Language and Cognition (2024), 1–24
doi:10.1017/langcog.2024.1

ARTICLE

Backchannel behavior is idiosyncratic

Peter Blomsma[1], Julija Vaitonyté[1], Gabriel Skantze[2] and Marc Swerts[1]

[1]Tilburg University, Tilburg, The Netherlands
[2]KTH Royal Institute of Technology, Stockholm, Sweden
Corresponding author: Peter Blomsma; Email: peter.blomsma@gmail.com

(Received 04 November 2022; Revised 31 October 2023; Accepted 04 January 2024)

Abstract

In spoken conversations, speakers and their addressees constantly seek and provide different forms of audiovisual feedback, also known as backchannels, which include nodding, vocalizations and facial expressions. It has previously been shown that addressees backchannel at specific points during an interaction, namely after a speaker provided a cue to elicit feedback from the addressee. However, addressees may differ in the frequency and type of feedback that they provide, and likewise, speakers may vary the type of cues they generate to signal the backchannel opportunity points (BOPs). Research on the extent to which backchanneling is idiosyncratic is scant. In this article, we quantify and analyze the variability in feedback behavior of 14 addressees who all interacted with the same speaker stimulus. We conducted this research by means of a previously developed experimental paradigm that generates spontaneous interactions in a controlled manner. Our results show that (1) backchanneling behavior varies between listeners (some addressees are more active than others) and (2) backchanneling behavior varies between BOPs (some points trigger more responses than others). We discuss the relevance of these results for models of human–human and human–machine interactions.

Keywords: backchannels; consensus sampling; head nod; listener feedback; multimodal; O-Cam paradigm

1. Introduction

A spoken conversation can be operationalized as a highly interactive form of cooperative activity between at least two individuals. In that sense, it is more than an exact data transfer process, whereby a sender simply transmits information to a receiver, who then decodes the incoming message. The latter characterization of a spoken interaction does not do justice to the observation that an addressee is often more than a passive listener and is, in fact, co-responsible for a successful exchange of information (Clark, 1996). Indeed, communication via speech can sometimes be a fuzzy endeavor, for example, because of a noisy channel or the fact that a speaker may not correctly estimate a listener's prior knowledge about a specific state of affairs. As a result, it is typically the case that speakers and addressees seek and provide feedback

on the smoothness of the interaction, to check whether information has successfully arrived at the other end of the communication chain. Accordingly, there is a growing interest in current models of spoken interaction regarding the systematicity of various types of feedback behavior.

In this article, we are specifically interested in the brief responses, called *back-channels* (Yngve, 1970), that addressees return during an interaction. Such back-channels, which can be verbal and non-verbal, serve as cues to show a speaker that an addressee is engaged and listening. Backchannels thus convey attention and interest to the speaker, and they can also regulate turn-taking (Gravano & Hirschberg, 2011). While verbal backchannels include vocalizations (laugh, sigh, etc.), paraverbals ('mm-hmm', 'uh-huh', etc.) and short utterances ('really', 'yeah', 'okay'), non-verbal backchannels consist of facial expressions, nodding, eye gaze and gestures. It has been shown that there is a marked difference between signals that serve as 'go-on' cues, that is, to make clear that the addressee has correctly processed the incoming message, and signals that highlight a possible communication problem so that a speaker–sender may have to repair a potential error (Granström et al., 2002; Krahmer et al., 2002; Shimojima et al., 2002).

In the literature, backchannels are distinguished from turn-taking cues. The intention of a speaker, when backchanneling, is to signal that the current speaker is still in charge of the turn, while the intention of a turn-taking cue is to interrupt the speaker and to take the speaking turn. Thus, backchannels can be viewed as a form of cooperative overlap or, from a turn-taking perspective, as a turn-yielding cue (Bertrand et al., 2007).

1.1. Backchannel-inviting cues

It has been shown that the timing of backchannels is crucial to guarantee a smooth interaction (Gratch et al., 2006; Poppe et al., 2011). For instance, Gratch et al. (2006) demonstrated that a wrongly timed head nod from a listener can disrupt a speaker, which suggests that addressees typically are efficient at producing backchannels at the right points in an interaction. Indeed, research shows that backchannels occur at specific points in a conversation, for example, after the speaker gives a so-called backchannel-inviting cue (Gravano & Hirschberg, 2011), also called backchannel-preceding cues (Levitan et al., 2011).

The specific behaviors that the speaker produces to transmit backchannel-inviting cues to elicit backchannel behavior from an addressee come in different forms, including the usage of specific prosodic patterns. Gravano and Hirschberg (2009) found that speakers use rising and falling intonations to elicit feedback. Similarly, Cathcart et al. (2003) and Ward and Tsukahara (2000) showed that listeners often provide a backchannel after speakers have lowered their pitch for at least 110 ms, and Cathcart et al. (2003) showed that pauses in the speaker's speech and also certain parts of speech are predictive of backchannels. Furthermore, Duncan (1972) observed that backchannels occur after syntactically complete sentences, while Bavelas et al. (2002) revealed that mutual gaze often occurs prior to a backchannel being produced. In line with this, Hjalmarsson and Oertel (2012) found that listeners were more likely to identify a backchannel-inviting cue when the speaker (an embodied conversational agent (ECA) in this case) made direct eye contact with the camera, as opposed to gazing away.

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

The probability that a listener will backchannel after a cue will increase when backchannel-inviting cues are stapled to form more complex signals (Gravano & Hirschberg, 2011). In a similar vein, Hjalmarsson (2011) showed that it appears to be the case with turn-taking and turn-yielding signals (signals closely related to backchannel-inviting cues, yet distinct) that the more cues are used to comprise the signal, the faster the reaction time of the interlocutor becomes. Speakers may not be aware of sending out backchannel-inviting cues, but listeners and observers are capable of picking up on those signals. Bavelas et al. (2000) showed that listeners are even able to provide backchannels at the right moment when not attending to the content of the speech.

1.2. Backchannel opportunity points

Although speakers provide backchannel-inviting cues, it is up to the addressee to pick up on these cues and identify relevant moments in a conversation to produce backchannels. Those moments in a conversation, where it is appropriate for an addressee to provide some kind of listener feedback, are referred to as backchannel opportunity points (BOPs) (Gratch et al., 2006). BOPs, which are also known as jump-in points (Morency et al., 2008) and response opportunities (de Kok, 2013), are points in the interaction where an addressee could or would want to provide feedback in reaction to the speaker (de Kok & Heylen, 2010). Prior studies show that not all BOPs are used by addressees to provide a backchannel (Kawahara et al., 2016; Poppe et al., 2011). However, we lack detailed insight into the extent to which there is variability in the way addressees return feedback and regarding the different types of BOPs.

1.3. Current work

The goal of this study is to shed light on the variation that exists in backchannel behaviors across addressees and within an individual addressee. Specifically, we ask the following: (1) What types of behaviors are utilized by addressees to give feedback during BOPs? (2) How does feedback behavior differ across different addressees? (3) To what extent differs the behavior of addressees for the same BOP?

The fact that we expect there to be variability between and within addressees in their feedback behavior is in line with the previous findings that human beings do not have a fixed communication style. Speakers have been shown to adapt their way of speaking depending on the situational context, such as the type of addressee or the specific environment. Typically, speakers talk differently to children or adults and switch to a different style when they notice that their partner experiences some problems of understanding (e.g., because that person is not a native speaker) (Bortfeld & Brennan, 1997). Along the same lines, there may be differences across addressees, for example, depending on personality traits or the mere fact that some addressees have more developed communicative skills (Williams et al., 2021). It could be expected that addressees may vary in how they produce backchannel behaviors, with some spots in the interaction eliciting stronger or more backchannels than others (e.g., because such a cue is felt to be more needed). Also, some addressees may be more extraverted or engaged so that one could expect differences across addressees as well.

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

Furthermore, the characteristics of a BOP can influence the type of behavior it elicits. A BOP placed at the end of a complete syntactically complete phrase is more likely to be seized than a BOP at the end of a syntactically incomplete phrase (Skantze et al., 2013). The dynamics of the interaction could also play a role. Benus et al. (2007) show that the liveliness of an interaction may influence the type of verbal backchannels a participant uses. In their study, mm-hm and uh-huh were more used during lively interactions, while okay and yeah were used more during less animated interactions. Orthogonal to this, the reason why not every BOP is seized could also be due to idiosyncratic differences between listeners. Huang and Gratch (2012) examined the personalities of backchannel coders and explored the connection between these personalities and the frequency of identified BOPs. The results revealed a positive association between a higher number of identified BOPs and elevated levels of agreeableness, conscientiousness and openness. This is in line with the results of an earlier study that showed that different types of backchannel behavior correlate with various impressions of people's specific personalities (Blomsma et al., 2022).

Insight into the variability of audiovisual backchannel behavior is not only informative to understand how human–human communication proceeds, but it is also relevant for practical applications, such as models of human–computer interaction, specifically social robots and ECAs (Cassell et al., 2000), also known as socially interactive agents (SIAs) (Lugrin et al., 2021). In a similar manner to human–human interaction, it could be useful for ECAs to vary in the extent to which they backchannel, for example, depending on the type of user, context and application. It is also likely that inducing variability may render the interaction style of an ECA more natural and less monotonous, similar to the efforts to synthesize variability in speech and language generation systems (Gatt & Krahmer, 2018). However, modeling natural backchannel behavior for artificial entities is a non-trivial task for at least two reasons. One of the difficulties lies in detecting and appropriately responding to backchannel-inviting cues. Another difficulty is that due to backchannel behavior being idiosyncratic, it is not easy to define what a typical backchannel behavior should consist of for an ECA.

