expressing misalignment/misunderstanding) have been shown to be more frequent (Dideriksen et al., 2022).

Against this backdrop, the present study adopts a comprehensive approach to feedback (positive and negative, specific and generic) in a director–matcher game using tangrams (Clark & Wilkes-Gibbs, 1986). In this highly collaborative, task-oriented setting, the listener is more active and has to identify abstract geometrical shapes described by the speaker. We also manipulated cognitive load by giving the listener a secondary mental task, in order to measure its effect on the production of

different types of commentaries. Our view of cognitive load follows Mattys et al.'s (2012) definition as 'any factor placing unusually high demands on central attentional and mnemonic capacities'. It relates to divided attention since these demands have been shown to increase when an individual performs multiple simultaneous tasks (such as speaking while driving) (Baldock et al., 2019; Strayer et al., 2011). Our research questions are thus the following: What types of feedback are produced in task-oriented dialogue? How is feedback quality impacted by divided attention?

1.1. Listener feedback impacts speaker's behaviour

It has long been established that the metalinguistic contributions typically produced by a listener, such as *mhmm* or *okay*, bear a crucial impact on the success of an interaction. Clark and Wilkes-Gibbs (1986) famously proposed the principle of 'least collaborative effort', according to which interactants try to minimize their joint difficulty. In their tangram game, they argued that the participants did this by agreeing on the most efficient way to refer to complex shapes. This negotiation happens through listener feedback that can either accept or reject descriptions and pinpoint potential problems: "addressees minimize collaborative effort by indicating *quickly* and *informatively* what is needed for mutual acceptance" (1986: 27), for instance, by repeating a description which they cannot identify (A: "Uh, person putting a shoe on." – B: "putting a shoe on?", 1986: 22). Thanks to this feedback, efficient references are jointly constructed, which leads to a drastic reduction of description length throughout the interaction. This efficiency is thus achieved not only by the sheer presence of feedback but also by its timing and informativeness. In a more recent study building on Garrod and Anderson's (1987) Maze Task, Healey et al. (2018) found that successful coordination between speakers depended more on devices that cope with misunderstandings (which Pickering and Garrod, 2021 would call negative commentaries) than on confirmations of understanding (or positive commentaries). This suggests that negative feedback is more informative and more impactful on the success of the joint task.

Even in the absence of a common goal or task, listener feedback has a direct impact on the interaction and on the quality of the speaker's contributions. Bavelas et al. (2000) conducted dialogue experiments in story-telling contexts in English, in which they manipulated the level of distraction of the listener: they were either focused on the story (Experiment 1: just listening vs listening to summarize vs listening to retell the story) or distracted by a second task (Experiment 2: listening while counting the number of words starting with *t*). The narrators' stories were then rated for the quality of their plot and delivery by independent raters, and the scores showed lower story-telling quality when the listener was distracted: "No matter how good the story plot is, a good listener is crucial to telling it well" (2000: 947).

Corpus-linguistic approaches to the effect of feedback also reveal differences in the speech that follows various types of feedback. Tolins et al. (2014) took up Bavelas et al.'s (2000) distinction between specific and generic backchannels (i.e. closely tied to the previous context versus reusable in any context) and applied it to analyze how speakers continue their narrative. They found that speakers produce new information (the next event in the narrative) after generic backchannels, while they elaborate on the previous information after specific backchannels. This suggests that specific

feedback highlights story elements that deserve an explanation or an elaboration, while generic feedback only cues the speaker to go on.

In the context of repair (that occurs when communication goes off-track, in other words misalignment), feedback has an even stronger role of identifying what went wrong and how it should be repaired. Dingemanse et al. (2015: 2) distinguish between ‘specific repair initiation techniques’ (e.g. *who?*) and more generic ones (e.g. *huh?*). The former can be either a ‘restricted request’ (about a specific component of the trouble source) or a ‘restricted offer’ (the listener offers a candidate and asks for confirmation), while the latter are called ‘open requests’ (signalling trouble without specifying what it is). Dingemanse et al. (2015) noted that interjections, question markers, prosody, or repetition can signal a repair; they observe that repetitions are responsible for about half of their data, which “point[s] to the importance of other-initiated repair as a mechanism for achieving interactive alignment” (2015: 5). Their corpus study shows that open requests lead to longer repair solutions than restricted offers, presumably because the little information in generic feedback does not allow the speaker to make the most efficient (condensed, to the point) contribution to bringing the dialogue back on track.

More recently, Dideriksen et al. (2022) and Fusaroli et al. (2017) showed that types of repair vary across tasks: specific repairs (Dingemanse et al.’s 2015 ‘restricted offers’) were more frequent in the asymmetrical Map Task than in a more symmetrical game. Healey et al. (2018) further used a text-chat tool to manipulate the explicitness of the negative (and positive) feedback that matchers write while performing the Maze Task. The tool automatically (and secretly) transformed clarification requests such as ‘top left?’ to ‘what?’ – in other words, transforming negative feedback from specific to generic. They found that directors shifted towards abstract descriptions of the maze after such generic feedback more often than controls with no manipulation of the repair type. These studies provide evidence that repair use shapes the interlocutor’s contributions.

There is thus strong evidence, from both psycholinguistic and corpus-linguistic studies, that listener feedback is not a secondary feature of dialogue but has a strong impact on the quality of the interaction. But what are the factors that impact the quality of feedback itself?

1.2. Factors impacting feedback in dialogue

As mentioned above, the type of feedback varies across contexts and tasks (e.g. Dideriksen et al., 2022). Knutsen et al. (2018) further explored the use of procedural coordination devices, which indicate steps in a joint task such as turn-taking or moving on to the next subtask, during a card-matching game. They compared a classic condition that always involves the same cards with a more difficult condition in which participants had to arrange new cards at each trial. The authors found that matchers were affected by this difference in difficulty: they produced more coordination devices related to card placement (‘it’s the second card’), to the task progress (‘I see it’), and to the general logistics of the game (‘are we done?’) in the new cards condition than in the easier classic condition; generic acknowledgement devices (typically discourse markers like *yeah*, *okay*, *mhm*, which they refer to as ‘generic implicit coordination’) were not affected by the task. They conclude that, in the classic condition, the ability to coordinate semantically through the establishment

of ‘conceptual pacts’ (i.e. tacit agreements about referring expressions; Brennan & Clark, 1996) across trials triggers a higher use of feedback for procedural coordination: when semantic demands are high, procedural demands remain high as well. In other words, it appears that a demanding task requires more feedback, although this does not apply to generic backchannels.

While Knutsen et al.’s (2018) manipulation of difficulty prevents the matching task from becoming easier round after round, it does not substantially increase the cognitive load of the matcher, for instance, by adding an extra concurrent task. By contrast, the effect of divided attention on feedback has been documented by Bavelas et al. (2000) for non-task-oriented conversation. Their study not only showed the effect of feedback on story quality, but also revealed that, in story-telling, different types of feedback are impacted by external factors. The main finding from their Experiment 2 is that both specific and generic responses were significantly reduced in the condition with the mental counting task compared to the baseline condition with no extra load. This effect was stronger for specific responses, which were almost completely suppressed under high cognitive load. Divided attention has thus an important suppressing impact on feedback in story-telling contexts. This result stands in contrast with Knutsen et al.’s (2018) observed increase in feedback in the more difficult (new cards) condition. However, it should be noted that the feedback investigated in Bavelas et al.’s (2000) study roughly corresponds to Knutsen et al.’s generic implicit devices, which were not impacted by the manipulation, and that the data are very different in the two studies (matching game versus story-telling).

Another interesting result from Bavelas et al. (2000) is that there was no effect of cognitive load on feedback when the secondary task was focused on the story content (Experiment 1), contrary to the reduction in feedback observed in the unrelated counting task (Experiment 2). The authors concluded that “as long as listeners are attending to the meaning of the story, they are capable of handling several simultaneous demands and responses” (2000: 947). In other words, the type of mental task (and its related cognitive demands) does not matter as long as the listener does not have to divide her attention between the story and an unrelated task. But note that in this study, listeners are relatively inactive since they do not need to collaborate with the speaker to achieve a common goal.

Colman and Healey’s (2011) corpus study also provides some insight into factors that impact the production of feedback. They compared the production of self-repair and other-repair in two corpora: ordinary conversation in the British National Corpus and task-oriented dialogue from the Map Task corpus (Anderson et al., 1991). Although they did not systematically distinguish between self-repair and other-repair or between speakers in their analyses, it appears that, in general, “the arguably greater processing demands of the task-oriented dialogues are reflected in more repair” (Colman & Healey 2011: 1565–1566). The specific effects of processing demands on other-repair produced by the listener (i.e. ‘matchers’ in a tangram game, ‘route followers’ in the Map Task) remain to be examined.

In pursuit of a similar objective, Dingemanse et al. (2015) looked at when repair occurs in conversation. Although they did not systematically investigate cognitive load, they manually coded trouble-prone contexts such as noise, distraction, or a parallel activity during the dialogue. Their analysis shows that generic feedback (‘open requests’ in their words) is more frequent when the listener has trouble

hearing or processing what the speaker says. Specific feedback, in turn, is more frequently used if the source of the repair is an answer (rather than a question) or if it is relatively long. In other words, within negative feedback, specific responses tend to be produced in relation to informative linguistic content, while generic responses may relate more to external interference.

