Feedback and activity in dialogue: signals or symptoms?

Andrew Gargett

Linguistics Department, UAE University, Al Ain, UAE

andrew.gargett@uaeu.ac.ae

Abstract

This paper presents new approaches to modelling both linguistic and non-linguistic feedback during instruction giving in a virtual domain. Our approach enables fine-grained investigation of how language and actions are conditioned by task-level and domain-level features of dialogue. In a preliminary study, we examine the interaction between pauses in linguistic and non-linguistic activity. As far as we know, ours is the first analysis of pauses across modalities. In the longer term, we aim to use these techniques as a window on the underlying processes conditioning feedback, and for such applications as the generation of situated forms of listening, such as instruction following.

Index Terms: feedback, linguistic and actional pauses, virtual worlds

1. Introduction

During instruction giving, backchanneling is a normal part of feedback behaviour. However, an instruction giver may well take all kinds of instruction following behaviour as feedback, e.g. should an instruction giver consider an instruction follower who has stopped talking or moving to be indicating understanding or lack of understanding? For their part, an instruction follower can deliberately signal problems arising from lack of understanding, by stopping when faced with difficulties, or even waiting for further clarification.

Given that such silence and inaction is ubiquitous in everyday conversation (e.g. [5]), then clearly instruction givers need to be very good at detecting and dealing with such evidence about instruction follower behaviour. Instruction followers for their part should impart the correct signals when necessary. However, such phenomena as silence and inaction lack content (quite literaly), which raises the question: how do we in fact construe meaning of pauses in **both** action and language?[1] And given the apparent involvement of pausing in feedback behaviour, how does such lack of action or language affect other feedback channels?

The long-term aim of our work is to answer such questions by developing a method for gathering fine-grained information about interactive language behaviour in multi-modal settings. To this end, in this paper we present some preliminary work on a multi-modal corpus, the SCARE corpus ([3]), to see if, using this method, we are able to discover interesting interactions between pausing in action and language and other forms of feedback, specifically, backchannels. We hope that by examining the interaction between such cross-modal phenomena, insights can be gained into the "meaning" of such interactional phenomena.

2. Previous work

The literature on pausing behaviour is well-established, but typically does not consider actional alongside linguistic pausing. Pauses involve unusually lowered levels of activity in the production systems of a single language user, in our case lowered activity levels in language and/or actions. How low such levels must go for a pause (to be perceived) to have occurred is a difficult question, and some useful progress has been made toward answering this (e.g. [1]). Pauses have recently been of some interest in research on interaction. Heldner and Edlund ([5]), in particular, provide a thorough typology from different speaker's perspectives, echoing [8], that while gaps are between-speaker silences, pauses are within-speaker silences. We adopt this definition, and extend it to define pauses in actions to be within-actor inactivity, and gaps in action to be between-actor inactivity.

In line with a growing body of research, we take it that pauses provide a window on language production and cognitive processing (e.g. [10]). However, such work has until now been largely linguistically oriented. We seek to extend this to other production systems, in a way which is in line with previous literature on linguistic pauses within interaction (e.g. [5]).

Now, numerous established corpora of instruction giving dialogues (e.g. TRAINS[2]) are strictly text based. This has led to a paucity of information regarding the use of situational features made by interlocutors. Extending models of interaction to incorporate such information may provide qualitatively distinct accounts of what is going on in dialogue. In this paper, we will offer a preliminary proof-of-concept study, suggesting the usefulness of

[1] While pauses lack referential content, there is certainly some sense that can be made of them, albeit wholly gained from the context.

[2] http://www.cs.rochester.edu/research/speech/trains.html

such information.

Corpus collections of multimodal dialogues, like the SCARE[3] ([3]) and GIVE-2[4] ([4]) corpora, are crucial for approaches of the kind we are proposing. The availability of such corpora provide an opportunity to make fine-grained investigations of the situational features conditioning interaction.

Our approach is as far as we know the first to empirically model the interaction of different forms of feedback across modalities. This enables us to make uniquely cross-modal comparisons of dialogue phenomena.