To investigate variation in backchannel behaviors and to answer the research questions above, we conducted a computational study based on the data collected in a human experiment that used the so-called O-Cam paradigm (Goodacre & Zadro, 2010). The current study is the first one in which the paradigm is used to examine backchannel behavior. The O-Cam paradigm was set up to allow comparisons between multiple addressees who are exposed to identical conversational data from the same speaker stimulus. The computational study consisted of two analyses. Analysis I examines the speaker stimulus, specifically the identification of BOPs, the categorization of those BOPs and the prosodic properties of the backchannel-inviting cues preceding the BOPs. Analysis II investigates the addressee's behavior during the BOPs. We compared the behavior of the addressees across multiple channels (i.e., facial expressions, head movement and vocalizations) to examine the degree of variability between and within addressees.

2. Dataset

This study employed the materials of a database previously recorded during an experiment conducted by Brugel (2014). The database consisted of (1) one video

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

recording of the stimulus, henceforth 'speaker', and (2) the video recordings of 14 participants who were filmed during the experiment, henceforth 'addressees'. Each video was 8.42 minutes long and contained 6.25 minutes of conversation, and the remaining time was used for game-related tasks such as preparing and answering questions (see explanation below). The number of participants is comparable to similar backchannel studies, including Krogsager et al. (2014) and Poppe et al. (2010).

The recorded experiment was based on the O-Cam paradigm (Goodacre & Zadro, 2010), an experimental design that combines the advantages of online paradigms (i.e., highly controllable environment, easy to run) with the advantages of offline settings (i.e., high ecological validity). The core concept of the O-Cam paradigm is that a participant thinks that he/she is having a computer-mediated conversation with another participant (i.e., an interaction via a video conferencing setting), while, in reality, the other participant is a confederate whose video is pre-recorded. Certain manipulations are used in the setup to make a participant think it is a real-life conversation (Goodacre & Zadro, 2010). The O-Cam paradigm has been previously utilized to, for example, study the relationship between gender and leadership capabilities (Hong et al., 2014) and investigate the influence of smiling behavior (Mui et al., 2018).

The experiment reported by Brugel (2014) was aimed to elicit feedback behavior from the participants. Each addressee played a Tangram game with the speaker (who was a pre-recorded confederate) via computer-mediated connection. During the experiment, the addressee was presented with four Tangram figures for 5 seconds, followed by a description of one of those Tangrams provided by the speaker. The participant's task was to choose the figure from the four Tangram figures based on the description by the speaker. See Figure 1 for a visual illustration of the experiment. The experiment consisted of 11 rounds in which each time a different quadruple of Tangram figures would be used. The participants were told that the experiment was related to abstract thinking and that they were not allowed to ask questions since

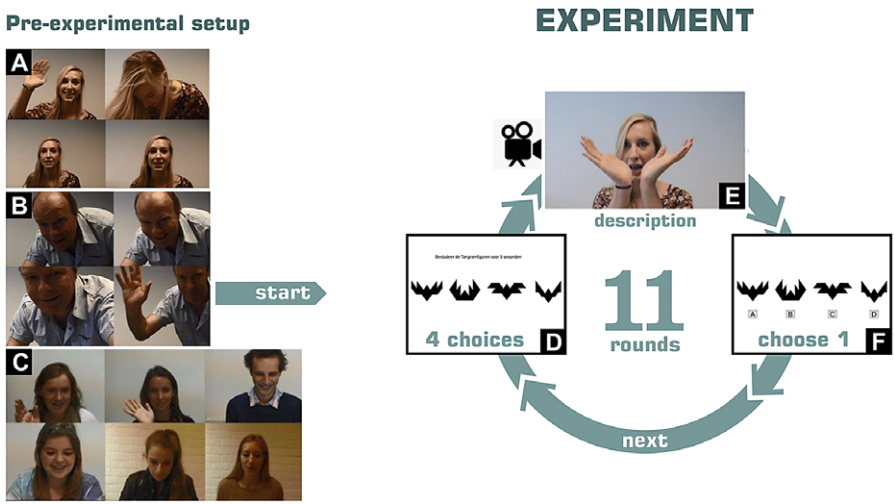

Figure 1. Visual impression of the o-cam experiment. First, the participant is prepared (A–C); after that, 11 rounds are played: In each round, the participant is shown four figures (D), followed by a description of one of those figures (E) after which the participant indicates which figure is described (F).

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

asking questions would make the game too simple. The confederate (the speaker) was not informed about the goal of the study in order to keep the experiment as ecologically valid as possible. Although task success was not measured, the primary objective of the game was to create a challenging experience with a task success rate close to 100%. This was intended to ensure that participants would fully concentrate on the speaker without feeling the need to ask additional questions for clarification, which would have been disruptive to the experimental setting, as the participant would then notice that the recorded confederate would not be responding to his/her questions. After the experiment, participants were asked whether they suspected that instead of a live interaction they were presented with a pre-recorded video of another person. The data of five participants were discarded because they answered positively, whereas one participant asked a question during the experiment, and thus, their data were also discarded.

3. Analysis I: speaker's behavior

The first analysis regards only the speaker's behavior to identify the BOPs and to analyze the audiovisual behavior of the speaker during the backchannel-inviting cues preceding the BOPs. The identified BOPs are subsequently used in Analysis II to investigate the addressee's feedback behavior. An obvious approach to identify the BOPs would be to annotate the backchannel behavior for each of the addressee videos separately. However, such an approach comes with at least two disadvantages. As addressees do not necessarily utilize all BOPs to provide feedback, analyzing the addressees would thus not necessarily result in the identification of all BOPs. Furthermore, using the same data for selection and selective analysis would result in a circular analysis also known as 'double dipping' (Kriegeskorte et al., 2009). Therefore, we identified the BOPs based on the speaker stimulus.

3.1. Methods

3.1.1. BOP identification

We used parasocial consensus sampling (Heldner et al., 2013; Huang et al., 2010), which takes the advantage of the fact that humans, especially as a third-party observer, can aptly point out BOPs in a conversation (de Kok, 2013). The approach consisted of two steps: identification of possible BOPs by a jury of multiple judges, followed by the aggregation of the output of the jury to determine genuine BOPs. Genuine BOPs are those BOPs that are identified by at least a certain percentage of judges.

For the identification of BOPs, we used a human jury that consisted of 10 judges. Each judge watched the speaker video and identified each moment that he/she thought was appropriate to backchannel. Each judge was instructed in the same way. First, they were explained what backchanneling behavior is; namely, the listening signals one gives during a conversation include head nods and sounds like 'uh-uh', 'hmm' and 'hm-hm' and combinations of nods and sounds. Next, they were asked to watch the speaker video and to make a sound (e.g., 'yes') when he/she thought it was appropriate to backchannel, either verbally, non-verbally or both. The audio of the judge was recorded.

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

The aggregation of all the recordings of judges allowed us to determine, for each data point in the stimulus, the percentage of judges that thought that a specific moment was a BOP. BOPs that were agreed upon by a minimum percentage of judges were classified as genuine BOPs and selected for further analysis.

The minimum percentage is based on the expected number of backchannels in the recording. Poppe et al. (2011) state that one could expect from 6 to 12 backchannels per minute. Since our recording was 6.25 minutes, we therefore expected between 38 and 77 backchannels. The appropriate consensus level is determined as follows. First, the number of BOPs is calculated for each potential consensus level. That is, the number of BOPs that would be marked as genuine BOPs if that consensus level were used. Next, the final consensus level is selected based on the resulting BOP count. In this case, the BOP count should fall within the range of 38 and 77. In general, the relationship between consensus levels and number of BOPs could be seen as a monotonic non-increasing function: When the consensus level increases, the number of genuine BOPs either increases or stays constant; it never decreases.

All the recordings of judges were preprocessed with audacity (Audacity Team, 2021): We used a noise gate filter (250 ms attack and −12.50 dB grate threshold) to remove background noise and a 20 dB audio amplification to ensure that a judge was audible. Each recording was then converted to a binary time series with a resolution of 25 frames per second (FPS), in such a way that frames that contained a sound with an amplitude above 0.1 were converted to 1 and, otherwise, to 0. Although Huang et al. (2010) used a resolution of 10 FPS, we decided to use 25 FPS as this matched with the FPS of both our video recording and the FaceReader encodings (as described in the subsequent section).

Because judges had to vocally indicate visual backchannels, which start on average 202 ms before a vocal backchannel (Wlodarczak et al., 2012), the onset of each indication was set to 202 ms before the actual onset in order to correct for a potential delay. Each onset of a judge's indication was converted to a potential BOP of the length of 1000 ms in line with Huang et al. (2010). Finally, a time series was created with a resolution of 25 FPS, where each frame (i.e., sample) contained the number of judges that indicated a BOP for that frame.