Experimental and corpus-based findings once more converge to suggest that language-external factors (nature of the task, cognitive load) affect the quantity and type of feedback. The studies by Bavelas et al. (2000) and Dingemans et al. (2015) reviewed above differ, however, on a major feature of feedback: story-telling contexts mainly (almost exclusively) lead to ‘positive’ commentaries (expressing alignment), while repair signals are by definition ‘negative’ (expressing misalignment) – or at least those are the only types of feedback that were discussed in these papers. As for Knutsen et al. (2018), their data seem to include both positive and negative feedback, although no quantitative information is provided on their distribution across tasks. They also refer to these devices as ‘generic’ and ‘specific’, but it is difficult to tell how their categories exactly correspond to the distinction in Bavelas et al. (2000) (cf. also ‘continuers’ versus ‘assessments’ in Goodwin, 1986). In sum, the way that the specific versus generic distinction interacts with the positive versus negative feature in a single type of context remains to be uncovered.

1.3. Feedback quality and cognitive load: predictions for task-oriented dialogue

For all we know about different types of listener responses and the impact of cognitive load on feedback, we still lack a comprehensive account that integrates all these factors and explores their interaction. In this study, we bring together elements from the studies reviewed above and measure the effect of divided attention in a task-oriented context, the tangram game. This setting is prone to both positive and negative commentaries on alignment. We introduce for the first time a manipulation of cognitive load in this task to refine Bavelas et al.’s (2000) findings on story-telling, to extend them to French dialogue, and to explore the following research questions: Does divided attention affect generic versus specific responses in task-oriented dialogue in the same way that it does in story-telling? Is there also an effect on the distribution of positive versus negative responses? Furthermore, we test the effect of cognitive load not only on the frequency of these commentaries but also on their timing (pause duration before the listener’s turn) and on their novelty (rate of other-repetitions in the listener’s commentaries), and the evolution of these features throughout the course of the dialogue.

If task-oriented dialogue is similar to story-telling as far as the listener’s behaviour is concerned (which is not necessarily the case), then we can expect to replicate Bavelas et al.’s (2000) finding of divided attention reducing the number of specific responses. The effect of positive versus negative commentaries is unknown, but here we can also hypothesize that the most informative responses, that is, negative ones, will be reduced under high cognitive load. Our analyses of feedback timing and novelty are more exploratory as no such variables have been studied experimentally before. Overall, listeners are expected to be less efficient conversational partners under divided attention, whether this affects the frequency of their input, their nature (specific–generic, negative–positive), or their timing (or all of the above). Our measure of novelty through the rate of other-repetitions, in turn, aims at testing

whether listeners resort to strategies of production economy by reusing the speaker's words: we can expect this strategy to be more useful under high cognitive load, where fewer cognitive resources are available to produce novel feedback, which would result in a larger number of other-repetitions in the high load condition.

In addition to this in-depth analysis of the effect of cognitive load on listener feedback, this study will provide a detailed portrait of commentaries on alignment, exploring the interaction between the specific–generic and positive–negative distinctions brought forth in the literature, as well as the specific devices that are used in each category of commentaries, with a focus on other-repetitions as important signals of both negative (Dingemanse et al., 2015) and positive (Pickering & Garrod, 2021; Pöldvere et al., 2021) feedback. In doing so, we will complement previous studies that focus either on the positive versus negative value (Diederiksen et al., 2022; Healey et al., 2018) or on the generic versus specific distinction (Bavelas et al., 2000; Knutsen et al., 2018).

2. Dialogue experiment: data and method

2.1. Participants

We recruited 18 pairs of participants through word of mouth (23 females and 13 males). Participants were native French speakers between 20 and 55 years old (average 36yo) from France (mainly Paris and Aquitaine regions) and Belgium (mainly Namur Province), who did not know each other and had no experience in linguistics or psychology. They gave informed consent through the self-recruitment questionnaire.

2.2. Procedure

The dialogue experiment was conducted online during a recorded Zoom call: cameras were always off; participants interacted orally, without written chat. Participants received individual instructions in separate break-out rooms. Once ready, they opened a URL that directed them to the experiment. The task consisted of four rounds of 12 tangrams to identify: the director saw one tangram at a time and described it so that the matcher could recognize it among a set of 16 tangrams. The tangrams were selected arbitrarily from the array of 20 used by Branigan et al. (2011). There was no restriction on the duration and content of the descriptions. Once the matcher believed that they had identified the correct tangram, they had to type the number of the tangram on their keyboard and move on to the next one. After the first 12 tangrams, participants had a 10-second silent break and then started over for three more rounds with the same roles and the same tangrams in a new random order.

Participants were divided into three experimental groups. In the *baseline* condition, matchers performed only the main task of identifying tangrams. In the *counting* condition, in addition to the main task, matchers had to mentally count the number of words produced by the director that started with the letter 'd' and type the number next to the tangram number at the end of each trial. The choice of the letter 'd' (as opposed to 't' in Bavelas et al., 2000) was based on its greater frequency in French (e.g. in prepositions, articles, the adjective *droite* 'right'). This condition corresponds to the manipulation of divided attention used in Bavelas et al.'s (2000) Experiment

2. Finally, in the *memory* condition, matchers had to memorize five simple words (e.g. for Round 1: *bille* ‘marble’, *fleur* ‘flower’, *louve* ‘she-wolf’, *hiver* ‘winter’, *ongle* ‘nail’) before each round of 12 tangrams. At the end of a round, they had to type the words from memory. Then, a new set of five words appeared during the silent break before the start of the next round of tangrams.

Once the game was completed, participants were debriefed. None of them guessed the purpose of the experiment.

2.3. Transcription and coding

The experiment was programmed in PsychoPy (version 2020.2.6; Peirce et al., 2019). The answers and game data (tangram numbers, answers to mental tasks, trial time) were recorded on a server. The recordings were transcribed and audio-aligned in EXMARaLDA (Schmidt & Wörner, 2012). Directors’ and matchers’ speech was first segmented into ‘turns’, defined as stretches of talk between speaker changes. Pause duration before matchers’ turns was automatically calculated by the transcription software, while pauses within matchers’ turns were manually included in the transcription. The transcriptions are highly accurate and reproduce hesitations (*mmh*, *eum*), laughter, and other sounds (*mhm*, *mh*). No other information was recorded in the transcriptions, such as intonational patterns or lengthening, as this requires extensive manual coding or specific tools, and detailed prosodic analysis was outside the scope of the study. When prosody was used during the annotation (see below), it was thus strictly based on the annotator’s perception.

Once the 18 recordings were transcribed, matchers’ turns were annotated. We first distinguished turns that were commentaries on alignment (most turns; everything that is directly related to the game) from turns that were talking about the logistics of the game (e.g. typing the answer, technical issues), as in Example (1), which did not count as a commentary.

- (1) <D1> je passe à la suivante (*I’m moving on to the next one*)
<M1> (1000) ouais (527)¹ donc là on est bien sur quatre sur douze? (*yeah so now we’re on four out of twelve?*)

The commentaries were then classified for their type (generic vs specific) and for their value (positive vs negative). A single matcher turn can be divided into several commentaries if it includes commentaries of different types and/or values.

- (2) <D1> non c’est un carré qui correspond à la taille euh de l’image précédente tu vois (*no it’s a square that corresponds to the size uh of the previous image you know*)
<M1> (700) d’accord (522) donc euh c’est comme si tu dis que le carré c’est sa tête quoi (*okay so it’s as if you’re saying that the square is his head*)
- (3) <D3> avec un triangle qui sort du carré vers la droite (*with a triangle that pops out of the square towards the right*)
<M3> euh attends donc il y a un carré sur le dessus? (*uh wait so there is a square on the top?*)

¹In the examples, the figures in brackets indicate pause duration in milliseconds.

Table 1. Decision process for commentary annotation

Generic versus specific annotation	Positive versus negative annotation
Is the commentary lexically specific to what was said before (lexical repetition or same field)? If yes, specific If not, can you move it to a different place in the conversation without change in meaning? Could it be used multiple times? If not, specific Otherwise, generic, including conventions established by participants to move on to the next trial (e.g. <i>je valide</i> 'I confirm (my answer)')	Does the commentary signal communicative success (end of trial)? If yes, positive If not, does it acknowledge an intermediate step in the description ('got it, move on')? If yes, positive Otherwise, negative

In Example (2), the matcher's turn starts with a generic positive commentary 'd'accord', while the rest of the turn is specific positive. Similarly, in Example (3), the turn starts with generic negative signals ('euh attends'), while the rest is specific negative. Our definition thus seems more inclusive than, for instance, what counts as specific versus generic responses in Bavelas et al. (2000). However, we would argue that this difference in definition is mainly due to the difference between the story-telling and tangram description tasks, rather than a conceptual difference in the way we define listener feedback: our matchers are more actively involved in negotiating the description of the tangrams and thus produce more elaborate confirmation or clarification requests. Still, our extended view of specific commentaries, in particular, is actually highly similar to Goodwin's (1986) 'assessments', defined as "an analysis of the particulars of what is being talked about" (1986: 210) and to Bavelas et al.'s (2000) 'specific responses', which "permit listeners to become, for the moment, co-narrators who illustrate or add to the story" (2000: 944). The major difference between our approach and these previous studies is that we include negative specific commentaries, which either do not occur or have not been analyzed in story-telling data.

This annotation was performed independently by two of the authors, after piloting the coding scheme on one transcription. Table 1 reproduces our decision process for these two categories.

Acknowledgement signals (*ok* 'okay', *d'accord* 'all right', *ouais* 'yeah', etc.) were considered positive based on their lexical meaning of alignment if the annotators disagreed or hesitated; restricted offers (e.g. *so it looks like someone dancing*) were considered negative if they presented a rising intonation contour (asking a question) but positive with a falling contour (summarizing, stating a fact); other-repetitions (the matcher directly repeating words produced by the director) were also disambiguated thanks to their intonation (negative if low voice, slow pace, interrogative or uncertain; positive if louder, faster, assertive). In general, perceptual prosodic information was used in the decision process through playing the audio context as well as reading the transcript, for a better interpretation of the context.

