Push-to-talk ain't always bad! Comparing Different Interactivity Settings in Task-oriented Dialogue

Raquel Fernández, David Schlangen and **Tatjana Lucht**
Department of Linguistics
University of Potsdam, Germany
{raquel,das,lucht}@ling.uni-potsdam.de

Abstract

Restrictions of interactivity in dialogue are often seen as having negative impact on the efficiency of the dialogue, as they affect the ability to give immediate feedback (Whittaker, 2003). We have conducted experiments with one such restriction common in spoken dialogue systems, namely *push-to-talk*. While our results confirm many predictions from the literature (fewer but longer turns; reduction of positive feedback), we found no significant impact on task-efficiency. Our analysis of the grounding strategies of the subjects shows that the restriction actually induced a more cautious strategy that proved advantageous for our matching task, and that giving *negative* feedback in the form of clarification requests was not affected by the restriction.

1 Introduction

Natural, freely regulated turn-taking as described for example in the seminal paper (Sacks et al., 1974) is still a long way off for spoken dialogue systems. Unable to interpret in real-time the various information sources that have been investigated as influencing turn-taking (see e.g. (Caspers, 2003) on the role of syntax and prosody in Dutch turn-taking), dialogue systems resort to simpler strategies like using *time-outs* (where a silence by the user is interpreted as the intention to yield the turn) and *push-to-talk*, where the turn is held explicitly by pushing a button when speaking (see e.g. (McTear, 2004) for a discussion of these methods).

In the work reported here, we wanted to investigate in isolation the effect of the latter strategy, *push-to-talk*, on the shape of task-oriented dialogue. For this we conducted an experiment where we let subjects do a conversational task (a variant of the matching tasks of (Krauss and Weinheimer, 1966; Clark and Wilkes-Gibbs, 1986) either with free turn-taking or with turn-taking controlled by *push-to-talk*. The theoretical literature makes clear predictions about such settings (fewer, longer turns with less efficient descriptions; see next section). While our findings confirm some of those, we found no negative impact on task success, which on further analysis seems due to a different grounding strategy induced by the restriction.

The remainder of the paper is structured as follows. In the next section, we briefly review some of the theoretical predictions of effects of interactivity restrictions. We then describe our task and the experimental conditions, procedure and method. In Section 5 we describe our analysis of the turn and dialogue act structure of the collected dialogues. The puzzling result that the restricted dialogues were not less efficient than the unrestricted ones is further analysed in Section 6 by looking at more global strategies used by the participants. We close by briefly discussing our results and possible further work that could be done to corroborate our findings.

2 Interactivity and the Shape of Dialogue

In pragmatics it is common to assume that conversation, like any other collaborative and interactive action, is governed by economy principles such as the Gricean maxims (Grice, 1975) or the more recently formulated *principle of least collaborative effort* (Clark and Wilkes-Gibbs, 1986). The latter states that participants will try to maximise the success of their collective purpose while minimising costs. As (Clark and Brennan, 1991) point out, the costs of communicative actions are dependent on features of the medium used, like copresence, visibility, audibility, cotemporality or simultaneity. For instance, using short feedback acts like “uhu”, which is effortless in face-to-face communication, becomes slightly more costly when communicating via email, while their cost is definitely much higher when communicating via non-electronic letters.

Mediums in which participants communicate by speaking (as opposed to for instance typing), receive messages in real time (cotemporality) and can communicate at once and simultaneously (simultaneity) afford full *interactivity* (Whittaker, 2003).

Interactivity plays a central role in theories of grounding like those of Clark and colleagues (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989). It enables speakers to interrupt and complete each other’s utterances and allows for constant feedback in the form of often concurrent backchannels, which help to determine whether the conversation is on track and facilitate quick repair of misunderstandings.