3.1.2. BOP types: continuer and end-of-turn

To gain further insight into whether specific BOPs or BOP types affect the average addressee's behavior, we subdivided the BOPs into two categories. Although each BOP functions as a moment for the addressee to acknowledge certain information, we conjecture that the urge to acknowledge is the strongest at the end of each game round. After all, no further information will follow the last BOP of a game round, and thus, the addressee should have enough information to answer the question at that point. And if not, the addressee should indicate that at that last BOP. Therefore, we estimate that the most expressive addressee's behaviors will be observable at the last BOP of a round. Hence, we have created the following categories: (1) All BOPs that are the last of a round, we called this category the last backchannel of round (LBR), and (2) all other BOPs that are placed during a round, we called this category continuer. Given this categorization, the LBR category contained 11 cues and the continuer category contained 42 cues.

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

3.1.3. Backchannel-inviting cues

To verify that indeed the (visual) prosody is different for backchannel-inviting cues compared to the prosody used during the remaining part of the conversation, we analyzed the pitch properties, facial behavior and head movement of the speaker's backchannel-inviting cues that preceded the identified BOPs. The cues were isolated by selecting the last 1000 ms of the speaker stimulus sound before the start of each BOP. However, there is no consensus on the length of such samples in literature; for example, Skantze (2012) analyzed the last 200 ms of the voiced region for pitch, while Levitan et al. (2011) reported longer sample lengths including 1000 ms. We choose 1000 ms to be on the safe side of finding a voiced part in the sample.

The pitch properties were extracted with Praat (Boersma & Weenink, 2022). Of each sample, the F0 values (i.e., the fundamental pitch values) were extracted with a precision of 100 FPS. Trailing and leading frames that did not contain pitch information were discarded. For each sample, the average, minimum, maximum, amplitude (which is the maximum minus the minimum), average and form were obtained. The form was calculated by subtracting the average pitch of the second half of the sample from the average pitch of the first half of the sample, such that a negative number for form means an increasing pitch and a positive number means a decreasing pitch.

The facial behavior and head movements were analyzed based on the output of FaceReader 8 software (Noldus, 2019). The stimulus video was encoded with action units (AUs) based on the Facial Action Coding System (Ekman & Friesen, 1978). Every frame of the videos was encoded with the following AUs: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, 26, 27 and 45, and X, Y and Z coordinates were extracted for head orientation. Each AU can be scored for intensity on an ordinal scale from 0 (i.e., absence of an AU) to 5 (i.e., maximum intensity). For some frames in the dataset, FaceReader was unable to detect a face and thus was also unable to encode head position and/or AU activations.

Head nods were quantified for all backchannel-inviting cues following Otsuka and Tsumori (2020). Specifically, for head nods, we extracted amplitude and frequency. Amplitude equals the maximum tilt angle, that is, the difference between the minimum and maximum X rotation angles. Frequency is the sum of upward and downward peaks per second of the X rotation angle. To prevent that small noise-related changes in elevation direction would influence the frequency, we ignored upward and downward peaks that differed a maximum of 1 degree. In order to verify whether the backchannel-inviting cues differed from non-backchannel-inviting cues, each backchannel-inviting cue was paired with a randomly selected voice sample from the speaker stimulus. Paired *t*-tests were conducted between the obtained pitch properties, head movements and the average AU activation of the backchannel-inviting cues and the non-backchannel-inviting cues. The Bonferroni correction was applied for the multiple pairwise comparisons. Subsequently, the analyzed properties of the backchannel-inviting cues of the LBR category were compared with those of the continuer category. The two categories were compared with Welch's *t*-test for significance and also corrected with Bonferroni.

3.2. Results

3.2.1. BOP identification

The number of identified backchannels per response level is depicted in Figure 2. Genuine (i.e., definite) BOPs were based on a consensus level of 30% (three coders)

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

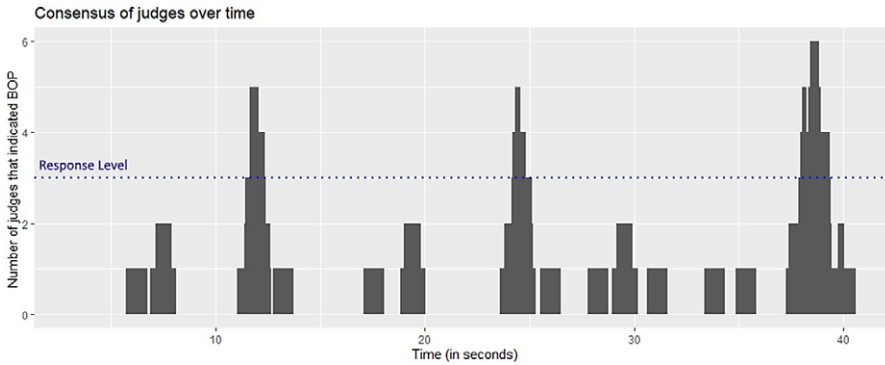

Figure 2. Illustration of a part of the speaker stimulus, with at each point in time the number of judges that indicated the presence of a BOP. If three judges or more indicated a BOP at a certain point, then this point is considered as a genuine BOP.

such that 53 BOPs were taken into account. The average duration of the 53 genuine BOPs was 934 ms (*SD* = 403 ms). The duration of a BOP was calculated starting from the initial timepoint with a consensus level of at least 30% and ending at the last timepoint where the consensus level was at least 30%.

3.2.2. Backchannel-inviting cues

The backchannel-inviting cues had a higher maximum pitch and a larger F0 range, compared to the randomly selected samples. There were no significant differences for average pitch, minimum and form. The highest pitch observed in backchannel-inviting cues was on average 350.36 Hz (*SD* = 106.94 Hz), while the highest pitch in the random samples had a lower average of 201.30 Hz (*SD* = 70.88 Hz). The F0 range for backchannel-inviting cues was on average 156.07 Hz (SD = 111.84 Hz), while the random samples had a lower average F0 range of 102.34 Hz (*SD* = 71.77 Hz). See Table 1 for all the results. The speaker's head movements and facial behavior did not differ significantly between cues and non-cues and also not between LBR and continuer-related inviting cues (see Tables 2 and 3). For all comparisons, the Bonferroni correction was applied.

The backchannel-inviting cues that preceded BOPs from the LBR category had a significantly lower average pitch, as compared to the cues that preceded the continuer

Table 1. Pitch properties of backchannel-inviting cues, compared to those of non-cues

	Cue (1)	Non-cue (2)	Diff (1) − (2)	df	Cohen's *d*	*p*-value
Average	246.95 (37.23)	250.72 (40.63)	−3.77	48	0.10	.740
Min	194.29 (43.44)	303.64 (54.23)	−109.35	48	0.14	.443
Max	350.36 (106.94)	201.30 (70.88)	149.06*	48	0.51	.006
Amplitude	156.07 (111.84)	102.34 (71.77)	53.73*	48	0.57	.006
Form	16.10 (52.87)	16.66 (49.74)	−32.76	48	0.45	.057

Note: Statistics are based on paired *t*-test analysis. All values are in Hertz. The Diff score is the result of subtracting the mean cue value from the mean random value. The Bonferroni correction was applied for the multiple pairwise comparisons with an alpha level of 0.01 (0.05/5 = 0.01). **p* <. 05, ***p* <. 01, ****p* <. 001.

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

Table 2. Averages of different channels (and standard deviations) over backchannel-inviting cues and non-backchannel-inviting cues

	Cue (1)	Non-cue (2)	Diff	df	Cohen's *d*
Head movement					
Frequency	4.80(2.41)	4.65(3.37)	0.15	49	0.05
Amplitude	11.39 (7.43)	9.99 (7.37)	1.41	49	0.19
Facial gestures					
Inner Brow Raiser (AU1)	0.33 (0.49)	0.32 (0.48)	0.02	49	0.04
Outer Brow Raiser (AU2)	0.23 (0.52)	0.09 (0.29)	0.14	49	0.34
Brow Lowerer (AU4)	0.53 (0.84)	0.38 (0.65)	0.15	49	0.19
Upper Lid Raiser (AU5)	0.05 (0.24)	0.02 (0.14)	0.03	49	0.16
Cheek Raiser (AU6)	0.64 (0.73)	0.40 (0.56)	0.24	49	0.36
Lid Tightener (AU7)	0.00 (0.03)	0.06 (0.22)	−0.05	49	0.34
Nose Wrinkler (AU9)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00
Upper Lip Raiser (AU10)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00
Lip Corner Puller (AU12)	0.06 (0.23)	0.03 (0.14)	0.03	49	0.16
Dimpler (AU14)	0.66 (0.91)	1.05 (1.20)	−0.39	49	0.39
Lip Corner Depressor (AU15)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00
Chin Raiser (AU17)	0.07 (0.27)	0.05 (0.21)	0.01	49	0.06
Lip Puckerer (AU18)	0.02 (0.12)	0.04 (0.16)	−0.01	49	0.08
Lip Stretcher (AU20)	0.00 (0.00)	0.01 (0.06)	−0.01	49	0.20
Lip Tightener (AU23)	0.00 (0.00)	0.00 (0.02)	−0.00	49	0.20
Lip Pressor (AU24)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00
Lips Part (AU25)	0.00 (0.03)	0.01 (0.06)	−0.01	49	0.17
Jaw Drop (AU26)	2.43 (0.89)	2.53 (0.87)	−0.10	49	0.12
Mouth Stretch (AU27)	0.09 (0.27)	0.17 (0.45)	−0.07	49	0.20
Eyes Closed (AU43)	0.00 (0.00)	0.00 (0.00)	0.00	49	0.00

Note: Statistics are based on paired *t*-test analysis. The Diff score is the result of subtracting the mean BOP value from the mean non-BOP value. No significant results were found in this analysis.

category. The form was also markedly different, and LBRs have a downward going pitch on average, while the other cues have an upward going pitch on average. There were no significant differences for minimum, maximum and amplitude. For an overview of the results, see Table 4.