The inter-annotation agreement was first analyzed on the first file in order to adjust the annotation scheme. In the rest of the recordings, agreement reached 97% and an almost perfect kappa score of $\kappa = 0.95$ for the generic versus specific variable and a 95% agreement rate and $\kappa = 0.85$ for positive versus negative. All disagreements were resolved through discussion.

Finally, we performed an additional analysis of other-repetitions, that is, whether the commentary consisted of or included quasi-exact repetition of the director's words. Criteria for the identification as a repetition were the following:

- it repeats at least one verb or noun (function words alone do not count);
- the repeated fragment is not syntactically nor pragmatically integrated into the rest of the turn; the turn can include other elements, but the repetition forms a separate utterance within the turn, whose main role is to repeat, not to add precision or reformulate;
- slight modifications of the words were allowed (e.g. change of verb tense, subject agreement).

This approach to other-repetitions is similar to that in Bertrand et al. (2013) and Perrin et al. (2003), who combine formal criteria with an “ostensive character”, that is, an “intention of quotation” (2013: 14). In other words, in these approaches and the one in the present study, sequences are identified as repetitions if their main intention is to repeat or quote the other speaker. In Example (4), the matcher repeats part of the director’s turn but adds a precision (‘à droite’) in order to ask a question. The intention is not to comment on alignment through repetition but to add more information, and this example was therefore not considered a repetition.

- (4) <D3> alors là on dirait quelqu’un qui est adossé contre un mur avec euh les genoux euh vers son torse et la tête un peu penchée (*so here it looks like someone who has his back against a wall with uh the knees uh towards his chest and the head a little tilted*)
 <M3> (1000) euh il est adossé vers le mur à à droite? (*uh he has his back against the wall on the on the right*).

The identification of repetitions was performed by two of the authors, and we reached an agreement rate of 94.4%, leaving only 88 cases of disagreement, which were individually resolved through discussion.

Repetitions are by definition specific commentaries, but they can be either negative (5) or positive (6).

- (5) <D3> ça c’est celui qui montre le triangle au bout du doigt (*that one is the one who shows the triangle at the end of its finger*)
 <M3> (1600) le triangle au bout du doigt euh... (*the triangle at the end of its finger uh*)
- (6) <D3> ... comme un poisson avec un pied (*like a fish with a foot*)
 <M3> (1400s) oula (347) euh ah (307) ah oui un poisson avec un pied oui si si si j’ai euh donc il y a il y a un triangle sur la droite (*wow uh ah ah yes a fish with a foot yes yes yes yes I got it uh so there is there is a triangle on the right*).

The matcher’s turn in Example (6) contains five different commentaries:

- the first interjections ‘oula euh’, generic negative
- the following interjections with an acknowledgement signal ‘ah ah oui’, generic positive
- the other-repetition ‘un poisson avec un pied’, specific positive
- agreement signals ‘oui si si si j’ai’, generic positive
- a specific positive commentary where the matcher summarizes the tangram in his own words (‘euh donc il y a ...’).

The positive or negative value of the repetition was attributed during the first annotation phase and corresponds to the value attributed to the commentary as a whole.

All annotations were then extracted from the concordancer module of EXMAR-aLDA and then processed for export in R.

2.4. Data analysis

The data were analyzed using linear and logistic mixed-effects models in R on our numerical and categorical (binary) variables, respectively, with {lme4} and {lmerTest} packages (Bates et al., 2015; Kuznetsova et al., 2017). In line with Barr et al.'s (2013) recommendation, we always included participants and picture items (i.e. the specific tangram being discussed) as random effects (random intercepts) whenever possible – this information was not available for all analyses because of the annotation extraction procedure. We never included random slopes since participants were involved in only one condition (between-participants design) and there is no reason why some tangram pictures would be more difficult in one condition or the other. No model failed to converge so we always report the most exhaustive model. For dialogue-level variables (e.g. total duration of the recording) where no clustering by participant is possible (one observation per participant), we used the non-parametric Kruskal–Wallis test, since our sample size is too small to meet the assumptions of analysis of variance.

All data files and the R script are available at https://osf.io/my4u9/?view_only=96f483935c25464c80a7a32dd0b445a0. The study was approved by the PPLS Research Ethics Committee of the University of Edinburgh on 22 February 2021 (reference number 215–2021/3).

3. Results

3.1. The effect of cognitive load on performance

The participants completed the tangram description game in 29 minutes on average (16'–53'). The duration of each round decreased steadily from Round 1 to Round 4 for all 18 recordings, as shown in Fig. 1.

We start by exploring the effect of divided attention on durations over an entire dialogue, following the hypothesis that participants take longer to complete the game if they are distracted by a secondary task. We first test for a difference across conditions in total duration. While Fig. 2 suggests that total duration is shorter in the memory condition, the Kruskal–Wallis test is not significant ($\chi^2 = 2.95$, $df = 2$, $p = .22$). We then ran a mixed-effects model on the duration of each trial over an entire dialogue, with participant and picture item as the random effects (intercept only): there is no significant difference between the baseline (37.94 s) and counting (40.86 s; $\beta = 2.92$, $SE = 6.02$, $t = .48$, $p = .63$) conditions or between the baseline and memory conditions (30.05 s; $\beta = -7.89$, $SE = 6.02$, $t = -1.31$, $p = .21$).

We then focus on durations within Round 1, which should be the most demanding round of the game since all tangrams are new to the director, and according to Brennan and Clark (1996), no 'conceptual pact' has already been established between participants. The mean duration is significantly shorter in the memory condition (Fig. 3) according to a Kruskal–Wallis test on the total Round 1 duration ($\chi^2 = 5.98$,

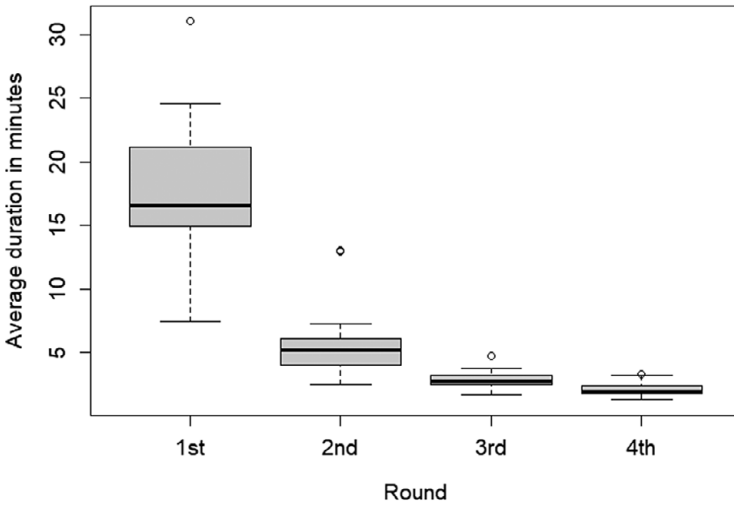


Figure 1. Average round duration.

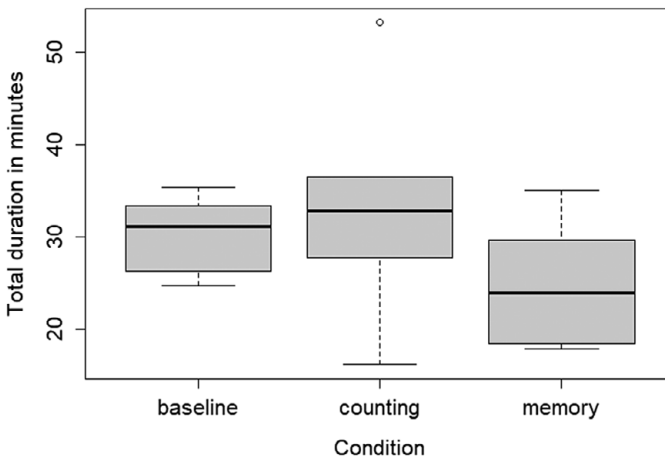


Figure 2. Total duration across conditions (in minutes).

$df = 2, p = .05$). There was a significant effect on trial duration within Round 1: a mixed-effects model with participant and tangram item as the random effects (intercept only) showed that trial duration is significantly shorter in the memory condition (67.84 s; $\beta = -32.5, SE = 14.67, t = -2.21, p < .05$) than in the baseline condition (100.34 s), as can be seen in Fig. 4.

We now turn to a second measure of performance, namely the number of errors that matchers made (wrongly identified tangrams). Although the mean number of errors is smaller in the baseline condition (2.67) than in the conditions under high cognitive load (4.33 and 4), the mixed-effects model, with participant and picture item as random effects (intercept only), was not significant ($\beta = 0.57, SE = .64, z = .89, p = .37$). Therefore, we can conclude that cognitive load only has an effect on task

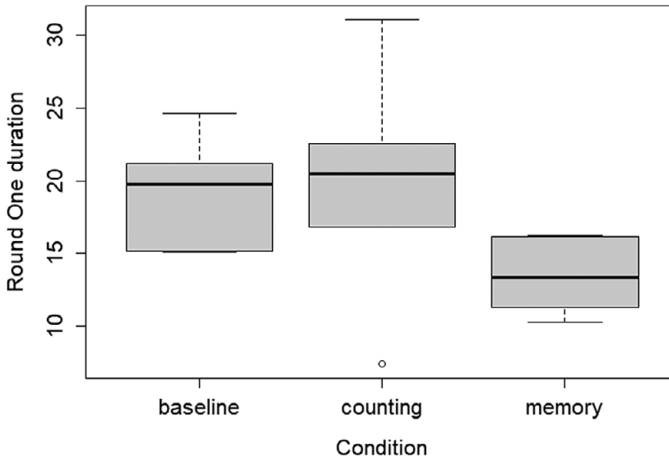


Figure 3. Duration of Round 1 across conditions (in minutes).