One of the predictions of these theories is that settings that preclude or restrict interactivity, like half-duplex channels, will disrupt understanding and quick repair and show less incremental content, thereby leading to more time and errors. This has been confirmed by several studies, like (Krauss and Weinheimer, 1966; Clark and Krych, 2004), that have investigated non-interactive settings that lack cotemporality and simultaneity. In these studies speakers, who are engaged in a referential communication task, talk to a tape recorder for *future* addressees. Interactivity is completely precluded and therefore speakers do not get any form of feedback. (Krauss and Weinheimer, 1966) found that speakers who do not get feedback from addresses take longer and make more elaborate references. Similarly, (Clark and Krych, 2004) showed that references designed without feedback are “inferior in quality” and some are even impossible to grasp.

The experiments we report here investigate the effects of restricting interactivity by using a half-duplex channel managed by *push-to-talk*, which allows cotemporality but inhibits simultaneity. As will be seen in subsequent sections, our results confirm many predictions from the literature, like the presence of fewer but longer turns and a significant reduction of positive feedback (as observed in other studies that used half-duplex channels like e.g. that of (Krauss and Bricker, 1967)). Surprisingly, however, we found that this did not lead to any significant impact on task-efficiency (Fernández et al., 2006). One of the aims of the present paper is to shed some light on the reasons behind this puzzle.

3 Task and Experimental Setting

The task we have asked our experimental subjects to do is a variant of the reference tasks pioneered by (Krauss and Weinheimer, 1964; Krauss and Weinheimer, 1966). In our task, a *player* instructs an *executor* on how to build up a *Pentomino* puzzle (see below). The player has the full solution of the puzzle, while the executor is given the puzzle outline and the set of loose pieces. The solution and the outline of the puzzle are shown in Figure 1.

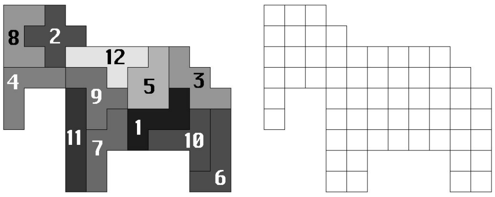

Figure 1: Solution and Outline

The player is asked to tell the executor how the puzzle is assembled following a particular order of the pieces, as given by the numbers on the solution in Figure 1. This enforces a reconstruction process common to all collected dialogues, which allows for more systematic comparisons. The pieces that the executor manipulates are not numbered and are all the same colour. Both player and executor were aware of the information available to each other.

During the experiment the player and the executor were in different rooms and communication between them was only verbal. They could not see each other and they did not have any visual information about the state of the task (i.e. the player could not visually monitor the progression of the reconstruction process).

We investigate two different conditions that differ in degree of interactivity. In a first fully interactive condition, player and executor communicate by means of headsets and the channel is continuously open, as it would be for instance in a telephone conversation. In the second condition interactivity is restricted. Here subjects communicate using walkie-talkies that only offer a half-duplex channel that precludes simultaneous communication. Speakers have to press a button in order to get the turn, hold it to keep it, and release it again to yield it (a 'beep' is heard when the other party yields the turn). We refer to these two conditions as *free turn-taking* (FTT) and *push-to-talk* (PTT), respectively.

4 Procedure and Methods

The experiments involved 20 subjects, 11 females and 9 males, grouped in 10 player-executor pairs. Five pairs of subjects were assigned to each of the two conditions: two female-female pairs, one male-male pair, and two female-male pairs used FTT, while two female-female pairs, two male-male pairs, and one female-male pair used PTT. All subjects were German native speakers between 20 and 45 years old, and the conversations were in German.

The 10 dialogues collected make up a total of 194.54 minutes of recorded conversation. The recordings were transcribed and segmented using the free software Praat (Boersma, 2001). The transcribed corpus contains a total of 2,262 turns and 28,969 words.

To keep a visual record of the progression of the task, the board with the outline and the pieces that the executor manipulated was videotaped during task execution. This gives us a corpus of 10 videos, which have been informally analysed but not systematically annotated yet.

5 Analysis 1: Turn & Act Structure

5.1 Coding

We used MMAX2 (Müller and Strube, 2001) to annotate each utterance with one or more dialogue acts (DAs). We distinguish between task and grounding acts. Task acts are further classified into task-execution (including a tag for description acts where a piece or a location are described) and task-management acts, while grounding acts include different types of feedback acts, as well as clarification requests (CRs). Table 1 shows an overview of the DA taxonomy used.