4. Analysis II: addressee's behavior

In the following subsection, we first compare audiovisual feedback behavior at BOP and non-BOP spots in the spoken messages. Then, we focus on BOPs only to see to what extent we can observe variability in audiovisual feedback behavior within and between addressees.

4.1. Methods

4.1.1. Semi-automated measures of audiovisual behavior

The videos from the addressees were all encoded for facial expressions, head movements and vocal backchannels as follows. The head movements and facial behavior were analyzed analog to how the backchannel-inviting cues were analyzed (see Section 3.1.3). The vocal backchannels of the addressee videos were manually encoded by one coder with ELAN 6.0 encoding software (Wittenburg et al., 2006). The coder indicated the moments that an addressee made a sound and its duration. The vocal backchannels were quantified for all identified BOPs as follows: If an

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

Table 3. Averages of different speaker channels (and standard deviations) of backchannel-inviting cues that precede LBRs vs cues that precede continuers

	LBR (1)	Continuer (2)	Diff	df	Cohen's *d*
Head movement					
Frequency	5.27 (2.72)	4.68 (2.34)	0.60	14.30	0.25
Amplitude	13.70 (8.33)	10.76 (7.15)	2.94	14.30	0.40
Facial gestures					
Inner Brow Raiser (AU1)	0.50 (0.58)	0.29 (0.47)	0.21	4.20	0.44
Outer Brow Raiser (AU2)	0.44 (0.65)	0.17 (0.47)	0.26	12.94	0.52
Brow Lowerer (AU4)	0.96 (1.22)	0.41 (0.68)	0.54	11.73	0.67
Upper Lid Raiser (AU5)	0.00 (0.00)	0.06 (0.27)	−0.06	39.00	0.27
Cheek Raiser (AU6)	0.96 (0.90)	0.55 (0.67)	0.41	13.17	0.56
Lid Tightener (AU7)	0.00 (0.00)	0.01 (0.04)	−0.01	39.00	0.18
Nose Wrinkler (AU9)	0.00 (0.00)	0.00 (0.00)	0.00	0.00	0.00
Upper Lip Raiser (AU10)	0.00 (0.00)	0.00 (0.00)	0.00	0.00	0.00
Lip Corner Puller (AU12)	0.00 (0.00)	0.07 (0.25)	−0.07	39.00	0.31
Dimpler (AU14)	0.98 (1.12)	0.58 (0.84)	0.40	13.28	0.45
Lip Corner Depressor (AU15)	0.00 (0.00)	0.00 (0.00)	0.00	0.00	0.00
Chin Raiser (AU17)	0.02 (0.07)	0.08 (0.30)	−0.06	48.72	0.21
Lip Puckerer (AU18)	0.02 (0.07)	0.03 (0.13)	0.00	30.90	0.03
Lip Stretcher (AU20)	0.00 (0.00)	0.00 (0.00)	0.00	0.00	0.00
Lip Tightener (AU23)	0.00 (0.00)	0.00 (0.00)	0.00	0.00	0.00
Lip Pressor (AU24)	0.00 (0.00)	0.00 (0.00)	0.00	0.00	0.00
Lips Part (AU25)	0.00 (0.00)	0.00 (0.03)	0.00	39.00	0.18
Jaw Drop (AU26)	2.53 (0.82)	2.40 (0.92)	0.13	17.60	0.15
Mouth Stretch (AU27)	0.25 (0.44)	0.06 (0.18)	0.19	10.97	0.75
Eyes Closed (AU43)	0.00 (0.00)	0.00 (0.00)	0.00	0.00	0.00

Note: Statistics are based on paired *t*-test analysis. The Diff score is the result of subtracting the mean BOP value from the mean non-BOP value. No significant results were found in this analysis.

Table 4. Pitch properties of backchannel-inviting cues that precede LBRs vs cues that precede continuers

	LBR (1)	Continuer (2)	Diff (1) − (2)	df	Cohen's *d*	*p*-value
Average	218.68 (13.05)	253.68 (38.02)	−35.00***	43.30	1.00	1.48e-05
Min	159.37 (46.51)	202.61 (38.79)	−43.24	12.16	1.07	.350
Max	318.66 (117.47)	357.91 (104.37)	−39.25	12.60	0.37	.018
Amplitude	159.28 (129.43)	155.30 (108.99)	3.98	12.22	0.04	.929
Form	23.25 (21.97)	−25.47 (53.88)	48.72***	36.72	0.98	6.68e-05

Note: Statistics are based on Welch's *t*-test analysis. All values are in Hertz. The Bonferroni correction was applied for the multiple pairwise comparisons with an alpha level of 0.01 (0.05/5 = 0.01). *p <. 05, **p <. 01, ***p <. 001.

addressee made a sound during a BOP, the BOP was represented by 1 for the addressee or else by 0.

4.1.2. Comparisons of audiovisual behavior in BOPs versus non-BOPs

To understand whether the behavior of addressees differed between the BOPs and the rest of the conversation, we paired each BOP with a random non-BOP of the same length. A non-BOP is a moment in the conversation for which none of the judges thought it was a BOP. We compared the behavior of all addressees for a specific BOP with the behavior exhibited at the same non-BOP. Paired *t*-tests were carried out over all encoded channels. Pairs that contained frames that FaceReader was unable to encode were discarded. To determine how backchannel behavior differs across

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

addressees, we calculated the average behavior per addressee and reported the average behavior across all addressees. The Bonferroni correction was applied for the multiple pairwise comparisons.

4.1.3. BOP types: continuer and end-of-turn

The differences in behavior between continuer BOPs and LBR BOPs were quantified with Welch's *t*-test and corrected with the Bonferroni method.

4.2. Results

Overall, the behaviors during BOPs and non-BOPs differed markedly, except that we did not find any differences regarding minute facial expressions related to the AUs (see Table 5 and Figures 3–5). Even though the standard deviations for amplitude and frequency were high, there was a significant difference between the head movement of an addressee during a BOP and a non-BOP. On average, the frequency of head movement during a BOP was 3.43 upward/downward peaks per second, being 0.68 higher than the frequency in a non-BOP. The average amplitude was 5.95 degrees, which was 1.87 higher than that in a non-BOP. Across all BOP instances, 28% of the time, vocalizations were produced, while during non-BOP instances, this occurred only 3% of the time. The behavior of the facial muscles was generally the same during

Table 5. Averages of different channels (and standard deviations) over BOPs and non-BOPs

	BOP	non-BOP	Diff	df	Cohen's *d*	*p*-value
Head movement						
Frequency	3.43 (3.09)	2.75 (2.76)	0.68***	705	0.23	8.57e-06
Amplitude	5.95 (5.69)	4.07 (3.74)	1.87***	709	0.39	3.57e-16
Vocalizations	0.28 (0.45)	0.04 (0.18)	0.25***	741	0.72	2.2e-16
Facial gestures						
Inner Brow Raiser (AU1)	0.01 (0.12)	0.01 (0.08)	0.00	712	0.01	.812
Outer Brow Raiser (AU2)	0.00 (0.03)	0.00 (0.02)	0.00	712	0.02	.741
Brow Lowerer (AU4)	0.07 (0.26)	0.08 (0.27)	0.00	712	0.01	.635
Upper Lid Raiser (AU5)	0.00 (0.04)	0.01 (0.06)	0.00	712	0.04	.471
Cheek Raiser (AU6)	0.32 (0.65)	0.39 (0.75)	−0.07	712	0.09	.056
Lid Tightener (AU7)	0.11 (0.28)	0.19 (0.28)	0.01	712	0.05	.291
Nose Wrinkler (AU9)	0.00 (0.00)	0.00 (0.00)	0.00	712	0	–
Upper Lip Raiser (AU10)	0.11 (0.30)	0.08 (0.27)	0.03	712	0.09	.032
Lip Corner Puller (AU12)	0.78 (1.00)	0.87 (1.10)	−0.09	712	0.08	.087
Dimpler (AU14)	0.01 (0.08)	0.02 (0.14)	−0.01	712	0.11	.036
Lip Corner Depressor (AU15)	0.01 (0.10)	0.01 (0.09)	0.01	712	0.05	.356
Chin Raiser (AU17)	0.23 (0.55)	0.23 (0.57)	0.00	712	0.00	.992
Lip Puckerer (AU18)	0.00 (0.04)	0.01 (0.09)	0.01	712	0.09	.105
Lip Stretcher (AU20)	0.00 (0.02)	0.01 (0.08)	−0.01	712	0.09	.074
Lip Tightener (AU23)	0.02 (0.16)	0.02 (0.12)	0.01	712	0.04	.439
Lip Pressor (AU24)	0.17 (0.39)	0.15 (0.38)	0.02	712	0.05	.237
Lips Part (AU25)	0.31 (0.85)	0.30 (0.88)	0.01	712	0.01	.797
Jaw Drop (AU26)	0.05 (0.30)	0.02 (0.17)	0.02	712	0.09	.065
Mouth Stretch (AU27)	0.00 (0.02)	0.00 (0.04)	0.00	712	0.03	.528
Eyes Closed (AU43)	1.39 (1.50)	1.35 (1.49)	0.04	712	0.03	.250

Note: Statistics are based on paired *t*-test analysis. The Diff score is the result of subtracting the mean BOP value from the mean non-BOP value. The Bonferroni correction was applied to correct for 23 comparisons. Alpha was set to 0.002 (=0.05/23). **p* <. 05, ***p* <. 01, ****p* <. 001.