3. Method

3.1. Data

We used data from the SCARE corpus ([3]), a collection of instruction giving dialogues in a virtual world (created using QuakeII gaming software) made up of two levels, each with between 7 and 9 rooms, and these rooms having buttons for opening cabinets that contained objects to be retrieved (see Figure (1) for screenshots). The corpus consists of 15 sessions, with interlocutors taking roles of either instruction giver (IG) or instruction follower (IF). They had to complete a series of 5 simple tasks (retrieving objects), with the IG verbally guiding the IF through the world, but only the IG having access to a map of the world, and a list of the tasks to be completed. The 19 male and 11 female participants had an average age of 30, and identified as native speakers of North American English. Sessions ranged from 10 minutes in length to over half an hour.

From this corpus, we collected from 12 of the 15 SCARE sessions, objective correlates of pauses in actions (i.e. complete cessation of physical activity),[5] while for pauses in language we relied on the judgements made by the original annotators of the linguistic aspects of the SCARE corpus. Accompanying the SCARE corpus were detailed recordings of information about the fifteen game sessions, including position and orientation of the IF, as well as locations of objects in the SCARE world, such as buttons, cabinets and doors.[6] From the data streams recorded in these log files, information about events could be extracted, such as whether the instruction follower was or was not moving or turning - we took pauses in actions to be those periods between when the follower was turning and/or moving, as well as the context of such activity or inactivity. Note that due to the way the SCARE corpus is recorded, only the instruction follower both moves and talks, while the instruction giver simply talks.

We developed a Scala tool for re-building the SCARE corpus as a stand-off corpus using the Nite NXT toolkit ([7]). The Nite NXT approach is particularly useful for us due to its rich structuring of data, including a data set model for structuring a corpus in terms of (i) observations, (ii) agents, (iii) the interaction, as well as (iv) the signal. In particular, the observations can be multi-layered, either directly aligned to the timing level, or else symbolically linked to other levels (e.g. annotations of dialogue acts can be linked to actual utterances, which in turn can be directly aligned with the timing of the original audio and video signal). Aside from allowing us to adequately model the information contained in the SCARE dialogues, this also allowed access to a very useful library of Java classes bundled with the Toolkit (e.g. for searching NXT-formatted corpus files).

For backchannels, we examined acknowledgements like "ok" or "yeah", tacit agreements like "mhm", and fuller expressions of agreement like "yep" or "alright", as well as interjections "um" and "uh" (e.g. [9]). We will not consider here the role of "gaps", as defined by [5] (although, it would be interesting to pursue this further in the future).

3.2. Procedure

As a proof-of-concept of our approach to modelling multimodal elements of feedback in interaction, we carried out the four studies reported in Table (3.2), comparing instruction giver (IG) and instruction follower (IF) behaviour.

In each case, we attempt to determine whether such forms of feedback, in both instruction givers and followers, are independent or not of the activity of the instruction follower. In the context of instruction giving, the inactivity of the instruction follower is a direct indication of trouble in completing the task. For the instruction giver, given the need to finish the task as quickly as possible, the instruction follower's lack of movement is likely a symptom of misunderstanding (where inaction=inability to act), while, the instruction follower could well intentionally signal their lack of understanding in this way (cf. "Sorry, what was that?").

3.3. Results

3.3.1. Studies 1 & 2

Table (2) reports the percentage of backchannel tokens as used by speakers in each role, interjections being reserved for the fourth study (see Section (3.3.3) below).[7] We carried out a Pearson chi-squared test on this data, with result $\chi^2 = 106.5$ ($p < 0.01, df = 3$), suggesting that the use of backchannels by Instruction Giver and Instruction Follower do not have the same distribution, in other

[3] http://slate.cse.ohio-state.edu/quake-corpora/scare/

[4] www.give-challenge.org/research/page.php?id=give-2-corpus

[5] Ignoring sessions 1, 5, and 15, which present various problems for such data collection.

[6] Thanks to Alexander Koller and Krystof Drys for making available code, some of which was adapted in the data retrieval process. The use we made of this modified code is of course our own responsibility.

[7] Recall, IG=instruction giver, IF=instruction follower.