DA Tag	Meaning
`Task`	
⊢ `Task-Execution`	
`descr_piece`	Description of piece
`descr_pos`	Description of position
`req_info`	Request of task-related info
`req_action`	Request for action
`sugg_error`	Suggest error in task
⊢ `Task Management`	
`dis_sett`	Discuss setting
`dis_stra`	Discuss strategy
`coor_task`	Coordinate task execution
`Grounding`	
⊢ `pos_fback`	Acknowledgement
⊢ `neg_fback`	Rejection or correction
⊢ `ask_conf`	Request for acknowledgement
⊢ `CR`	Clarification request
`Other`	Incomplete and other acts

Table 1: DA Taxonomy

5.2 Results

An analysis of turn patterns shows that our PTT dialogues contain roughly half as many turns as the FTT dialogues, with the turns however being on average twice as long as the FTT turns (in seconds: 7.21 sec and 3.71 sec on average respectively; this difference is statistically significant at $p < 0.01$; in number of words: 20.2 vs 11.3 on average; $p < 0.05$).[1]

Figure 2 plots the number of turns per dialogue in each condition and for each participant role. The diagram allows us to see that the number of turns is rather constant across PTT dialogues, with equal number of contributions by player and executor. This indicates that in this condition player and executor do indeed take turns; i.e. each contribution by one is followed by one by the other. In the FTT dialogues there is a higher variation among pairs of participants and the number of turns contributed by the executor is higher. This in turn indicates that often executors' contribu-

[1] Unless otherwise stated, all significances reported in this paper are calculated with a t-test.

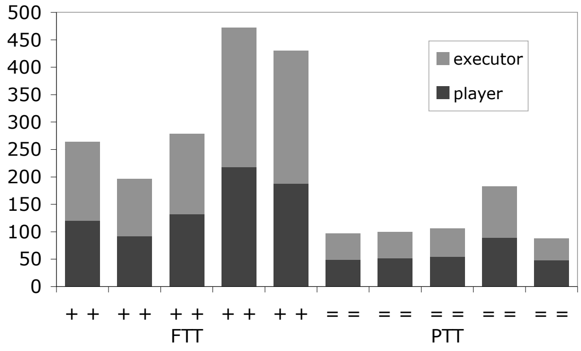

Figure 2: Number of turns per dialogue

tions are provided concurrently to those of the player. On average, around 35% of FTT turns are given in complete overlap; even when these turns are not counted, the number of turns in FTT is significantly higher ($p < 0.02$).

Despite the differences in turn patterns, pairs of participants in both conditions were able to finish the task in roughly the same time (18.7 min in PTT and 19.8 min in FTT on average; no significant difference). However, pairs in the PTT condition were able to do so using significantly fewer words (2253.6 vs 3540 on average; $p < 0.05$). Table 2 shows the mean number of words per condition and speaker role. As is common in this kind of instructional tasks (e.g. (Clark and Krych, 2004)), instruction givers (players) talk markedly more than instruction followers (executors).

	FTT	PTT
player	2127	1551.2
executor	1413.2	702.4

Table 2: Mean num of words per dialogue

The distribution and length of dialogue acts also helps to highlight some further differences between conditions. Distribution is shown in Table 3. The most significant difference regarding distribution is found in the amount of positive feedback acts, like backchannels and acknowledgements, which is consistently higher in FTT (33.8% vs 25.7% on average; $p < 0.01$ on a χ^2 test on raw numbers). This is still the case when ovelapping turns are not taken into account. The distribution of other grounding acts like negative feedback and CRs, however, is similar in both conditions. As for task acts, PTT dialogues contain a higher proportion of task-related acts than FTT dialogues (45.4% vs 36.7% on average; $p < 0.01$ on a χ^2 test on raw numbers).