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

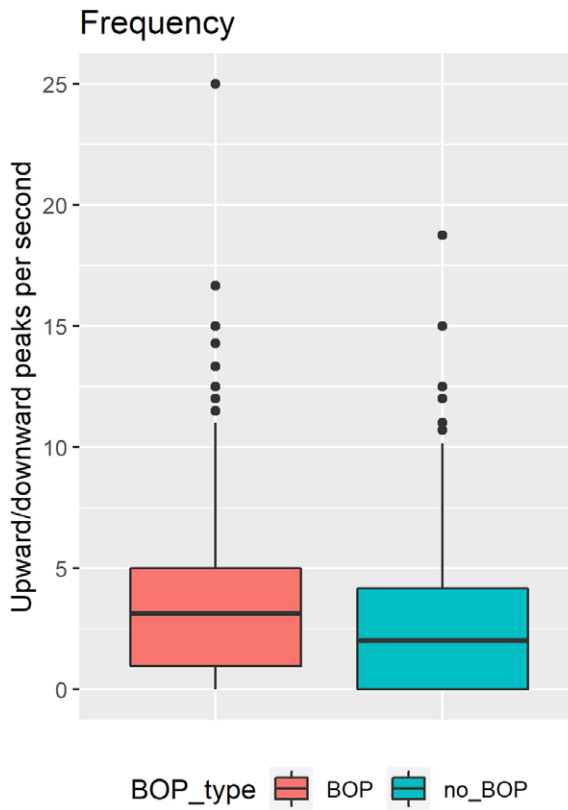

Figure 3. Head movement frequency during BOPs and outside of BOPs.

BOPs and non-BOPs (see Table 5 for an overview) and contained no significant differences.

4.2.1. Variation of backchannel behaviors across addressees

There was substantial variation regarding different behaviors across addressees. Head movement differed among the addressees. Although the mean frequency of head movement was 3.46 upward/downward peaks per second across addressees, the most nodding addressee showed 5.47 upward/downward peaks per second on average, compared to 1.49 upward/downward peaks per second on average for the least nodding addressee. Amplitude was on average 5.97 degrees, with the addressee on the lowest end having an amplitude of 3.34 degrees on average, while the addressee on the highest end showed an amplitude of 9.65 degrees on average. Addressees vocalized 28% of the BOPs on average, while the least vocal addressee only vocalized 4% of the BOPs and the most vocal addressee vocalized 59% of all BOPs. AU activations also varied; for example, the AU with the highest variation (*SD* = 1.39) was Eyes Closed (AU43), followed by Lip Corner Puller (AU 12) (*SD* =. 77). See Table 6 for a complete overview and Figure 6 for a visual inspection.

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

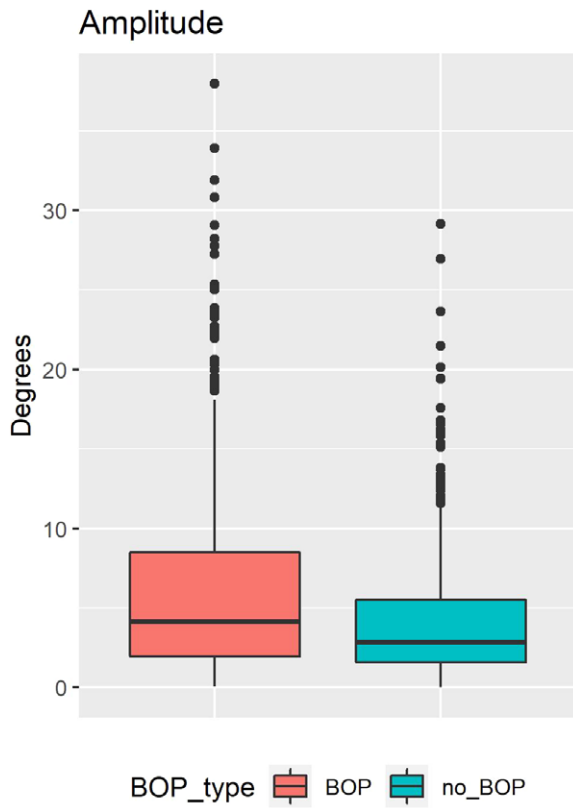

Figure 4. Head movement amplitude during BOPs and outside of BOPs.

4.2.2. Variation within addressees

The average addressee's behavior also differed across the different BOPs. Figure 7 shows the distribution of behavior per BOP. On average, the frequency was 3.42 upward/downward peaks per second across BOPs. However, BOP 35 elicited an average frequency of 1.10 upward/downward peaks per second, while at BOP 51, addressees showed an average frequency of 6.25 upward/downward peaks per second. The amplitude also varied; the mean amplitude across all BOPs was 5.96, while the minimal average amplitude was 0.65 degrees at BOP 51, and the maximum average amplitude was 14.1 degrees at BOP 11. Some BOPs (e.g., 12, 16, 17) were never vocalized, while other BOPs were vocalized by 93% of the addressees (e.g., BOP 26). The effect of addressee-dependent behavior is visually inspected in Figure 8. For a full overview of the numbers, see Table 7.

4.2.3. Variation within addressees for different BOP types

The BOPs that are marked as LBR BOP elicit higher nodding amplitudes from the addressees than the continuer BOPs; furthermore, LBRs let to more vocalizations, on average 60% of the time, while during the remaining BOPs, addressees vocalized 20%

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

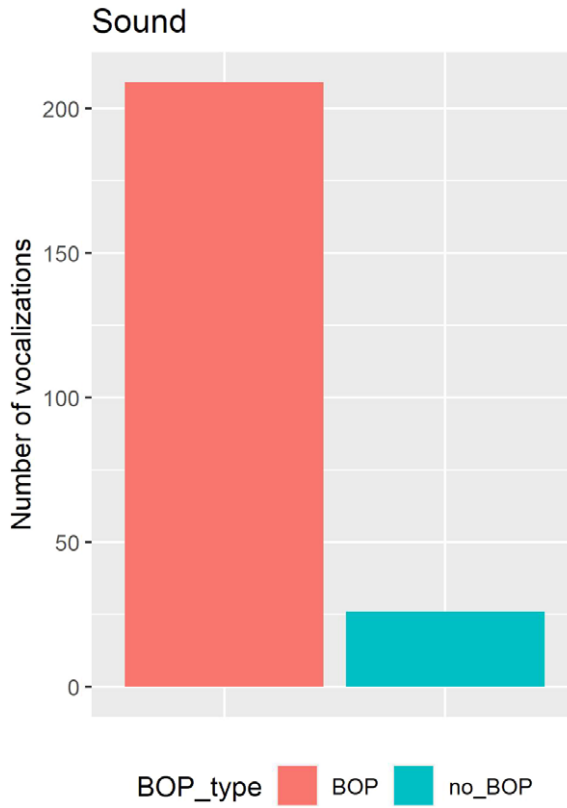

Figure 5. Number of times addressees vocalized BOPs versus the number of vocalized non-BOPs.

of the time, on average. Nodding frequency is not different between the two types of BOPs. For all the results, see Table 8.

5. Discussion

In this study, we were interested in a computational examination of the variability in backchannel behaviors among addressees. We looked at whether and how behavior varied during BOPs across and within addressees, specifically focusing on head movement, vocalizations and facial expressions produced by 14 addressees in a Tangram game. The game setup used the O-Cam paradigm, meaning that each addressee was exposed to exactly the same behaviors produced by the speaker. We showed that in general head movement and vocalization behavior significantly differed between BOPs and non-BOPs.

