‘Are you sure you’re paying attention?’ – ‘Uh-huh’ Communicating understanding as a marker of attentiveness

Hendrik Buschmeier[1], *Zofia Malisz*[2], *Marcin Włodarczak*[2], *Stefan Kopp*[1], *Petra Wagner*[2]

[1]Sociable Agents Group, CITEC and Faculty of Technology
[2]Faculty of Linguistics and Literary Studies
Bielefeld University, Bielefeld, Germany

{hbuschme,skopp}@techfak.uni-bielefeld.de,
{zofia.malisz,petra.wagner,mwlodarczak}@uni-bielefeld.de

Abstract

We report on the first results of an experiment designed to investigate properties of communicative feedback produced by non-attentive listeners in dialogue. Listeners were found to produce less feedback when distracted by an ancillary task. A decreased number of feedback expressions communicating understanding was a particularly reliable indicator of distractedness. We argue this finding could be used to facilitate recognition of attentional states in dialogue system users.

Index Terms: communicative feedback; dialogue; distraction; engagement; attention; dual task

1. Introduction

Short feedback utterances (e.g., ‘uh-huh’, ‘m’, ‘yeah’, ‘okay’) are very characteristic of listener behaviour. The presence of feedback is necessary to facilitate grounding as dialogue partners must be reassured they are heard and understood. It can be said that typical feedback utterances are the minimal required spoken signals that sustain interaction. Nonetheless, conversational situations exist where listeners are being distracted by simultaneous tasks (browsing the Internet, reading documents, etc.) or disengaged for other reasons. We are interested in if and how listener behaviour changes in such situations.

Previous research focused mainly on speaker behaviour, particularly on the effect of impoverished or unexpected listener feedback on the speaker. [1] devised a method that made it possible to manipulate and control a listener’s state of attention over the course of a dialogue without the speaker knowing. By comparing the resulting conversations with dialogues recorded under normal conditions (where the listeners were not distracted experimentally), the authors found that distracted listeners produced less context-specific feedback, which in turn had a substantial influence on the speakers’ behaviour as well as the quality of the storytelling.

These findings were recently refined by [2], who distracted listeners with the same method. In the study speakers were instructed to tell two jokes and were informed (rightly or falsely) whether the listener had already heard the joke. Both listeners’ attention and speakers’ expectations thereof had a significant effect on the storytelling. Speakers told more vivid stories when they expected an attentive listener and in fact interacted with one. Speakers also spent more time telling their stories when their expectations of listeners’ attention states matched reality.

Both studies showed that distracted listeners had an influence on speakers and their behaviour. Consequently, speakers must somehow be able to notice that their dialogue partners are distracted. As [2, p. 582] note, speakers are “painfully aware when their conversational partners [...] are inattentive, and they can often tell when their partners are only pretending to pay attention.”

In this paper, we assume distractedness should manifest in the listeners’ communicative behaviour. We look for and analyse differences in the behaviour of distracted as opposed to attentive listeners, specifically differences in the way different types of feedback occur depending on listeners’ attentional states. For classification of feedback types we adopt a framework proposed by [3] which distinguishes four basic feedback functions: *contact* (willingness and ability to continue the interaction), *perception* (willingness and ability to perceive the message), *understanding* (willingness and ability to understand the message) and *attitudinal reactions* (willingness and ability to respond to the message).

The paper is organised as follows: Section 2 presents the experiment design and Section 3 gives an overview of our annotation scheme. In Section 4 we describe and discuss results of the study. Finally, Section 5 summarises our findings, and outlines prospects for further research.

2. Study design

In order to analyse how feedback behaviour changes in situations where listeners are distracted, we carried out a face-to-face dialogue study. One of the dialogue partners (the ‘storyteller’) told two holiday stories to the other participant (the ‘listener’), who was instructed to listen actively, make remarks and ask questions.