	FTT	PTT
task_related	871 (36.7%)	444 (45.4%)
pos_fback	804 (33.8%)	250 (25.7%)
other fback	211 (8.9%)	70 (7.1%)
CRs	361 (15.2%)	161 (16.5%)
other acts	127 (5.4%)	52 (5.3%)

Table 3: Distribution of DAs

The diagram in Figure 3 shows the mean length in words of the four main DA types for each of the two conditions. As can be seen, the length in words of positive and negative feedback acts is roughly the same in PTT and FTT dialogues. CRs tend to contain more words in PTT, although this is not statistically significant. Finally, description acts (which are the lion's share of task acts) contain significantly more words in PTT dialogues than in FTT dialogues (19.8 vs 14.2 on average; $p = 0.05$).

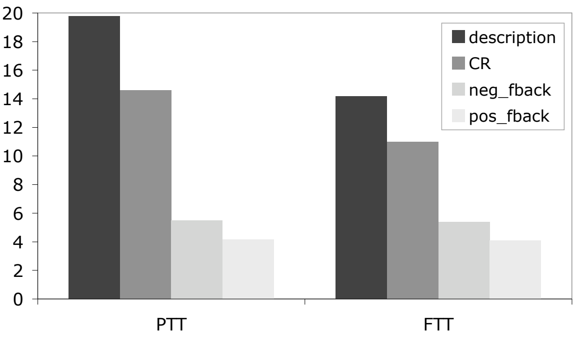

Figure 3: Mean num of words per DA type

5.3 Discussion

Our results confirm the predictions from the literature (e.g. (Krauss and Bricker, 1967; Whittaker, 2003)) that using a unidirectional channel produces less speaker switching and longer turns. We have also seen that description acts contain significantly more words in the PTT condition, which confirms the observation that contributions in non-interactive conditions tend to be more elaborate.

In Section 2, we pointed out that the lack of concurrent bidirectional communication is predicted to disrupt grounding behaviour leading to less shared understanding, which should have negative consequences at the task level. The analysis of dialogue acts has shown that grounding behaviour is certainly disrupted in the PTT condition. Although grounding acts do not vary in number of words across conditions, PTT dialogues show a significant reduction of the amount of positive feedback acts. This is presumably because positive feedback acts like acknowledgements, being very short acts and hence having a relatively high speaker-change overhead, are too costly in this condition. Interestingly, however, the proportion of other grounding acts like negative feedback acts and CRs (that also tend to be shorter) is not affected by the restriction. It seems that for our subjects, giving negative feedback was more essential, while positive feedback could presumably be taken as the default in a condition that made it coslty.

More surprising is perhaps the fact that the restricted interactivity of the PTT condition, with its lack of concurrent turns and its reduced positive feedback, did not lead to overall longer dialogues. Not only were pairs in the PTT condition not slower, but they were able to solve the task using significantly fewer words (see Table 2).

These observations pose a puzzle: Why does the reduction of interactivity in PTT dialogues not have a negative effect in terms of task efficiency (measured w.r.t. length of dialogue and number of words used)? To find an answer to this question, in the next section we analyse the dialogues on a level higher than individual acts, that of task-related moves.

6 Analysis 2: Task & Move Structure

6.1 Coding

The task of reconstructing the Pentomino puzzle can be divided into 12 *moves* or cycles, one for each of the pieces of the puzzle. A *move* as defined here covers all speech that deals with a particular piece, from the point when the player starts to describe the piece (*"Okay, so the next piece looks like a stair case"*) to the point when participants have agreed on the piece and its target location to their satisfaction and move on to the next piece. Sometimes moves are not successful and contain errors that are discovered later on in the dialogue. We call any stretch of speech that deals with the repair of a previous move that had already been closed a *repair sequence*.

Each dialogue contains 12 moves, while the number of repair sequences varies depending on the amount of errors and the uncertainty with which previous moves were grounded.

The video recordings of the board during task execution allow us to determine the grounding status of moves. By looking at the state of the board when a move is considered closed, we can determine whether the move has been successfully grounded or else whether there is a mismatch in common ground.