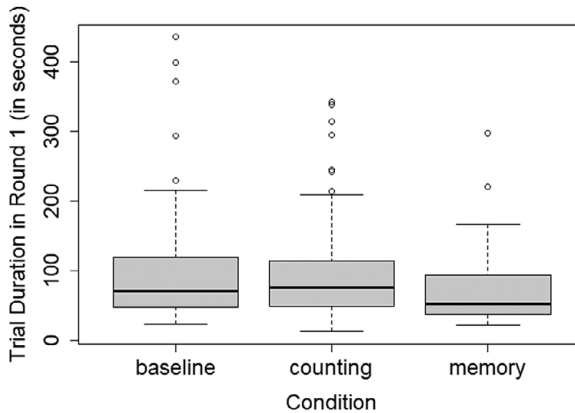


Figure 4. Trial duration in Round 1 (in seconds).

performance measured by duration within the first round, as the other measures (number of errors, durations over the entire dialogue) did not reach significance.

3.2. The effect of cognitive load on feedback quantity

We now turn to the effect of cognitive load on the frequency of commentaries. Overall, no effect reached statistical significance across conditions on the total number of commentaries or the total duration of commentaries, as can be seen in Figs. 5 and 6. With 220, 234, and 232 commentaries on average in the baseline, counting, and memory conditions, respectively, the Kruskal–Wallis test was not significant ($\chi^2 = .18$, $df = 2$, $p = .91$). Neither was the Kruskal–Wallis test on total duration of commentaries, with on average 6 min 6 s in baseline, 8 min in counting, and 5 min 58 s in memory ($\chi^2 = 2.26$, $df = 2$, $p = .32$) conditions.

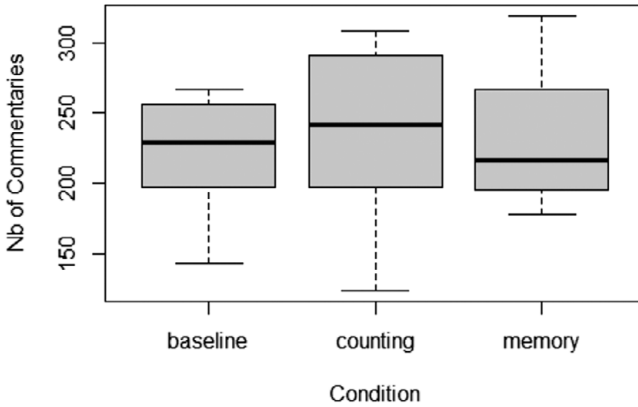


Figure 5. Average total number of commentaries per condition.

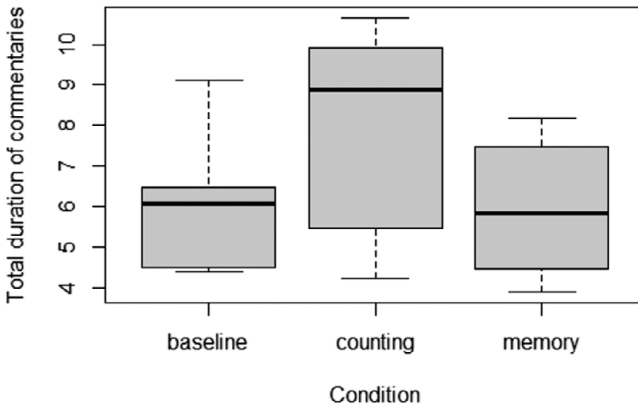


Figure 6. Average total duration of commentaries per condition in minutes.

We also tested whether commentaries were distributed differently with or without cognitive load, following the hypothesis that less efficient participants under high cognitive load would produce fewer specific and fewer negative commentaries, since these are the most informative types of commentaries. We first tested this on the binary variables of commentary type (generic versus specific) and commentary value (positive versus negative). The mixed-effects regression models, with condition as the fixed effect and participant as the random effect (intercept only), did not return a significant effect of condition on type ($\beta = -.10$, $SE = .24$, $z = -0.44$, $p = .66$) or value ($\beta = -0.05$, $SE = .26$, $z = -0.19$, $p = .85$). We then ran a Kruskal–Wallis test for all four combinations, but none of them reached statistical significance: generic positive (Fig. 7; $\chi^2 = .15$, $df = 2$, $p = .93$), generic negative (Fig. 8; $\chi^2 = 4.28$, $df = 2$, $p = .12$), specific positive (Fig. 9; $\chi^2 = .15$, $df = 2$, $p = .93$), and specific negative (Fig. 10; $\chi^2 = .08$, $df = 2$, $p = .96$).

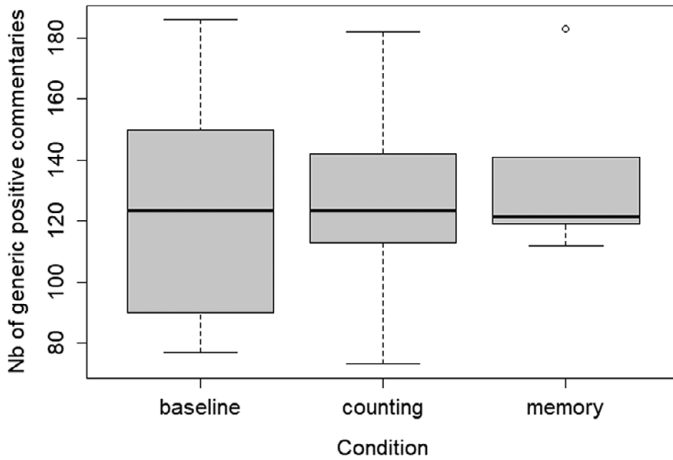


Figure 7. Mean number of generic positive commentaries across conditions.

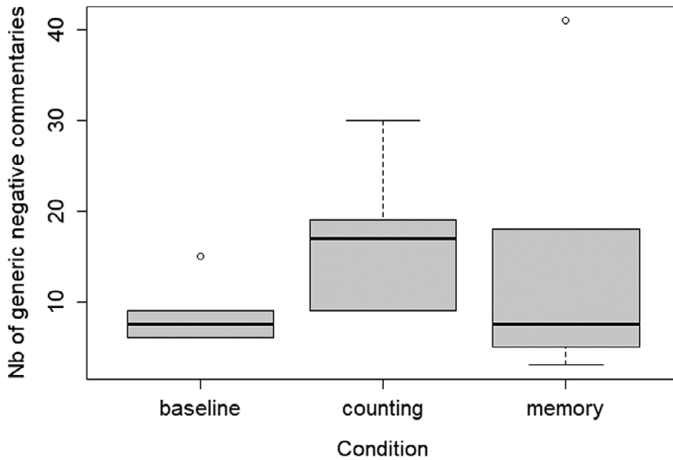


Figure 8. Mean number of generic negative commentaries across conditions.

3.3. Pressure to speak under speech-related versus unrelated tasks

As Clark and Wilkes-Gibbs (1986) pointed out, it is not only the frequency of feedback but also its timing that matters in order to be maximally efficient in dialogue. We thus tested the effect of divided attention on the timing of commentaries, taking as input data the duration of the interval between the end of a director's turn and the start of a matcher's turn. This information thus tells us how long the matcher waits on average before they start producing feedback. This measure is more reliable than total pause duration, which varies strongly with total speech duration and speech rate. While predictions are only tentative here, we can expect that the matcher's reactivity will be impacted by their divided attention, with potentially longer pauses under high cognitive load.

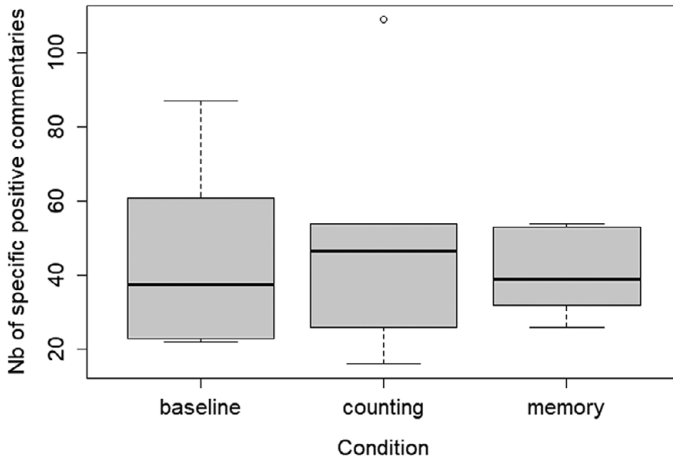


Figure 9. Mean number of specific positive commentaries across conditions.

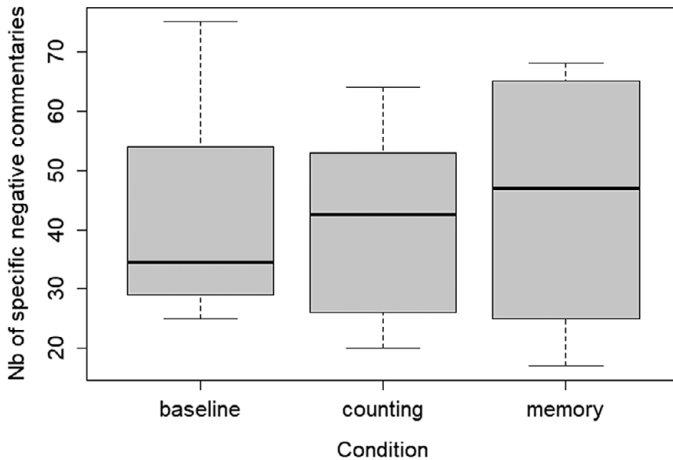


Figure 10. Mean number of specific negative commentaries across conditions.