Building upon the paradigm of [1], we distracted the listeners by instructing them to press a button on a hidden remote control every time their dialogue partner uttered a word starting with the letter ‘s’ (which is the second most common German word-initial letter and usually corresponds to perceptually salient sibilants). They also had to count the total number of ‘s-words’ they heard. Storytellers were informed that their partners would be listening for something in the dialogue but they did not know during which story.

To enable a direct comparison between the listeners’ feedback behaviour in the distracted condition and their normal behaviour, experimental condition varied within subject: sto-

The first three authors are listed in alphabetical order.

Figure 1: The experimental setup with a dyad interacting in the distracted condition (note the listener using the remote control with her left hand).

rytellers had to tell two different holiday stories and listeners only engaged in the distraction task for either the first (in even-numbered sessions) or the second story (in odd-numbered sessions). Participants were assigned to their roles randomly.

Participants were positioned approximately three metres apart to minimise crosstalk. Close talking high-quality headset microphones were used. Furthermore, another microphone captured the whole scene and a fourth audio channel was used to record the 'clicks' synthesised by a computer when listeners pressed the button on the remote control. Interactions were recorded from three camera perspectives: medium shots showing the storyteller and the listener – enabling future fine-grained analysis of their head and arm gestures as well as their facial expressions – and a long shot showing the whole scene. Figure 1 shows one of the dyads from all three perspectives.

A total of fifty students (34 female and 16 male native speakers of German) were recruited at Bielefeld University to participate in the study, receiving either course credit or 4 euro as payment. They were assigned to one of 25 same-sex dyads. Most dialogue partners were unacquainted, however, four participant pairs knew each other before the study.

3. Annotation

Existing annotation schemes are often limited in their characterisation of feedback meaning. [1], for instance, differentiated between two categories of listeners responses: 'generic' and 'specific'. Similarly, [5] categorised affirmative feedback expressions into 'backchannels' and 'acknowledgement/agreement'.

Given our focus on the listener, we needed to describe feedback in more finely grained ways. We devised an annotation scheme (see Table 1) in which the first three levels of positive and negative feedback largely correspond to basic communicative functions of feedback as defined by [3, 4]. Our category P1 corresponds to the broad definition of the backchannel as a 'continuer', category P2 signals successful interpretation of the message, and category P3 indicates acceptance, belief and agreement. These levels can be treated as a hierarchy with increasing value of judgement and 'cognitive involvement' or 'depth' of grounding. Categories N1–N3 are the negative counterparts of the respective functions.

Following the theoretical implications of [4], emotional and attitudinal evaluation of the message was incorporated as a modifier A to the main categories (leading to labels such as P2A).

Table 1: Inventory of feedback functions. Categories P1–P3, N1–N3 and the modifier A are based on [3, 4]. Modifiers C and E were adopted from [5].

C / M	Definition of category or modifier
P1	The partner signals perception of the signal. *'I hear you and please continue.'*
N1	The partner signals problems with perception. *'What are you saying?'*
P2	The partner signals perception and understanding of the message content. *'I understand what you mean.'*
N2	The partner signals perception of the message and problems with understanding the content. *'I do not understand what you mean.'*
P3	The partner signals perception, understanding and acceptance of the message or agreement with the message. *'I accept / agree / believe what you say.'*
N3	The partner signals perception and understanding but rejection or disagreement of the message. *'I disagree / do not accept what you say.'*
A	The partner expresses an attitude towards the message, e.g., surprise, excitement, admiration, anger, disgust.
C	The partner introduces a new discourse segment or topic.
E	The partner ends the current discourse segment or topic.
?	Unresolved.

Two other modifiers, C and E, indicating turn or discourse topic boundaries can be used in similar ways and were adopted from [5]. Each of the modifiers could also be used as a category on its own. The category A, for instance, is used when a feedback signal is solely evaluative.