Using this visual information, we classified moves according to four categories: `correct`, `correct_rep`, `incorrect_inf` and `incorrect_rep`. Moves classified as `correct` were successful moves that did not require any subsequent repair nor double checking. Moves classified as `correct_rep` were successful but were grounded with low confidence and therefore required a repair sequence to confirm their correctness (usually after encountering problems with subsequent pieces). Moves classified as `incorrect_inf` were not successful but problems were discovered by inference by the executor after dealing with other pieces and the repair did not trigger an explicit repair sequence. Finally, moves classified as `incorrect_rep` were not successful and a repair sequence was performed at a later point in the dialogue to deal with the mismatch and repair the problems.

6.2 Results

The diagram in Figure 4 illustrates task progression with respect to the grounding success of the 12 moves (left to right) for each of the 5 dialogues in each of the two conditions.

We can compute a global *error score* for each dialogue by assigning values from 3 to 0 to moves classified as `incorrect_rep`, `incorrect_inf`, `correct_rep` and `correct`, respectively. The score of a dialogue is then the sum of the values obtained in each of the 12 moves, on a scale from 0 to 36. For instance, the top PTT dialogue in Figure 4 has an error score of 3, while the error score of the top FTT dialogue is 7.

In general PTT dialogues obtain lower error

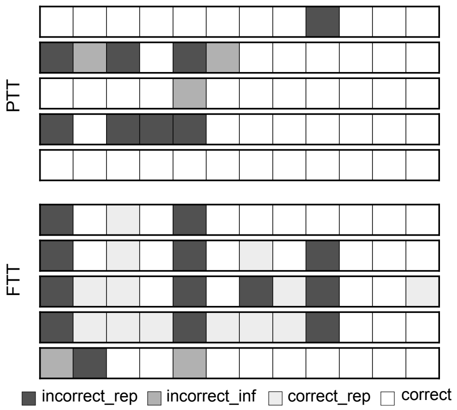

Figure 4: Move success

scores than FTT dialogues (5.8 vs 11.2 on average), although the difference is not statistically significant. This is probably not surprising given that in fact all pairs were able to finish the task successfully in roughly the same time. We find, however, that there is a correlation between error score and number of words in description acts per move (Pearson's correlation coefficient: $r = -0.7, p < 0.05$).

Further contrasts can be identified when looking at error score *per move*. The chart in Figure 5 plots the error score accumulated at each move for each of the two conditions. The score of a move within a condition is computed by adding the scores obtained in each of the five dialogues in that condition.

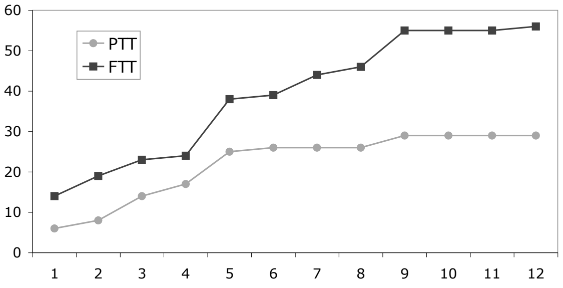

Figure 5: Error score per move

The chart allows us to see that after move 6 PTT pairs hardly make any more mistakes (the error score stays fairly constant from then on to the end of the task). Pairs in the FTT condition, on the other hand, keep on accumulating errors well until move 9. If we look at the amount of time spent on the first 6 moves in each dialogue, we see that, regardless of condition, the percentage of time spent on the first part of the task (up to the end of move 6) correlates with the global error score assigned to each dialogue (Pearson's correlation coefficient: $r = -0.69, p < 0.05$). For instance, the last PTT dialogue in Figure 4, which has an error score of 0, spends more than 8 minutes on the first 6 moves, while the third FTT dialogue, whose error score is 16, deals with the first 6 moves in less than 3 minutes. That is, more time leads to fewer errors.

6.3 Discussion

The analysis of task and move structure shows that, independently of conditions, a strategy whereby more time is spent on more detailed (=more words) descriptions making sure that moves are grounded before proceeding leads to fewer errors. The efficiency of PTT dialogues then can be explained by the fact that the restricted interactivity favours this kind of strategy. In Section 3 we showed that description acts contain a significantly higher number of words in PTT dialogues. Certainly, the fact that speakers can control the length of their turns allows for more detailed, perhaps better planned descriptions. Thus, what other studies of non-interactive settings have described as "overelaboration" (Krauss and Bricker, 1967) actually seems to be an advantage for the task at hand, which requires a fair amount of descriptive talent. The stricter control imposed by the turn-taking restriction on the interaction level leads to a stricter and better structured performance at the task level.