Nodding behavior and vocalizations were most pronounced during BOP instances, compared to non-BOP instances. However, it is notable that the amount of facial activity was generally the same during BOPs and non-BOPs, characterized by most AUs being activated at low-intensity levels. These low-intensity levels may be a

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

Table 6. Differences in feedback behavior between addressees

	Mean	SD	Min	Max
Head movement				
Frequency	3.45	1.21	1.49	5.45
Amplitude	5.97	2.10	3.34	9.65
Vocalizations	0.28	0.14	0.04	0.59
Facial gestures				
Inner Brow Raiser (AU1)	0.01	0.02	0	0.06
Outer Brow Raiser (AU2)	0.00	0.00	0	0.01
Brow Lowerer (AU4)	0.07	0.21	0	0.79
Upper Lid Raiser (AU5)	0.00	0.01	0	0.04
Cheek Raiser (AU6)	0.32	0.33	0	1.14
Lid Tightener (AU7)	0.11	0.14	0	0.44
Nose Wrinkler (AU9)	0	0	0	0
Upper Lip Raiser (AU10)	0.11	0.20	0	0.65
Lip Corner Puller (AU12)	0.77	0.51	0.19	1.82
Dimpler (AU14)	0.01	0.02	0	0.05
Lip Corner Depressor (AU15)	0.01	0.02	0	0.07
Chin Raiser (AU17)	0.25	0.41	0	1.29
Lip Puckerer (AU18)	0.00	0.01	0	0.05
Lip Stretcher (AU20)	0.00	0.00	0	0.01
Lip Tightener (AU23)	0.02	0.06	0	0.20
Lip Pressor (AU24)	0.17	0.25	0	0.91
Lips Part (AU25)	0.32	0.47	0.02	1.57
Jaw Drop (AU26)	0.05	0.15	0	0.58
Mouth Stretch (AU27)	0.00	0.00	0	0.01
Eyes Closed (AU43)	1.39	1.40	0	4.06

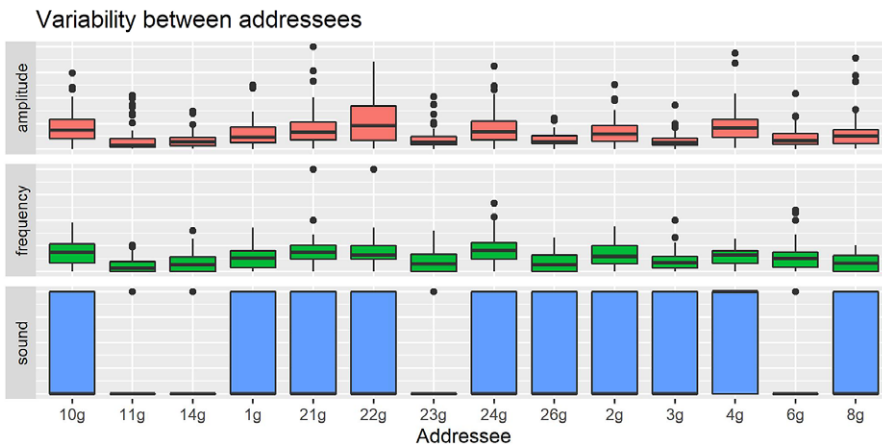

Figure 6. Values for head movement (frequency and amplitude) and vocalizations (sound) for each addressee. Frequency and amplitude are scaled, such that 1 represents the maximum value and 0 represents the lowest value.

consequence of the experimental setup, namely that addressees did not exhibit higher AU intensities because of the nature of interaction that the experimental setup (O-Cam paradigm) allowed. However, it is more likely that low facial activity during both BOPs and non-BOPs was the result of a general pattern, which is that during

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

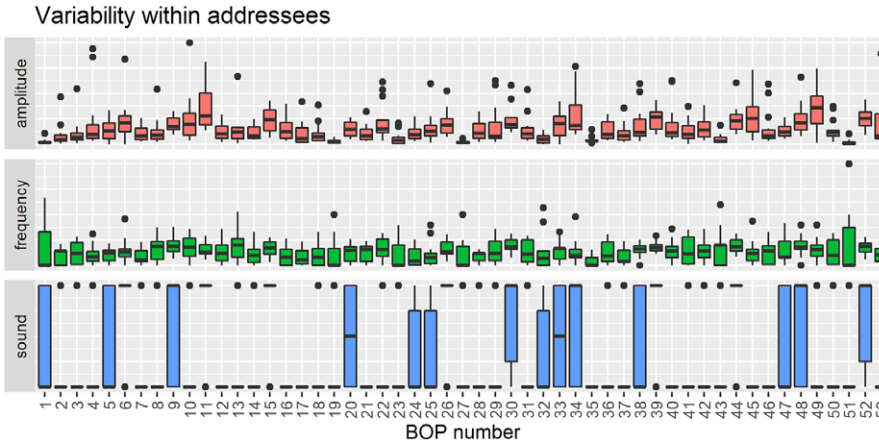

Figure 7. Values for head movement (frequency and amplitude) and vocalizations (sound) for each BOP. Frequency and amplitude are scaled, such that 1 represents the maximum value and 0 represents the lowest value.

natural interactions people rarely produce exaggerated facial expressions (Blomsma et al., 2020).

Further dissection of behavior during BOPs showed that there was person-specific variability. This between-addressee variability indicates that not every addressee demonstrated the same feedback behavior during BOPs. Some individuals were more discrete with their feedback behavior than others. In addition, the analysis indicated BOP-related differences. Some BOPs manifested more expressive behavior on average than others. Thus, in general, the timing of feedback behavior seems to adhere to certain rules. All addressees showed consistently different behavior during the BOPs than outside of the BOPs. However, the exact behavior seemed to be influenced by person-specific and BOP-related variables.

5.1. Variability between addressees

There was also variability between addressees. While all addressees nodded and vocalized during BOPs more than outside of them, there was variability in the extent to which addressees produced nodding and vocalizations during BOPs. Interestingly, the most vocal addressee produced a sound during more than half of the BOPs, a substantial difference from the least vocal addressee, who vocalized 14 times less. Likewise, the addressee with the smallest amplitude (addressee 14, with an average amplitude of 2.9) differed substantially from the person with the most pronounced amplitude (addressee 22 with an average amplitude of 7.4).

Given that all addressees were subject to the same experimental paradigm, the most likely source of this variation in backchannel behavior was the addressee's tendencies related to personality characteristics. In other words, while most BOPs were amenable to nods and vocalizations, addressees differed in the manifestation of their listening behaviors. Prior research shows that backchannel behavior can be

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

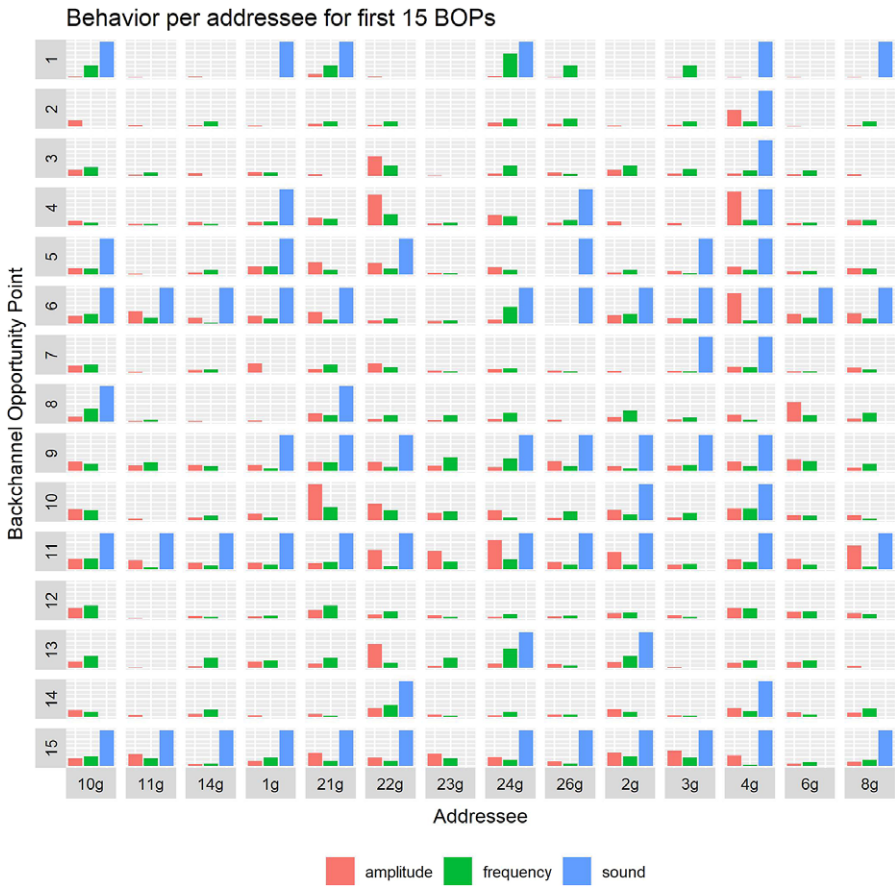

Figure 8. Behavior for each addressee per BOP. Only the first 15 BOPs due to visualization restrictions.

linked, to some extent, to the personality characteristics of a person as measured through the Big Five traits (Vinciarelli et al., 2015). In a follow-up experiment, we showed that the type of backchannel behavior indeed influences the personality perception of the listener. Listeners who produced head nods with a bigger amplitude are, for example, perceived as being more extraverted, compared to listeners whose head nods are smaller (Blomsma et al., 2022).

Other factors could include gender, and research showed that women tend to backchannel with a higher frequency than men and that backchanneling occurs more frequently in Japanese than in American English (Dixon & Foster, 1998; Furo, 2000; Maltz & Borker, 2018). Lastly, variability could also be (partly) caused by pure randomness.