So far, feedback utterances in 14 of the 25 sessions in our corpus were segmented and transcribed according to German orthographic conventions (where existent). Feedback functions were annotated independently by three annotators taking communicative context into account. Majority labels between annotators were then calculated automatically and problematic cases (110; roughly 10%) were discussed and resolved. Modifiers were appended to a resulting majority label if they were used by at least one annotator so that subtle (especially emotion-related) distinctions were preserved.

4. Results and discussion

The 28 dialogues annotated so far have a total length of 180 minutes and each dialogue has a mean length of 6 minutes 25 seconds (Min = 2:16; Max = 14:29; SD = 2:31). A total number of 1003 feedback signals was identified, resulting in a mean of 36 feedback signals per dialogue (Min = 7; Max = 93; SD = 23.1). High standard deviation of these values indicates substantial differences between sessions.

4.1. Distraction task effect on feedback count

The overall count of annotated feedback was analysed using linear mixed effects models[1]. The response variable 'Feedback Count' was log-transformed (Shapiro-Wilk test for normality:

[1] We used the lme4 R package, version 0.999375-39.

Table 2: Parameter estimates of the linear mixed effects model.

Fixed effects	**Estimate**	**Std. Err.**	***t*-value**	***p*-value**
(Intercept)	1.997	0.350	5.708	0.0000
Condition	0.221	0.077	2.850	0.0086
Duration	0.003	0.001	4.025	0.0005
Random effects				
Residual variance: Var = 0.15, Std.Dev.= 0.38				

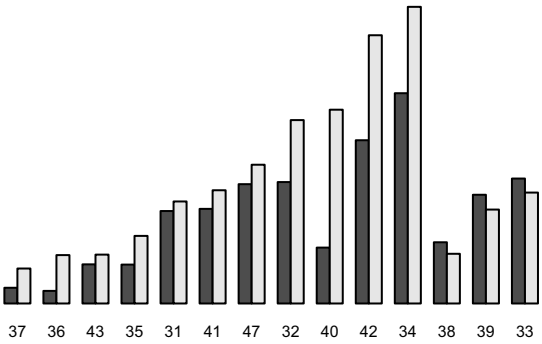

Figure 2: Time-normalised feedback rates for each session in the distracted (dark bars) and non-distracted (light bars) conditions. Numbers identify sessions.

$W = 0.96$, p-value $= 0.42$). 'Condition' (distracted vs. not distracted), 'Order' (of instruction) and 'Duration' (of each dialogue) as well as their interactions were entered in the model as fixed factors ('Duration' as a numerical fixed factor). Subject pair IDs were entered as a random factor. Non-significant interactions were eliminated and the model was refitted. Log-likelihood tests revealed that a simple model with two main effects of 'Condition' and 'Duration' ('Order' did not reach significance) without interactions and with the random effect of Session provides the best fit to our data. We report parameter estimates in Table 2. p-values were calculated by means of Markov Chain Monte Carlo (MCMC) sampling. The effect of dialogue duration points to the need of time normalisation in subsequent analyses.

The analysis shows that there is a significant effect of Condition on the number of produced feedback. Figure 2, showing time-normalised feedback rates for each session in the distracted and non-distracted condition, confirms this: more feedback is produced in the non-distracted condition in all but three sessions. This indicates that our experimental design worked as expected.

4.2. Distribution of positive feedback across conditions

We now turn to the analysis of individual positive feedback categories (P1–P3, A) across conditions. The results are presented in Figure 3. It shows the number of occurrences of each function in either condition normalised by the dialogue duration and divided into two groups depending on whether more feedback was produced in the distracted or non-distracted condition. Sessions plotted on the left of the dotted lines had higher feedback rates in the non-distracted condition, those on the right in the distracted condition.