We have seen that subjects in FTT dialogues tend to make more mistakes further ahead in the task. This is in part due to a cascading effect whereby earlier errors lead to more subsequent mistakes. However even when errors are made, they can be recovered relatively fast (there is no correlation between length of dialogue and error score). The time that is not spent on detailed moves is then used in repair sequences.

As the lack of constant feedback makes quick repair more costly in PTT dialogues, subjects in this condition tend to adopt a more cautious strategy where moves are better grounded on a first pass and hence require fewer subsequent repair sequences, or use inference to avoid explicit repair.

7 Conclusions

We have presented the results of experiments that compare two different turn-taking conditions that differ in degree of interactivity: a fully interactive free turn-taking condition and a restricted condition where subjects use a half-duplex channel managed by push-to-talk.

Our results confirm many predictions from the literature, like the presence of fewer but longer turns and a reduction of positive feedback in the restricted condition. Indeed, participants do not produce short acts like positive feedback backchannels when conditions make them expensive; negative feedback acts and CRs however (also being shorter) are produced even under adverse conditions.

The literature also predicts that a reduction of interactivity will disrupt shared understanding and ultimately lead to problems at the task level. However, we found that the restricted condition did not have any significant impact on task-efficiency. Our analysis of the grounding strategies employed by the subjects shows that the restriction in interactivity actually favoured a more adequate strategy (longer and more detailed descriptions) that proved advantageous for our task—a difficult task that requires identification of very abstract referents.

More generally, our results indicate that dialogue participants do not always use the grounding strategy that is best for the task at hand, and that a particular grounding strategy can be "primed" by imposing turn-taking restrictions.

We are currently analysing in detail the form and evolution of the referring expressions used by the subjects with the aim to provide a more qualitative analysis of the differences between the two interactivity settings. In the future we also plan to experiment with other tasks in order to determine to what extent the consequences of reducing positive feedback are dependent on the task to be carried out.

Acknowledgements. We are thankful to J. Dreyer for letting us use the lab at ZAS, to M. Stede and A. Corradini for early discussions of the setting, and to three anonymous reviewers for helpful comments. This work was supported by the EU (Marie Curie Programme) and DFG (Emmy Noether Programme).

References

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10).

J. Caspers. 2003. Local speech melody as a limiting factor in the turn-taking system in dutch. *Journal of Phonetics*, 31:251–276.

H. Clark and S. Brennan. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, chapter 7, pages 127–149. APA Books, Washington.

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

H. Clark and E. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

H. Clark and D. Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

R. Fernández, T. Lucht, K. Rodríguez, and D. Schlangen. 2006. Interaction in task-oriented human-human dialogue: The effects of different turn-taking policies. In *Proceedings of the first International IEEE/ACL Workshop on Spoken Language Technology*, Palm Beach, Aruba.

H. P. Grice. 1975. Logic and converstion. In *Syntax and semantics, Volume 3: Speech acts*, pages 225–242. Seminar Press, New York.

R. Krauss and P. Bricker. 1967. Effects of transmission delay and access delay on the efficiency of verbal communication. *Jounrnal of the aCoustic Society of America*, 41:286–292.

R. Krauss and S. Weinheimer. 1964. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1:266–278.

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

M. F. McTear. 2004. *Spoken Dialogue Technology.* Springer Verlag, London, Berlin.

C. Müller and M. Strube. 2001. MMAX: A tool for the annotation of multi-modal corpora. In *Proceedings of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems.*

H. Sacks, E. A. Schegloff, and G. A. Jefferson. 1974. A simplest systematics for the organization of turn-taking in conversation. *Language*, 50:735–996.

S. Whittaker. 2003. Theories and methods in mediated communication. In *The Handbook of Discourse Processes*, pages 243–286. Lawrence Erlbaum Associates.