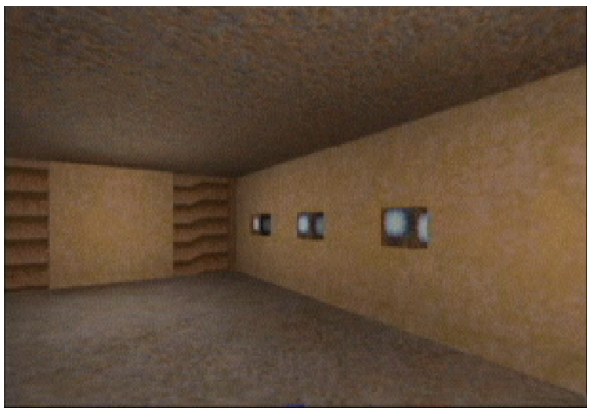

Figure 1: Screenshots of rooms within the SCARE world

Study question	**Purpose**
Study 1: In the context of pause in actions, what IG backchannels are most likely?	Evaluating overlap of backchannels of instruction givers with the activity of instruction followers
Study 2: In the context of pause in actions, what IF backchannels are most likely?	Evaluating overlap of backchannels of instruction followers with their own activity
Study 3: In the context of pause in actions **and** words, what is the likely backchannel to be used, and who is most likely to use it?	Evaluating overlap of backchannels and cessation of **both** actions and language in instruction givers and followers
Study 4: In the context of pause in actions **and** words, how does the choice between backchannels "um" vs. "uh" affect word pause duration?	Evaluating overlap of interjections with cessation of **both** actions and language

Table 1: Proof-of-concept studies

words that the use of backchannels is not independent of the role of speaker. Note that Table (2) also reports standardised residuals, which is useful in contrasting how the actual patterns of use of backchannel tokens differs from one where use of such tokens would be independent of the role of speaker (i.e. the null hypothesis case).

Token	IG	IF	Row totals
alright	77(.96)	30(-1.3)	107
mhm	4(-5.9)	61(7.9)	65
ok	381(.63)	191(-.84)	572
yeah	165(1.5)	63(-2.0)	228
Column totals	627	345	972

Table 2: Backchannels in the context of IF inactivity (including standardised residuals)

3.3.2. Study 3

Two investigations were conducted here, one that examined what happened in the context of linguistic pauses by each speaker preceding action pauses, and the other in the context of linguistic pauses following action pauses. It turns out that only tokens for "ok" and "yeah" with enough frequency for results to be significant. The comparisons in Table (3) yield results of far less significance, with $\chi^2 = 1.67$ ($p = .80$, $df = 4$). The data in Table (4), is similar, with $\chi^2 = 2.05$ ($p = .73, df = 4$). In neither case, can independence of use of tokens from speaker role be refuted.

Token	IG	IF	Row totals
ok	9(-.55)	22(.40)	31
yeah	6(.89)	6(-.65)	24
Column totals	30	56	172

Table 3: Use of backchannel tokens by Instruction Giver vs. Instruction Follower (word pause preceding action pause, including standardised residuals)

Token	IG	IF	Row totals
ok	14(-.34)	23(.29)	37
yeah	3(1.0)	1(-.88)	4
Column totals	17	24	82

Table 4: Use of backchannel tokens by Instruction Giver vs. Instruction Follower (action pause preceding word pause, including standardised residuals)

3.3.3. Study 4

Our reason for focusing here on the interjections "um" vs. "uh" is that, while their interaction with linguistic pauses is established ([11]), "um" projecting a longer pause in words than "uh", it would be interesting to consider both in the context of actional pauses. Indeed, we were able to confirm the distinction in projected word duration for those interjections overall, with two sample t-test results $t(147) = 3.49, p < .01$, indicating the null hypothesis of no distinction between word duration after different interjections can be rejected (word duration after "um" having $M = .56(SD = .68)$ $seconds$, "uh" having $M = .30(SD = .39)$ $seconds$). However, testing the distinction between subsequent word duration following these interjections, that occur during action pauses, we obtained two sample t-test results $t(28) = 1.76, p = .08$, indicating the same null hypothesis cannot be rejected in the context of action pauses (word duration after "um" having $M = .78(SD = .76)$ $seconds$, "uh" having $M = .48(SD = .55)$ $seconds$).[8]