In a future experiment, it would be valuable to take into account the characteristics of the addressee, such as personality, gender and cultural background, to identify factors that may play a role in producing the person-specific variability of feedback behavior. In addition, it would be beneficial to extend the length of the experiment to harvest more behavioral data from each addressee, which would allow to also shed

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

Table 7. Differences in feedback behavior within addressees

	Mean	SD	Min	Max
Head movement				
Frequency	3.42	1.01	1.10	6.25
Amplitude	5.96	3.11	0.65	14.1
Vocalizations	0.28	0.28	0	0.929
Facial gestures				
Inner Brow Raiser (AU1)	0.01	0.03	0	0.19
Outer Brow Raiser (AU2)	0.00	0.01	0	0.05
Brow Lowerer (AU4)	0.07	0.04	0	0.15
Upper Lid Raiser (AU5)	0.00	0.01	0	0.07
Cheek Raiser (AU6)	0.32	0.30	0	1.56
Lid Tightener (AU7)	0.11	0.08	0	0.27
Nose Wrinkler (AU9)	0	0	0	0
Upper Lip Raiser (AU10)	0.11	0.07	0	0.31
Lip Corner Puller (AU12)	0.79	0.53	0.14	2.51
Dimpler (AU14)	0.01	0.02	0	0.12
Lip Corner Depressor (AU15)	0.01	0.03	0	0.141
Chin Raiser (AU17)	0.24	0.11	0.05	0.45
Lip Puckerer (AU18)	0.00	0.01	0	0.05
Lip Stretcher (AU20)	0.00	0.01	0	0.03
Lip Tightener (AU23)	0.02	0.05	0	0.17
Lip Pressor (AU24)	0.17	0.08	0.04	0.42
Lips Part (AU25)	0.32	0.34	0.00	2.03
Jaw Drop (AU26)	0.04	0.07	0	0.25
Mouth Stretch (AU27)	0.00	0.00	0	0.03
Eyes Closed (AU43)	1.38	0.185	0.85	1.73

Table 8. Averages of head movement and vocalizations over continuer BOPs and last backchannel of round (LBR)

	LBR (1)	BOP (2)	Diff	df	Cohen's *d*	*p*-value
Frequency	3.19 (2.58)	3.06 (3.03)	0.13	509.23	0.04	.47
Amplitude	6.78 (5.99)	4.56 (4.48)	2.22***	372.78	0.46	8.97e-09
Vocalizations	0.60 (0.49)	0.20 (0.40)	0.40***	208.74	0.97	2.2e-16

Note: Statistics are based on Welch's *t*-test analysis. The Diff score is the result of subtracting the mean BOP value from the mean last backchannel of round (LBR) value. The Bonferroni correction was applied to correct for three comparisons. Alpha was set to 0.017 (=0.05/3). *p <. 05, **p <. 01, ***p <. 001.

light on potential intrapersonal variability, unrelated to BOP or person-specific characteristics. Although it is currently unknown what the time limits are of an o-cam experiment, we hypothesize that a longer experiment would result in more addressees that would find out that the speaker is pre-recorded.

5.2. Variability between BOPs

While nodding and vocalizations characterize spontaneous listening behavior, the high standard deviations regarding nodding behavior (i.e., amplitude and upward/downward peaks per second) suggest that different BOPs lead to the differing amount of nodding. This can be seen in Figure 8.

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

Regarding the current data, differing nodding patterns based on a BOP may partially be related to the fact that some Tangrams may have been more difficult to understand than others. That is, if an addressee quickly understood the description of a figure, they may have nodded more energetically compared to those instances where they doubted and hence nodded in a less pronounced fashion. This insight is related to the early research on non-verbal behavior conducted by Birdwhistell (1970), who showed that based on both the frequency of nods and their duration the involvement of an addressee was communicated differently. In particular, a single nod of 400 ms in duration acted as a strong affirmation of the speaker's behavior, while a nod of 800 ms or longer signaled disbelief and even elicited interruptions on the part of the speaker. Overall, this demonstrates that the nature of backchannels varies as the interaction unfolds. Our research also put forward a difference between behavior shown during the last BOPs of a round and BOPs that were located during a round. The last BOP of a round may have acted as a feedback point, but also as marking the end of a round. The addressee was signaled at this BOP that the moment of choosing the correct Tangram was near, and therefore, the function of the BOP was perhaps different than the other BOPs. The speaker was more 'asking' for a confirmatory signal from the addressee, than an acknowledging feedback signal. Indeed, the backchannel-inviting cues from the speaker were clearly different when signaling the last BOP of a round, compared to other BOPs. The speaker was using a downward inflection when signaling the last BOP of the round, compared to an upward inflection when signaling other BOPs, and used a lower pitch rate on average. In return, addressees were more expressive during the LBRs, in the sense that they vocalized more often and showed a higher amplitude in their nodding behavior. That backchannel-inviting cues have a lower pitch at the end of a round and have a downward inflection is in line with Geluykens and Swerts (1994), which show that speakers 'reserve' the low pitch to mark the end of a turn, while keep using a higher pitch in other cases to prevent that the turn is taken over by the opponent.

Given the variability in audiovisual behavior between various BOPs, we looked at a few cases in more detail to gain insight into possible reasons for the differences. In particular, we did some speculative analyses of BOP 26, which was vocalized by 92% of the addressees and received relatively frequent head nods (4.36), versus BOP 16, which was not vocalized by any of the addressees and not frequently marked by head nods (2.21). Comparing these two instances yields the impression that the strength of feedback cue (in terms of nodding and auditory backchanneling) is related to the degree to which the speaker signals that the information she provided is complete. BOP 26 occurs at the end of round 5, just after the speaker said, 'That's the one you have to pick. So, a square chimney and a triangle from the side of the house'. During this BOP of 1000 ms, the speaker is completely silent. The speaker appears to cue that she provided all the information the addressee needs to pick the correct Tangram figure and therefore expects a strong affirmative backchannel. BOP 16, on the contrary, occurs at the beginning of round 4, just at the end of short sentence from the speaker 'These are more like birds.', where it is clear that more details from the speaker are needed to be able to identify the Tangram she is describing. At this stage, a strong feedback cue from the addressee would seem less appropriate, given that the provided information is still incomplete, but an addressee may acknowledge that he/she is listening to the speaker and awaiting further details. Obviously, future work

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

is needed to determine whether these impressions would generalize to more conversational contexts.

5.3. Division of labor

Given the results described above, it is interesting to compare the audiovisual behavior of the speaker with that of the addressee. Admittedly, given that we only recorded one speaker, our claims related to her role would have to be explored further in future work, but based on our analyses so far, it appears that our speaker more consistently makes use of auditory cues than visual cues to elicit feedback from her addressees. Indeed, while we find some prosodic differences between BOPs and non-BOPs, there are no significant differences in facial activity. Conversely, the addressees appear to exploit visual cues more regularly than vocalizations to return feedback after BOPs. In other words, given the broader set of audiovisual cues that function within an interaction, these results suggest that a speaker is more often using auditory features and the listener is more often making use of silent, visual cues, except for BOPs that occur at the final edge of a turn where a speaker is basically signaling that she has arrived at the end of her turn and will stop talking.

While this would have to be explored further in the future, these results point to a division of labor between auditory and visual cues in the feedback mechanism of a conversation, with the former being more typical for the speaker and the latter for the addressee. The advantage of being able to access multiple channels is that their use can be distributed over conversation partners so that they can exchange information in parallel. For instance, while one person is talking, the other can return visual feedback, such as affirmative head nods or expressions of surprise or misunderstanding, that do not interfere with the speech produced by the other as these are produced in silence. If instead dialog partners were to produce speech simultaneously, miscommunication might well result from the overlapping sound streams, because the speech by one person might mask that of the other (Swerts & Krahmer, 2020).

5.4. Embodied conversational agents

Understanding variation in backchannel behaviors across addressees is important for applications in ECAs. If for a large portion of backchannels nodding and vocalizations can be produced to show that one is engaged and listening, future research could investigate the conditions under which these behaviors are necessarily produced and vice versa the conditions when there is a slim chance that either a nod or a vocalization will occur. Understanding this balance between variability and stability of backchannel behaviors across a human–human conversation can help make artificial agents that can give flexible feedback and that come across natural in human–computer conversations. Moreover, person-specific variability may be used in an ECA to augment gender, personality and cultural characteristics. In other research, we have shown that indeed specific backchannel behavior in an ECA can elicit specific personality perceptions by its audience. We copy-synthesized the feedback behavior of different addressees during various BOPs onto an ECA and asked participants to indicate the perceived personality characteristics of the ECA. Among other conclusions, we found that a higher nodding

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

amplitude during a BOP is perceived as more extroverted than a smaller nodding amplitude.

Previous studies show that when listening behaviors are missing or are poorly timed, communication is negatively affected and can go off the rails (Bavelas et al., 2000). The current findings suggest that there is no 'one listening behavior', but a variety of behaviors across different BOPs and across different addressees. And although nods and vocalizations are characteristic of spontaneous interactions, the degree to which they will be produced varies between addressees.

References

Audacity Team. (2021). *Audacity 3.1.3*. Audacity Team.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79, 941. http://doi.org/10.1037/0022-3514.79.6.941

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, 52, 566–580. http://doi.org/10.1111/j.1460-2466.2002.tb02562.x

Benus, S., Gravano, A., & Hirschberg, J. B. (2007). The prosody of backchannels in American English.