The proportions of P2 show that the distracted condition caused the listeners to signal understanding less frequently. We interpret the low P2 frequencies as a sign of decreased attention in the listener. It becomes clear when we consider that using P2

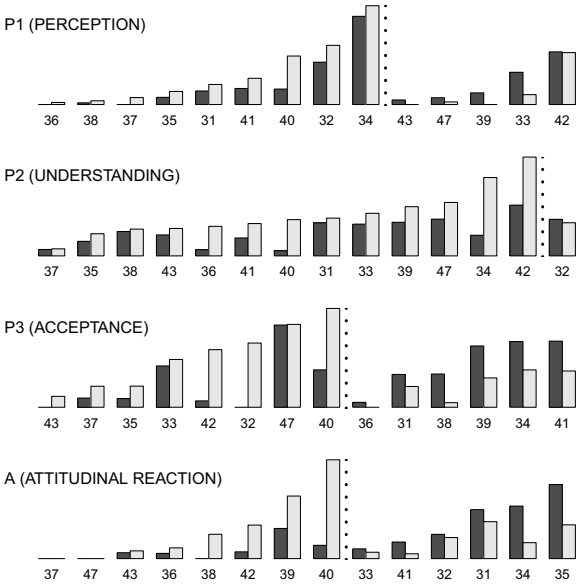

Figure 3: Time-normalised counts of functions P1–P3 and A for each session in the distracted (dark bars) and non-distracted (light bars) conditions. For each function the dotted line separates sessions with more feedback in the non-distracted condition from those with more feedback in the distracted condition. Numbers identify sessions.

evidences a high degree of genuine understanding (reconstructing the information structure) and fine attention to deep semantic features. It cannot be merely a reaction to prosodic features or specific keywords (e.g., 'New York', 'late' or 'cancelled' in the context of holiday stories). P2 is, therefore, a *response* rather than *reactive feedback* in the sense of [4].

Alternatively, one could adopt an explanation based on intentional behaviour. Subjects might not want to cause confusion by falsely signalling close comprehension of immediate details since acting so might reveal their distractedness. In fact, minimal social constraints on politeness and cooperation require they feign attentiveness. These constraints are hard to be met in P2. In fact, some distracted listeners prefer to choose feedback signals that only involve 'shallow evaluation' (P1 or A). Not only are these easier to use correctly (or to feign successfully) but they also adequately fulfil social constraints. Intuitively, dialogue partners who use many backchannels and emotional responses are considered 'good listeners' and more likely to establish rapport [6].

While it might be difficult to produce P3 feedback based on shallow cues only (particularly when no attitudinal or emotional component is present), acceptance, belief and agreement might be nonetheless safer to feign than understanding since speakers should normally have no grounds to doubt the sincerity of these acts and it is unlikely that their true character should be exposed in the course of conversation. It seems, therefore, that the larger number of higher-level feedback, such as P3, in the distracted condition does not rule out distractedness (as in [1]).

Interestingly, listeners who produced more P1 feedback in the distracted condition also tended to produce less P3 feedback when distracted, and *vice-versa*. This suggests that different listeners might use different compensatory strategies when distracted. While some rely on low-level P1 feedback, others turn

to signalling the higher-level P3 functions. However, almost all of them signal less understanding.

It should be also pointed out that some of the P3 feedback might in fact include borderline P1 cases. Annotators indeed reported difficulty distinguishing between P1 and P3 categories in some cases. This indicates that P1 and P3 are not as distant as might initially appear and might at least partly explain the similarities between the distributions of P1 and P3.

Nonetheless, being a sufficiently good listener does not require to display understanding (in the sense of P2 here) when P1 and P3 are used convincingly. Their increased use might, in fact, even reassure speakers that they should continue and that the listener already accepts or believes the message. In such cases understanding seems implied which reduces the pressure on the speaker to elaborate. The lower number of P2 can, therefore, be interpreted as a sign of cooperative action on the part of the distracted listener, who in this way tries to avoid disrupting the flow of conversation.

5. Conclusions and future work

The results reported here show that speakers produce less feedback when distracted by a simultaneous task. More importantly, we identified the reduced rate of signalling understanding as a consistent and predictable cue of distractedness in the listener. We proposed two explanations of this finding: one in terms of feedback-inviting cues, and one in terms of dialogue strategies aimed at concealing decreased attention. We intend to pursue these questions with a more fine-grained analysis of listener feedback (especially the P1 and P3 categories) in future work. We are also planning to include speakers' behaviour and its influence on listeners in the analysis.