The duration of pauses before a matcher's turn was significantly different across conditions: a mixed-effects model, with participant as the random effect (intercept only), showed that between-turn pauses were significantly shorter in the baseline condition (1,005 ms) than in the counting condition (1,131 ms; $\beta = .13$, $SE = .04$, $t = 2.87$, $p < .01$); they were longer in the baseline condition than in the memory condition (769 ms; $\beta = -.24$, $SE = .04$, $t = -5.27$, $p < .001$). This can be seen in Fig. 11.

3.4. Cognitive load leads to more economical feedback

We further tested the effect of cognitive load on another feature of feedback, namely the novelty of commentaries. We operationalized novelty by looking at the use of

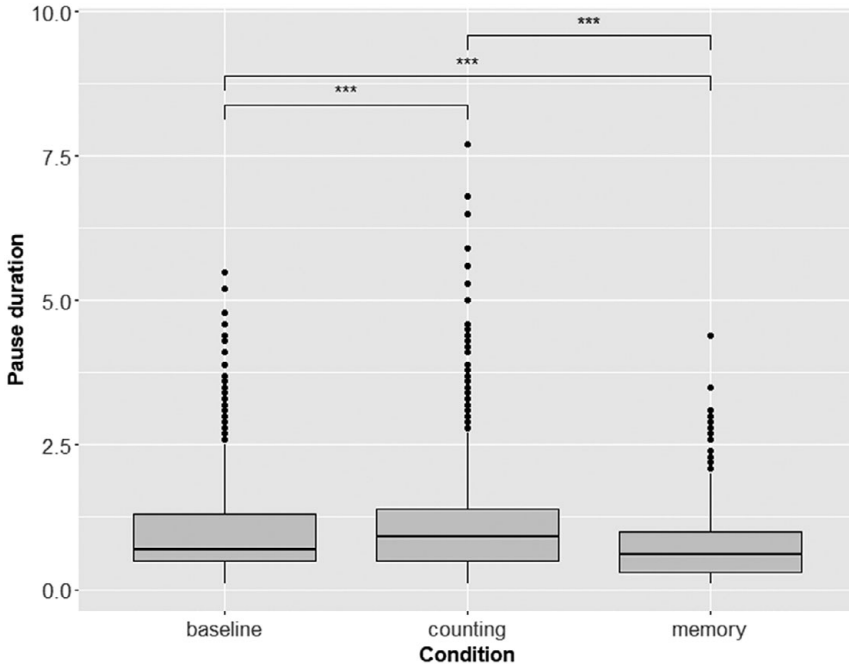


Figure 11. Average duration of pauses (in seconds) before a matcher's turn across conditions.

other-repetitions in commentaries, which occur only in specific responses by definition. Other-repetitions are not novel commentaries but are more economical in the sense that they recycle or reuse previously uttered words. By contrast, commentaries where the matcher reformulates the director's description in their own words, whether to express alignment or misalignment, are more novel.

A recent study by Pöldvere et al. (2021) showed that repetition ('resonance' in their words) is much more frequent in the expression of disagreement (negative stance) than that of agreement. However, their data include 'formal' resonance (i.e. repetitions) and 'semantic' resonance (i.e. reformulations) alike and focus on stance rather than on alignment in a broader sense. In our data, by contrast, repetitions take up a similar rate of specific positive and specific negative commentaries (15.57% and 15.95%, respectively). It is the intonation and surrounding context that determine whether they are positive or negative. In Example (7), for instance, the matcher (partly) repeats the director's turn twice, first with a rising intonation (negative value) and then with a falling intonation (positive value), as also confirmed by the markers surrounding this turn (the connective *alors*, which often occurs before a search; the agreement signals *oui* and *c'est bon*).

- (7) <D13> euh les deux triangles qui vont vers l'arrière (*uh the two triangles that go towards the back*)
 <M13> (1000) alors les deux triangles qui vont oui les deux triangles qui vont vers l'arrière donc euh pour moi c'est bon (*well the two triangles that go yes the two triangles that go towards the back so uh for me it's okay*).

By contrast, in the remaining 85% of commentaries, matchers produce novel feedback that does not repeat the director's turn, thus introducing new information (8) or reformulating the director's description through their own perspective (9):

- (8) <D3> et voilà je sais pas comment t'expliquer autrement <laughter> (*and that's it I don't know how to explain it to you differently*)
 <M3> euh est-ce qu'il y a un (380) un losange euh (247) dans (260) ou un carré couché dans? (*uh is there a a diamond uh in or a square laid down in?*)
- (9) <D3> comme une tu vois une flèche si tu veux en haut à gauche noire et en bas [...] (*like a you know an arrow if you will at the top left black and at the bottom [...]*)
 <M3> [...] ouais il y a vraiment une flèche au-dessus (*yeah there is really an arrow above*).

To measure the effect of cognitive load on this feature of feedback, we ran a mixed-effects logistic regression on specific commentaries, testing whether the presence of other-repetitions differed across experimental conditions (fixed effect) and across participants (random effect, intercept only). Results show that, compared to the baseline condition, participants in the memory condition were more likely to produce repetitions ($\beta = 1.01$, $SE = 0.39$, $z = 2.53$, $p = .011$). The difference between the baseline and counting conditions was not significant ($\beta = 0.71$, $SE = 0.39$, $z = 1.77$, $p = .08$). The average number of repetitions per condition is represented in Fig. 12.

3.5. The many forms of collaborative feedback

Leaving the effect of cognitive load aside, one of our goals is to provide an exploratory description of how these two dimensions of feedback interact, how they are distributed, and what types of linguistic devices they correspond to.

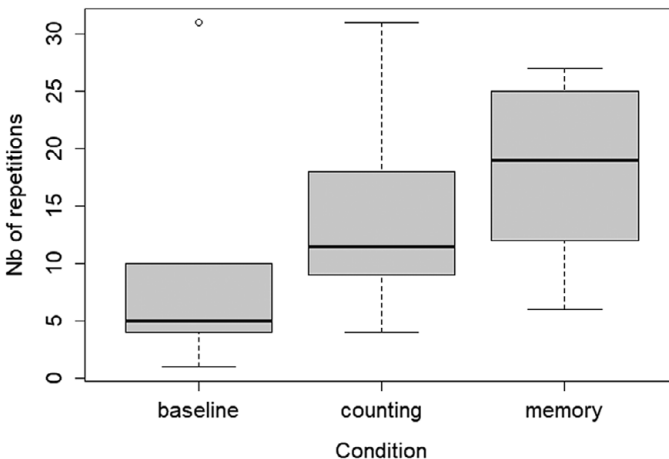


Figure 12. Average number of other-repetitions across conditions.

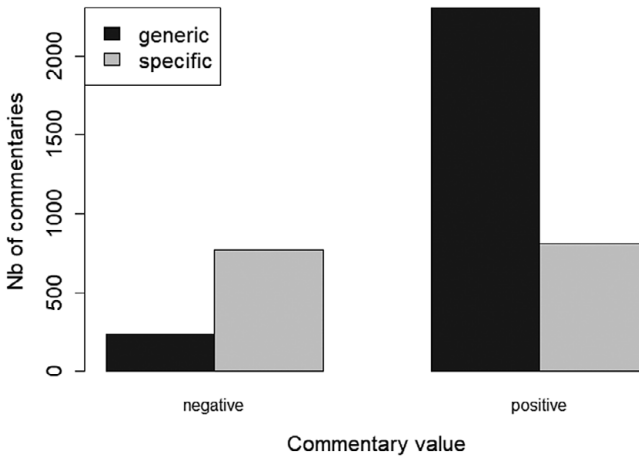


Figure 13. Distribution of response type and response value.

3.5.1. Mapping feedback type with feedback value

Fig. 13 shows the distribution of negative vs positive commentaries in relation to the generic vs specific distinction in our study.

This association between response type and response value is significantly different, according to a mixed-effects generalized linear model with type as the dependent variable, value as the fixed effect, and speaker as the random effect (intercept only) ($\beta = -2.29$, $SE = 0.09$, $z = -25.92$, $p < .001$). In particular, negative commentaries are mostly specific, while positive commentaries are mainly generic.

We can also see in this figure that there is a fairly similar number of positive and negative specific commentaries, in line with the similar number of positive and negative repetitions (which are themselves a subset of specific commentaries). In other words, specific feedback equally relates to alignment and misalignment, while generic feedback is a clearer signal of alignment.

3.5.2. Length of different types of commentaries

These findings do not mean, however, that all specific commentaries are alike. Looking at the mean length of commentaries, we ran a mixed-effects linear regression model with the interaction of commentary type and value as the fixed effect and participants as the random effect (intercept only) and found that all four combinations of commentary types and values significantly differ. In particular, Fig. 14 shows that specific positive commentaries are on average two words shorter than specific negative ones. This difference is significant according to a post-hoc pairwise comparison applying the Bonferroni correction ($\beta = -2.65$, $SE = 0.34$, $t = -7.79$, $p < .001$). In contrast, the difference between positive and negative generic commentaries is not significant ($\beta = .58$, $SE = 0.46$, $t = 1.26$, $p = .58$).