Bertrand, R., Ferré, G., Blache, P., Espesser, R., & Rauzy, S. (2007). Backchannels revisited from a multimodal perspective. In *Auditory-visual speech processing*(pp. 1–5). Kasteel Groenendaal.

Birdwhistell, R. L. (1970). *Kinesics and context: Essays on body motion communication*. University of Pennsylvania-Press. http://doi.org/10.9783/9780812201284

Blomsma, P., Skantze, G., & Swerts, M. (2022). Backchannel behavior influences the perceived personality of human and artificial communication partners. *Frontiers in Artificial Intelligence*, 5, 835298.

Blomsma, P. A., Vaitonyte, J., Alimardani, M., & Louwerse, M. M. (2020). Spontaneous facial behavior revolves around neutral facial displays. In *Proceedings of the 20th ACM international conference on intelligent virtual agents* (pp. 1–8). ACM. http://doi.org/10.1145/3383652.3423893

Boersma, P., & Weenink, D. (2022). *Praat: doing phonetics by computer (Version 6.2.10)*.

Bortfeld, H., & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23, 119–147.

Brugel, M. (2014). *Het effect van de eye gaze en lach van de spreker op het uitlokken van feedback bij de ontvanger*. PhD thesis, Master's thesis, Tilburg University.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied conversational agents*. MIT Press.

Cathcart, N., Carletta, J., & Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1* (pp. 51–58). Association for Computational Linguistics. http://doi.org/10.3115/1067807.1067816

Clark, H. H. (1996). *Using language*. Cambridge University Press. http://doi.org/10.1017/s0022226798217361

de Kok, I., & Heylen, D. (2010). Differences in listener responses between procedural and narrative task. In *Proceedings of the 2nd international workshop on social signal processing* (pp. 5–10). ACM.

de Kok, I. A. (2013). *Listening heads*. University of Twente.

Dixon, J. A., & Foster, D. H. (1998). Gender, social context, and backchannel responses. *The Journal of Social Psychology*, 138, 134–136.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23, 283.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: Investigator's Guide Consulting*. Psychologists Press.

Furo, H. (2000). Listening responses in Japanese and us English: Gender and social interaction. In *Social and cognitive factors in second language acquisition: Selected proceedings of the 1999 Second Language Research Forum* (pp. 445–457). Cascadilla.

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61, 65–170. http://doi.org/10.1613/jair.5477

https://doi.org/10.1017/langcog.2024.1 Published online by Cambridge University Press

Geluykens, R., & Swerts, M. (1994). Prosodic cues to discourse boundaries in experimental dialogues. *Speech Communication*, 15, 69–77.

Goodacre, R., & Zadro, L. (2010). O-cam: A new paradigm for investigating the effects of ostracism. *Behavior Research Methods*, 42, 768–774. http://doi.org/10.3758/brm.42.3.768

Granström, B., House, D., & Swerts, M. (2002). Multimodal feedback cues in human-machine interactions. In *Speech prosody 2002, international conference* (pp. 347–350). Laboratoire Parole et Langage.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R. J., & Morency, L.-P. (2006). Virtual rapport. In *International workshop on intelligent virtual agents* (pp. 14–27). Springer. http://doi.org/10.1007/11821830_2

Gravano, A., & Hirschberg, J. (2009). Backchannel-inviting cues in task-oriented dialogue. In *Tenth Annual Conference of the International Speech Communication Association* (pp. 1019–1022). ISCA.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25, 601–634. http://doi.org/10.1016/j.csl.2010.10.003

Heldner, M., Hjalmarsson, A., & Edlund, J. (2013). Backchannel relevance spaces. In *Nordic Prosody XI, Tartu, Estonia, 15–17 August 2012* (pp. 137–146). Peter Lang Publishing Group. http://doi.org/10.3726/978-3-653-03047-1/16

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53, 23–35.

Hjalmarsson, A., & Oertel, C. (2012). Gaze direction as a back-channel inviting cue in dialogue. In *IVA 2012 workshop on realtime conversational virtual agents*. Springer.

Hong, A., Schaafsma, J., van der Wijst, P., & Plaat, A. (2014). Taking the lead: Gender, social context and preference to lead. In *European conference on management, leadership & governance* (pp. 445–451). Academic Conferences International Limited.

Huang, L., & Gratch, J. (2012). Crowdsourcing backchannel feedback: understanding the individual variability from the crowds. In *Feedback behaviors in dialog*. ISCA.

Huang, L., Morency, L.-P., & Gratch, J. (2010). Parasocial consensus sampling: Combining multiple perspectives to learn virtual human behavior. In *Proceedings of the 9th international conference on autonomous agents and multiagent systems: Volume 1* (pp. 1265–1272). International Foundation for Autonomous Agents and Multiagent Systems.

Kawahara, T., Yamaguchi, T., Inoue, K., Takanashi, K., & Ward, N. G. (2016). Prediction and generation of backchannel form for attentive listening systems. In *Interspeech* (pp. 2890–2894). ISCA. http://doi.org/10.21437/interspeech.2016-118

Krahmer, E., Swerts, M., Theune, M., & Weegels, M. (2002). The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. *Speech Communication*, 36, 133–145. http://doi.org/10.1016/s0167-6393(01)00030-9

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, 12, 535–540.

Krogsager, A., Segato, N., & Rehm, M. (2014). Backchannel head nods in Danish first meeting encounters with a humanoid robot: The role of physical embodiment. In *International conference on human-computer interaction* (pp. 651–662). Springer.

Levitan, R., Gravano, A., & Hirschberg, J. (2011). Entrainment in speech preceding backchannels. In *ACL (Short Papers)* (pp. 113–117). ACL. http://doi.org/10.7916/D89Z9DCS

Lugrin, B., Pelachaud, C., & Traum, D. (eds.) (2021). *The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 1: methods, behavior, cognition* (vol. 37, 1 edn). Association for Computing Machinery.

Maltz, D. N., & Borker, R. A. (2018). A cultural approach to male-female miscommunication. In *The matrix of language* (pp. 81–98). Routledge.

Morency, L.-P., De Kok, I., & Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. In *International workshop on intelligent virtual agents* (pp. 176–190). Springer.

Mui, P. H., Goudbeek, M. B., Roex, C., Spierts, W., & Swerts, M. G. (2018). Smile mimicry and emotional contagion in audio-visual computer-mediated communication. *Frontiers in Psychology*, 9, 2077. http://doi.org/10.3389/fpsyg.2018.02077

Noldus. (2019). *FaceReader: Tool for automated analysis of facial expression: Version 8.0*. Noldus Information Technology.

Otsuka, K., & Tsumori, M. (2020). Analyzing multifunctionality of head movements in face-to-face conversations using deep convolutional neural networks. *IEEE Access*, 8, 217169–217195. http://doi.org/10.1109/access.2020.3041672

Poppe, R., Truong, K. P., & Heylen, D. (2011). Backchannels: Quantity, type and timing matters. In *International workshop on intelligent virtual agents* (pp. 228–239). Springer. http://doi.org/10.1007/978-3-642-23974-8_25

Poppe, R., Truong, K. P., Reidsma, D., & Heylen, D. (2010). Backchannel strategies for artificial listeners. In *International conference on intelligent virtual agents* (pp. 146–158). Springer.

Shimojima, A., Katagiri, Y., Koiso, H., & Swerts, M. (2002). Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses. *Speech communication*, 36, 113–132. http://doi.org/10.1016/s0167-6393(01)00029-2

Skantze, G. (2012). A testbed for examining the timing of feedback using a map task. In *Proceedings of the interdisciplinary workshop on feedback behaviors in dialog*. KTH.

Skantze, G., Hjalmarsson, A., & Oertel, C. (2013). Exploring the effects of gaze and pauses in situated human robot interaction. In *Proceedings of the SIGDIAL 2013 Conference* (pp. 163–172). ACL.

Swerts, M., & Krahmer, E. (2020). Visual Prosody Across Cultures. In *The Oxford handbook of language prosody* (pp. 477–485). Oxford University Press. http://doi.org/10.1093/oxfordhb/9780198832232.013.45

Vinciarelli, A., Chatziioannou, P., & Esposito, A. (2015). When the words are not everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls. *Frontiers in ICT*, 2, 4. http://doi.org/10.3389/fict.2015.00004

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32, 1177–1207. http://doi.org/10.1016/s0378-2166(99)00109-5

Williams, G. L., Wharton, T., & Jagoe, C. (2021). Mutual (mis) understanding: Reframing autistic pragmatic "impairments" using relevance theory. *Frontiers in Psychology*, 12, 1277. http://doi.org/10.3389/fpsyg.2021.616664

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)* (pp. 1556–1559). ACL.

Wlodarczak, M., Buschmeier, H., Malisz, Z., Kopp, S., & Wagner, P. (2012). Listener head gestures and verbal feedback expressions in a distraction task. In *Proceedings of the interdisciplinary workshop on feedback behaviors in dialog, INTERSPEECH2012 satellite workshop* (pp. 93–96). Universität Bielefeld.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting, 1970* (pp. 567–578). Chicago Linguistic Society.

Cite this article: Blomsma, P., Vaitonyté, J., Skantze, G., & Swerts, M. (2024). Backchannel behavior is idiosyncratic, *Language and Cognition*, 1–24. https://doi.org/10.1017/langcog.2024.1