4. Summary and future work

This paper reports results of what are essentially proof-of-concept studies, presenting a novel consideration of feedback across modalities, and thereby demonstrating the viability of our method for investigating situated dialogue. Our results show, first, that in the context of inactivity by an instruction follower, a range of forms of feedback become available for use, and that indeed the use of backchannels is dependent on role: instruction givers are far more likely to use "ok" in such contexts than instruction followers, while instruction followers are far more likely to say "mhm". However, during complete silence and inactivity, the use of specific backchannels becomes independent of speaker role. Finally, employing our approach, a tentative initial analysis suggests that an established distinction between backchannels "um" and "uh" may vanish in the context of action pauses.

For future work, we will broaden our investigation into this corpus, outside of the narrow range of pauses in instruction follower activity. For example, an important factor in the use of backchannels which we are planning to look at is their relationship with intonation contour ([12]), but also in the context of instruction follower inactivity. Further, given the highly adaptable means whereby vitual domains can be installed on mobile devices such as laptops, we are currently planning an Arabic version of the SCARE corpus with the aim of cross-linguistic investigation across modalities of the kind of phenomena explored by [12] and others. Finally, a key aim of our work is to develop from such studies, more natural and effective generation of listening behaviour on the part of artificial instruction following agents.

[8] However, a mixed effects account of this data is possible, with action pauses as fixed factors, and subjects and interjections as random factors. Preliminary ANOVA on by-subject vs. by-item means of word duration, suggests a significant effect for action pauses for the by-subjects analysis (F(1,26)=6.27, p<.05), but not for the by-items analysis. Typically, significance in both analyses is required to show overall significance; we are further examining this line of inquiry.

5. Acknowledgements

Many thanks to the "Feedback behaviours in dialogue" workshop organisers for their efforts. Background to this paper is ongoing research with Magda Wolska (Saarland University), on using virtual worlds to investigate dialogue. Of course, all errors, etc, here remain my own.

6. References

[1] Anna Danielewicz-Betz, "Silence and pauses in discourse and music", PhD., School of Philosophy, Saarland University, 1998.

[2] Patrick Ye, "Natural language understanding in controlled virtual environments", PhD. Department of Computer Science and Software Engineering, The University of Melbourne, 2009.

[3] Laura Stoia, Darla Magdalene Shockley, Donna K. Byron and Eric Fosler-Lussier, "SCARE: A Situated Corpus with Annotated Referring Expressions", Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), 2008.

[4] Andrew Gargett, Konstantina Garoufi, Alexander Koller and Kristina Striegnitz, "The GIVE-2 Corpus of Giving Instructions in Virtual Environments", Proceedings of the 7th International Language Resources and Evaluation Conference (LREC 2010), 2010.

[5] Mattias Heldner and Jens Edlund, "Pauses, gaps and overlaps in conversations", Journal of Phonetics, 28:555–568, 2010.

[6] Brigitte Zellner, "Pauses and the Temporal Structure of Speech", in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition, Chichester: John Wiley,pp. 41–62, 1994.

[7] Jean Carletta, Stefan Evert, Jonathan Kilgour, Craig Nicol, Dennis Reidsma, Judy Robertson and Holger Voormann, "Documentation for the NITE XML Toolkit", Online: http://http://groups.inf.ed.ac.uk/nxt/documentation.shtml.

[8] Harvey Sacks, Emmanuel Schegloff and Gail Jefferson,"A simplest systematics for the organization of turn-taking for conversation", Language, 50:696–735, 1974.

[9] Stefan Benus, Agustn Gravano and Julia Hirschberg, "The prosody of backchannels in American English", In ICPhS, 2007.

[10] Louis ten Bosch, Nelleke Oostdijk & L. Boves, "On temporal aspects of turn taking in conversational dialogues", Speech Communication, 47:80–86, 2005.

[11] Herbert H. Clark and Jean E. Fox Tree, "Using *uh* and *um* in spontaneous speaking", Cognition, 84:73–111, 2002.

[12] Nigel G. Ward, Rafael Escalante, Yaffa Al Bayyari and Thamar Solorio, "Learning to Show You're Listening", Computer Assisted Language Learning 20:385–407, 2007.