3.5.3. Typical forms of each commentary type: qualitative analysis

Table 2 includes translations of typical examples of the different types of commentaries.

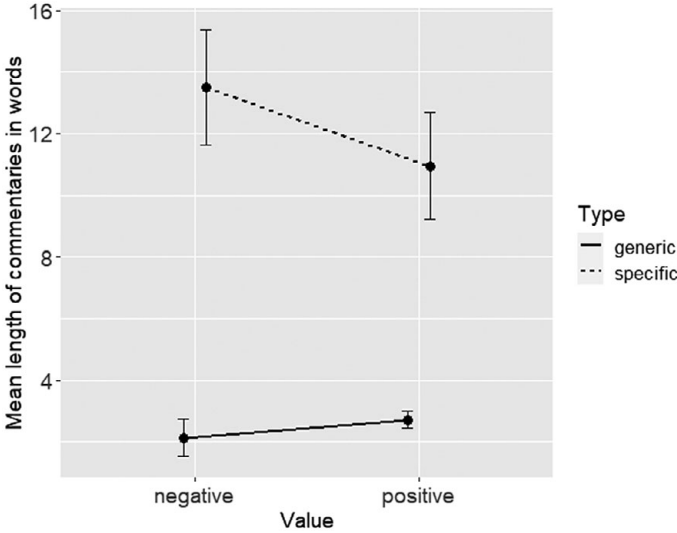


Figure 14. Mean length of commentaries across type and value, in number of words.

Table 2. Example commentaries per type and value (translated from French)

	Positive	Negative
Generic	<i>ouais</i> 'yeah' <i>ok je pense l'avoir</i> 'okay I think I got it' <i>ah oui</i> 'ah yes'a	<i>euh</i> 'uh' <i>non attends</i> 'no wait' <i>ah merde</i> 'ah damn' <i>comment?</i> 'what?'
Specific	<i>il y a effectivement un petit oeil là</i> 'there is indeed a small eye there' <i>donc un triangle avec un angle droit</i> 'so a triangle with a right angle' <i>le premier qu'on avait eu</i> 'the first one we had'	<i>est-ce que euh c'est une grosse forme sous la tête en losange?</i> 'and is it uh so a big diamond shape under the head?' <i>et on dirait comme s'il avait un pied euh il y a une pointe vers l'extérieur gauche?</i> 'and it looks like as if he had a foot uh is there a spike towards the outside left?'

As can be seen in this table, generic positive commentaries mainly consist of agreement markers (variations and combinations of *okay*, *yes*, *yeah*, *alright*, *got it*, etc.) and generic negative commentaries are primarily filled pauses (*uh*, *uhm*, *mmh*) as well as some other interjections (*no*, *wait*, *damn*).

Specific commentaries are much longer and more varied. There is no striking formal difference between positive and negative in this category, especially since both types can include other-repetitions. They are typically descriptive sentences about the tangrams, providing a reformulation of the director's description, some added detail, or a request for confirmation after a reformulation or after a new detail. They can also consist of direct metalinguistic comments indicating recognition ('that's the first one we did') and affiliation ('great description'), or alternatively confusion ('I'm completely lost', 'can you start over please?').

Positive and negative specific commentaries differ from each other by their intonation, although systematic prosodic analysis would be required to confirm this. They also differ by the typical markers that frame them: *alors* ‘so/well’ often launches a negative response, indicating a search; utterance-final *hein* ‘isn’t it?’ is almost exclusive to confirmation requests (negative); and *quoi* (‘you know’) often closes a positive response (typically a summary or reformulation said with confidence):

- (10) <M1> c’est comme si c’était un profil gauche comme tout à l’heure et comme si euh (393) la personne est (247) assise quoi (*it’s as if it was a left profile like before and as if uh the person is sitting ‘quoi’*)

We further observed that some generic commentaries consist of pragmatic markers such as *d’accord* ‘alright’ or *ok* ‘okay’ on their own. These pragmatic markers can also be used to frame (i.e. precede or follow) specific commentaries, in a single matcher’s turn. These framing markers are in bold in the following examples:

- (11) <M1> **d’accord** donc euh c’est comme si tu dis que le carré c’est sa tête quoi (*alright so uh it’s as if you say that the square is his head*)
 (12) <M1> donc un triangle qui est accroché dans le vide de sa tête **ok** (*so a triangle that is hanging in the back of his head okay*)

Generic and specific commentaries are therefore not mutually exclusive but often appear together.

It is also interesting to note that, during the course of the dialogue, interactants develop their own individual conventions to signal final alignment at the end of a trial: the matcher may say the number of the tangram on their screen, or say *validé* (they have entered and ‘validated’ their answer in the programme) or *c’est bon* (‘it’s all good’), thereby signalling to the director that they can move on to the next tangram.

3.6. The evolution of feedback over time

Finally, we can complement this comprehensive portrait of commentaries on alignment by looking at how the distribution and nature of feedback might change across the four rounds of the game. Our data show that specific commentaries become much less frequent in Round 4, as can be seen in Fig. 15. The overall distribution of commentary types is significantly different across rounds, according to a mixed-effects generalized linear model with type as the dependent variable, rounds as the fixed effect, and speaker as the random effect (intercept only): every round is significantly different from the first one in terms of the rate of generic versus specific commentaries (Round 2: $\beta = -.23$, $SE = 0.08$, $z = -2.77$, $p < .01$; Round 3: $\beta = -.39$, $SE = 0.10$, $z = -3.73$, $p < .001$; Round 4: $\beta = -.84$, $SE = 0.13$, $z = 6.41$, $p < .001$).

As might be expected, negative commentaries steadily decrease over time as the dialogue goes on and more and more alignment is achieved between the participants (see Fig. 16): most alignment issues appear to be solved in Round 1. Again, the distribution of commentary values significantly differs across rounds (Round 2: $\beta = .55$, $SE = 0.09$, $z = 5.68$, $p < .001$; Round 3: $\beta = 1.02$, $SE = 0.14$, $z = 7.45$, $p < .001$; Round 4: $\beta = 1.67$, $SE = 0.2$, $z = 8.34$, $p < .001$).

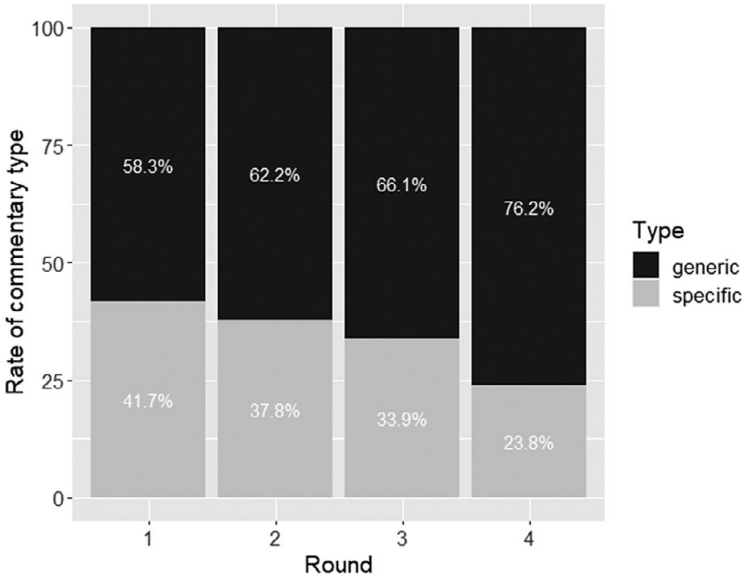


Figure 15. Generic and specific commentaries across rounds.

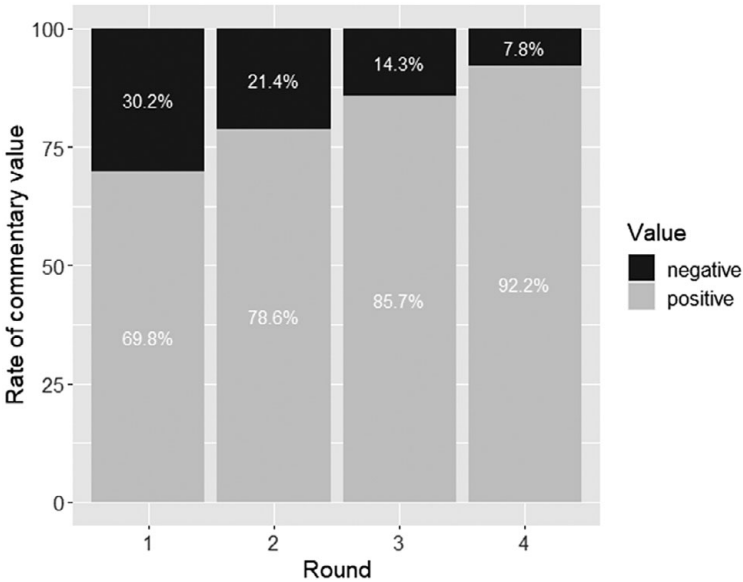


Figure 16. Positive and negative commentaries across rounds.

Commentary length is also greatly affected by time and decreases from an average of 7.4 words in Round 1 to 2.83 words in Round 4, a difference that is highly significant according to a linear mixed-effects regression analysis with round as the fixed effect and participant as the random effect (intercept only) ($\beta = 2.99$, $SE = 0.51$, $t = 5.83$, $p < .001$).