Additionally, these findings and the collected corpus open the way for further research and applications in various domains of interest to us. On the one hand, they add to engagement-related research done in the context of dialogue systems and human-computer interaction. They provide new cues to attentiveness in addition to those already identified: gaze direction [7], spatiotemporal [8] and prosodic features [9], low-level emotional state recognition [10], and the influence of engagement gestures in human-robot communication [11]. A comparison of multimodal, lexical and prosodic features of feedback in our study coupled with precise information about listener's task performance over time (by aligning listeners' button presses with speakers' 's-words') should provide more accurate correlates of distractedness in the signal. These efforts will help gain a better understanding of the role of engagement in communication.

On the other hand, in order to be able to use our findings in real world applications it is necessary to establish objective criteria for identifying functions of feedback expression. [5] and [12] ran a series of classification tasks of English cue words but their annotation scheme lacks the crucial category for signalling understanding, which corresponds to our category P2. We intend to address this problem in the future, this time using the whole corpus rather than a selection of sessions, and including an analysis of the nonverbal behaviour of dialogue partners (such as head movements, facial expressions and gaze).

Acknowledgements – This research is supported by the Deutsche Forschungsgemeinschaft (DFG) at the Center of Excellence in 'Cognitive Interaction Technology' (CITEC) as well as at the Collaborative Research Center 673 'Alignment in Communication'. We would also like to thank Robert Eickhaus for technical support in recording and editing the audio/video material.

6. References

[1] J. B. Bavelas, L. Coates, and T. Johnson, "Listeners as co-narrators," *Journal of Personality and Social Psychology*, vol. 79, pp. 941–952, 2000.

[2] A. K. Kuhlen and S. E. Brennan, "Anticipating distracted addressees: How speakers' expectations and addressees' feedback influence storytelling," *Discourse Processes*, vol. 47, pp. 567–587, 2010.

[3] J. Allwood, J. Nivre, and E. Ahlsén, "On the semantics and pragmatics of linguistic feedback," *Journal of Semantics*, vol. 9, pp. 1–26, 1992.

[4] S. Kopp, J. Allwood, K. Grammar, E. Ahlsén, and T. Stocksmeier, "Modeling embodied feedback with virtual humans," in *Modeling Communication with Robots and Virtual Humans*, I. Wachsmuth and G. Knoblich, Eds. Berlin: Springer-Verlag, 2008, pp. 18–37.

[5] A. Gravano, S. Benus, J. Hirschberg, S. Mitchell, and I. Vovsha, "Classification of discourse functions of affirmative words in spoken dialogue," in *Proceedings of Interspeech 2007*, Antwerp, Belgium, 2007, pp. 1613–1616.

[6] L. Tickle-Degnen and R. Rosenthal, "The nature of rapport and its nonverbal correlates," *Psychological Inquiry*, vol. 1, pp. 285–293, 1990.

[7] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi, "A model of attention and interest using gaze behavior," in *Proceedings of the 5th International Working Conference on Intelligent Virtual Agents*, Kos, Greece, 2005, pp. 229–240.

[8] D. Bohus and E. Horvitz, "Models for multiparty engagement in open-world dialog," in *Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue*, London, UK, 2009, pp. 225–234.

[9] T. Kawahara, Z.-Q. Chang, and K. Takanashi, "Analysis of prosodic features of Japanese reactive tokens in poster conversations," in *Speech Prosody 2010*, Chicago, IL, 2010, pp. 1–4.

[10] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," in *Proceedings of Interspeech 2004*, Jeju Island, Korea, 2004, pp. 1329–1332.

[11] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, pp. 140–164, 2005.

[12] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 2007, pp. 1065–1068.