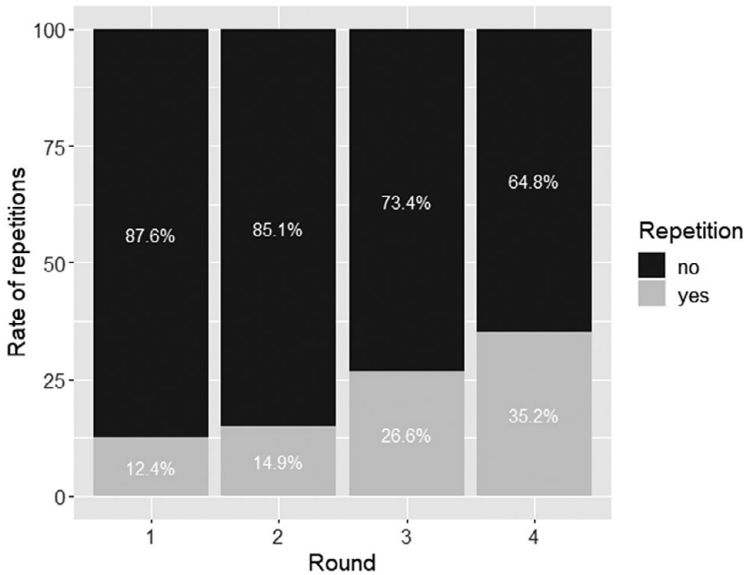


Figure 17. Proportions of other-repetitions in specific commentaries across rounds.

While specific commentaries generally decrease over time (Round 1: 991; Round 2: 328; Round 3: 173; Round 4: 88), the rate of repetitions within specific commentaries increases. Round 3 ($\beta = .94$, $SE = .21$, $z = 4.6$, $p < .001$) and Round 4 ($\beta = 1.38$, $SE = .25$, $z = 5.44$, $p < .001$) in particular show a significant boost of this alignment device compared to Round 1, as can be seen in Fig. 17.

A closer inspection of the data reveals that, in Round 1, repetitions often signal misalignment: 65.04% of repetitions are negative commentaries in the first round. This proportion progressively decreases and reaches only 22.58% in Round 4 (a statistically significant difference from Round 1 according to a z-test of independence: $z = 4.26$, $p < .001$): repetitions then are more positive, expressing alignment and recognition.

4. General discussion

This study used a tangram description game between matchers and directors to investigate the impact of divided attention on the production of feedback. The data resemble previous studies using this paradigm, where participants take less and less time to perform the task across rounds, presumably as a consequence of descriptions becoming shorter through negotiation and alignment (Clark & Wilkes-Gibbs, 1986; Hupet & Chantraine, 1992). We will now review and discuss our main findings.

4.1. Effect of divided attention on the timing and novelty of feedback

We started from the hypothesis that divided attention, in the form of a secondary mental task, would affect the production of listener feedback in task-oriented dialogue. Our data showed that enhanced cognitive load indeed had an effect on

two features of listener feedback: its timing, through measures of pause durations before matchers' turns, and its novelty, through the rate of other-repetitions in specific commentaries.

Looking at timing measures, the longest pauses are produced in the counting condition, which we interpret as a direct effect of divided attention: matchers must pay attention to the directors' words themselves (whether they start with a *d* or not), not only to the content. By contrast, the shortest pauses are produced in the memory condition, which may occur because matchers were rushing through the trials for fear of forgetting the five words to remember. Although the effect of condition on total duration was not significant, total Round 1 duration and trial duration within Round 1 were shorter under the memory task than in the other conditions, which suggests that matchers in the memory condition were indeed attempting to speak as fast as possible, but eventually took as long as the other participants to successfully complete the tangram game. Finally, pause duration in the baseline condition is intermediate and could correspond to the timing of commentaries that occurs without extra cognitive demands: not too fast, not too slow.

Between-turn pauses thus revealed opposite strategies between our two high-load conditions: the language-focused task of counting words that start with a *d* resulted in longer turn intervals, whereas the retention of unrelated words in memory throughout each round of the game might have prompted matchers to speak faster in order to move more quickly to the end of the round. The intermediate duration of pauses in the baseline condition seems to point to a 'sweet spot' of feedback timing, which is uttered not too late and not too soon after the previous director's turn (around one second later). The pressure to speak thus varies with the mental tasks that the matcher has to manage.

Let us also mention that between-turn pauses in our data are longer than the average duration of turn-taking reported in other studies (around 200 ms; e.g. Heldner & Edlund, 2010). This relatively long average duration contrasts with corpus-based studies on conversation, where listeners can start planning while listening (Levinson & Torreira, 2015). This could be due to various factors pointed out by Knudsen et al. (2020): the director's words may be less predictable (e.g. right versus left, square versus triangle) than in regular conversation; speakers cannot see each other as cameras were off, thus preventing prediction of turn ending; the dual task of listening and planning has to compete with the other tasks of the matching game (looking at tangrams, typing their answer). Further research is needed to explore these contrasts.

Turning to novelty, repetitions are least frequent in the baseline condition and most frequent in the memory condition, which confirms our hypothesis that high cognitive load leads to more economy in commentaries on alignment. This can be interpreted as an effect of divided attention, imposing greater cognitive demands on the matchers who then turn to more economical commentaries: matchers in the memory condition may not have a lot of available resources to produce new linguistic material and therefore recycle previously uttered material to express both alignment and misalignment. Repetitions thus appear to function as an economical device for indicating alignment. To conclude that matchers in these experimental conditions are less communicatively efficient than in the baseline condition might be a stretch considering that there were no differences between conditions in terms of performance, but more detailed analyses of the directors' speech (as in Dingemanse et al., 2015, or Tolins et al., 2014) might demonstrate an impact of 'recycled' or economical

feedback on how the dialogue continues, with perhaps more misunderstandings in the interaction because of the potential ambiguity of repetitions (although disambiguating the positive or negative value of feedback was not difficult during the annotation).

Regardless of condition, the function of repetitions changes throughout the interaction, from expressing mainly negative feedback to more and more positive feedback towards the end of the game, in accordance with the increasing level of interactive alignment.

4.2. *No effect of divided attention on performance and feedback quantity*

Contrary to the above findings on timing and novelty, and with the exception of Round 1 durations, the effect of cognitive load on performance and feedback quantity was not significant, although several parameters showed marginal trends (e.g. the duration of commentaries is longer in the counting condition; the total duration of commentaries is shorter in the memory condition). This overall lack of effect contrasts with Bavelas et al.'s (2000) findings on story-telling data, where they observed a different distribution of the various types of feedback as a result of divided attention. While this might be due to our relatively small sample size, we did find a significant effect of condition for other features of feedback, which suggests that our manipulation was indeed effective and our sample was large enough to observe meaningful differences.

Furthermore, while we had expected fewer negative commentaries under high cognitive load because of their higher informativeness (and therefore higher cognitive cost), an alternative hypothesis would be that matchers who experience more difficulties during the task encounter more occasions of misalignment and therefore produce more negative feedback. This account would be consistent with Knutsen et al.'s (2018) findings, where certain types of coordination devices were more frequent in the more difficult experimental condition of the matching game. In other words, there could be opposite forces that prevent frequency effects from reaching statistical significance.

Let us also stress the fact that the speakers in our setting did not see each other. The use of Zoom with cameras off prevents eye-contact and therefore restricts feedback to verbal responses only, thus excluding non-verbal cues such as head nods but also “looking sad, gasping in horror, mirroring the speaker’s gesture” (Bavelas et al., 2000, 943). The difference between our results and those of Bavelas et al. (2000) could therefore also be due to this restriction, in addition to the difference in languages (English versus French), although we see no clear reason why the latter should have much impact on our measures. It may also be that any effect of cognitive load was reduced and trumped by the Zoom setting, as has been shown in question–answer delays (Boland et al., 2022). However, in their corpus study, Colman and Healey (2011: 1566) found that “the level and distribution of repair types is relatively unaffected” by the possibility of eye-contact when comparing face-to-face and no-eye-contact settings. We would also argue that the Zoom setting was familiar to our participants at the time of data collection, more than a year after the onset of the COVID-19 pandemic, thus limiting a possible sense of strangeness from the online tool. Nevertheless, it cannot be ruled out that the absence of visual feedback

partly explains our findings, since divided attention might have affected the non-verbal cues of alignment more than the verbal commentaries.

All in all, the effects of cognitive load remain difficult to assess. Previous studies provided inconsistent findings, from positive feedback in story-telling contexts (Bavelas et al., 2000) to feedback focusing on procedural coordination in matching games using a different manipulation for task demands (Knutsen et al., 2018). The roles of listener and matcher are very different across these studies: the matcher does not merely listen and react to the director but has to actively look for specific tangrams among abstract similar-looking shapes, all the while making suggestions and proposing alternative descriptions. Therefore, it could be that tangram matching is so highly demanding already that divided attention does not add further difficulty. This would be in line with Bavelas et al.'s (2000, 947) conclusion that listeners are “capable of handling several simultaneous demands and responses” as long as they can attend to the content of the interaction. To sum up so far, the answer to our question on the effect of divided attention on feedback is more nuanced than we expected, and our experimental manipulation affected only some features of feedback such as timing and novelty.

4.3. *Forms and functions of specific commentaries*

This study is, to our knowledge, the first that offers quantitative data over both generic and specific commentaries and both positive and negative commentaries on alignment. It allowed us to refine our understanding of the intersection between these dimensions of feedback. In this respect, we showed that matchers give more information (with specific commentaries) when they are not aligned and less information (with generic commentaries) when they are aligned. These findings confirm Clark and Wilkes-Gibbs' (1986) observation that listeners try to be maximally collaborative by selecting the appropriate degree of informativeness for their feedback depending on the local contextual needs (not much information if the dialogue is on track, more information if alignment is not achieved yet).

We also showed that, within specific commentaries, positive feedback is shorter than negative feedback. This is in line with the findings of Dingemanse et al. (2015), who found that, for other-repair initiators, different forms of responses have different lengths and therefore different information contents: ‘open requests’ (our generic negative, 2.14 words in our data) are shortest with 3.7 words, and restricted requests and restricted offers (both specific negative, 13.5 words in our data) are 10.4 and 13 words long on average, respectively. Our data complement their description by showing that ‘restricted’ or specific feedback is shorter in the case of alignment.

In terms of distribution across rounds, specific commentaries become less frequent over time, which contrasts with Bavelas et al. (2000) who found an increase in specific commentaries in English story-telling, presumably because the matcher feels progressively more involved thanks to their knowledge of the story. Similarly, Bertrand and Espesser (2017) found in a corpus of French narratives that listeners speak more often and in longer turns (i.e. with specific responses) over time. Language alone thus cannot explain this difference in our results, which probably lies in the setting and task (passively listening to a story versus collaborating in a game). Overall, we found that time affects commentary length, with a general

decrease in the number of words across rounds, which indicates that most of the negotiation happens in the first round. The following rounds are mainly used to reactivate previously defined labels for tangrams (either because of priming or through the resort to ‘conceptual pacts’; Brennan and Clark (1996)), with occasional descriptions when these labels do not suffice or when no label has been defined yet.

One perhaps surprising conclusion of our study is that specific commentaries (and in particular repetitions within them) are functionally quite polyvalent: they can express alignment and misalignment with a similar frequency and formal features, contrary to generic commentaries, which are clearly different depending on their positive or negative value. During the annotation process, we relied not only on the explicit content of the commentaries but also on intonation in order to determine whether a commentary expresses confidence or doubt. Detailed prosodic analysis of specific commentaries is therefore an obvious avenue of research for this dataset.

This ambivalence of specific feedback is interesting since it begs the question of why do listeners ‘bother’ to provide a lot of informative content when they are expressing alignment. In other words, if everything is clear and the dialogue is on track, why not simply take the economical route and use short, generic responses like *okay* or *mhm*? There are of course uses of long specific commentaries that allow the matchers to make sure that they share the same perspective as the directors, often by summarizing and reformulating the descriptions through their own perspective or a shared one. But there are also commentaries that go beyond this purely task-oriented function: matchers sometimes praise the quality of the description or make joking comments that reflect their mutual understanding. Even summaries and reformulations sometimes feel ‘unnecessary’ or ‘overkill’ for the sole purpose of the game. Therefore, we propose that positive commentaries also perform a social function: they stress alignment, encourage the speaker to continue, and express affiliation and appreciation (cf. Du Bois’s, 2014 intersubjective view of alignment). This social function complements the cognitive account of other-repetitions as an economical resource for alignment that we sketched in the quantitative analysis. We therefore agree with Pöldvere et al., (2021, 665) who concluded that there is a “close interplay between intersubjective motivations and cognitive facilitation” in the use of resonating responses.

In sum, we propose that positive feedback strengthens the relationship between the interactants beyond its utilitarian function, which is why such feedback can be expected to occur in all kinds of interaction contexts. While most studies have looked at feedback in story-telling or in dialogue games, the universality of positive feedback for social purposes would need to be further investigated in other contexts such as general conversation, professional settings, or even in less interactive contexts (e.g. attending a lecture or a stage performance).

In addition to more data types, positive feedback could also be explored as a potential correlate of linguistic convergence. With a larger sample, individual differences could be investigated to test whether matchers who produce a lot of positive commentaries are also more likely to reuse the directors’ words. If that is the case, then commentaries on alignment would be correlated with lexical alignment. Additional levels of alignment (e.g. syntactic, prosodic, pragmatic, para-verbal; Bertrand et al., 2013; Hu & Degand, 2022) might all overlap and cluster in profiles of more or less alignment- and convergence-prone speakers. Such endeavours would bring together different lines of research on alignment both as a mental process and as an observable behaviour.

Data availability statement. All data and code used in this study can be found in open access on the following OSF repository: https://osf.io/my4u9/?view_only=96f483935c25464c80a7a32dd0b445a0.

Competing interest. The authors declare none.

References

- Anderson, A. H., Bader, M., Bard, G., Ellen, B., Elizabeth, D., Gwyneth, G., Simon, I., Stephen, K., Jacqueline, M. A., Jan, M., Jim, S., Catherine, T., Henry, S., & Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34(4), 351–366.
- Baldock, J., Kapadia, S., & van Steenbrugge, W. (2019). The task-evoked pupil response in divided auditory attention tasks. *Journal of American Academy of Audiology*, 30(4), 264–272.
- Bangerter, A., & Herbert, C. H. (2003). Navigating joint projects with dialogue. *Cognitive Science*, 27, 195–225.
- 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., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6), 941–952.
- Bertrand, R., & Espesser, R. (2017). Co-narration in French conversation storytelling: A quantitative insight. *Journal of Pragmatics*, 111, 33–53.
- Bertrand, R., Ferré, G., and Guardiola, M. (2013). French face-to-face interaction: repetition as a multi-modal resource. In N. Campbell, & M. Rojc (Eds.), *Coverbal synchrony in human-machine interaction* (p. 30). Science Publishers.
- Boland, J. E., Fonseca, P., Mermelstein, I., & Williamson, M. (2022). Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General*, 151(6), 1272–1282.
- Branigan, H. P., Catchpole, C. M., & Pickering, M. J. (2011). What makes dialogues easy to understand? *Language and Cognitive Processes*, 26(10), 1667–1686.
- Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6), 1482–1493.
- Clark, H. H. (1996). *Using Language*. Cambridge University Press.
- Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1), 1–39.
- Colman, M., & Healey, P. (2011). The distribution of repair in dialogue. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 33, 1563–1568. <https://escholarship.org/uc/item/7zd514km>
- Dideriksen, C., Christiansen, M. H., Tylén, K., Dingemanse, M., & Fusaroli, R. (2022). Quantifying the interplay of conversation devices in building mutual understanding. *Journal of Experimental Psychology: General*, 152(3), 864–889.
- Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladdottir, R. S., Kendrick, K. H., Levinson, S. V., Manrique, E., Rossi, G., & Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PloS ONE*, 10(9), e0136100. <https://doi.org/10.1371/journal.pone.0136100>
- Du Bois, J. W. (2014). Towards a dialogic syntax. *Cognitive Linguistics*, 25(3), 359–410. <https://doi.org/10.1515/cog-2014-0024>
- Tree, F., & Jean, E. (1999). Listening in on monologues and dialogues. *Discourse Processes*, 27(1), 35–53.
- Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M. H., & Dingemanse, M. (2017). Measures and mechanisms of common ground: backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society (CogSci 2017)* (pp. 2055–2060). Cognitive Science Society.
- Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2), 181–218. [https://doi.org/10.1016/0010-0277\(87\)90018-7](https://doi.org/10.1016/0010-0277(87)90018-7)
- Goodwin, C. (1986). Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies*, 9, 205–217.

- Healey, P., Mills, G. T., Gregory, J., Eshghi, A., & Howes, C. (2018). Running repairs: coordinating meaning in dialogue. *Topics in Cognitive Science*, 10, 367–388.
- Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38, 555–568.
- Hu, J., & Degand, L. (2022). “Alignment of conversational discourse units in English dialogues.” In *Paper presented at the SLE conference, Bucharest, Romania, 24–27 August, 2022*.
- Hupet, M., & Chantraine, Y. (1992). Changes in repeated references: Collaboration or repetition effects? *Journal of Psycholinguistic Research*, 21(6), 485–496.
- Knudsen, B., Creemers, A., & Meyer, A. S. (2020). Forgotten little words: How backchannels and particles may facilitate speech planning in conversation? *Frontiers in Psychology*, 11, 593671. <https://doi.org/10.3389/fpsyg.2020.593671>
- Knutsen, D., Bangarter, A., & Mayor, E. (2018). Procedural coordination in the matching task. *Collabra: Psychology*, 5(1), 3. <https://doi.org/10.1525/collabra.188>
- Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(5), 113–114. <https://doi.org/10.3758/BF03342817>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26.
- Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6, 1–17. <https://doi.org/10.3389/fpsyg.2015.00731>
- Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7–8), 953–978.
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Hochenberger, R., Sogo, H., Kastman, E., & Lindelov, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Perrin, L., Deshaies, D., & Paradis, C. (2003). Pragmatic functions of local diaphonic repetitions in conversation. *Journal of Pragmatics*, 35(12), 1843–1860.
- Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2), 169–190.
- Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.
- Pöldvere, N., Johansson, V., & Paradis, C. (2021). Resonance in dialogue: the interplay between intersubjective motivations and cognitive facilitation. *Language and Cognition*, 13, 643–699.
- Schmidt, T., & Wörner, K. (2012). EXMARaLDA. In J. Durand, G. Ulrike, & G. Kristoffersen (Eds.), *Handbook on corpus phonology* (pp. 402–419). Oxford University Press.
- Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21, 211–232.
- Strayer, D. L., Watson, J. M., & Drews, F. A. (2011). Cognitive distraction while multitasking in the automobile. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 29–58). Elsevier Academic Press.
- Tolins, J., Tree, F., & Jean, E. (2014). Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70, 152–164.

Cite this article: Crible, L., Gandolfi, G., & Pickering, M. J. (2024). Feedback quality and divided attention: exploring commentaries on alignment in task-oriented dialogue, *Language and Cognition*, 1–29. <https://doi.org/10.1017/langcog.2023.65>