INTERACTION IN TASK-ORIENTED HUMAN-HUMAN DIALOGUE: THE EFFECTS OF DIFFERENT TURN-TAKING POLICIES

Raquel Fernández, Tatjana Lucht, Kepa Rodríguez, David Schlangen

Department of Linguistics
University of Potsdam, Germany

ABSTRACT

In human–human dialogue, the allocation of turns between the participants is normally managed smoothly, without the participants paying much attention to it. In contrast, for spoken dialogue systems turn allocation is a difficult task, and often technical restrictions are introduced to simplify it. In this paper we investigate, by comparing two experimentally collected corpora of human–human task oriented dialogue, what the consequences are of imposing one particular kind of restriction, namely that of using a simplex channel managed by push-to-talk (PTT).

We found, as expected, a loss of interactivity in the PTT condition (fewer, longer turns; more silences), but surprisingly, no loss of efficiency; in fact, the subjects in the PTT condition were able to finish their task in roughly the same time, while using fewer words along the way. We analyse here the differences in the interaction patterns and the interplay of 'naturalness' and efficiency as relevant factors for practical system development.

Index Terms— Turn-taking, push-to-talk, interaction, task-oriented dialogue

1. INTRODUCTION

The following observation from the seminal 1974 paper on turn-taking [1] might seem trivial: *"[In any conversation,] speaker-change recurs, or at least occurs."* However, how exactly this speaker-change is managed by the participants is still the topic of an ever-growing body of literature.

What is clear, in any case, is that the participants in a conversation normally do not pay much attention to how they do this; and even in the (comparatively rare) cases where uncertainty about whose turn it is arose, recovery is normally swift [2]. In contrast, for spoken dialogue systems (SDSs), turn-taking management is a difficult task, as the information sources identified in the descriptive literature (e.g., prosody, predictions about syntactic structure, expectations about the content of the utterance) are not easily made available automatically in real-time. (But see [3, 4], *inter alia*, for work that begins to address these issues.)

There are some strategies that are commonly used in SDSs to simplify the process: for the decision of when to take the turn, systems either wait for pauses of a certain duration (see [4] for the problems of this strategy) or they impose a *push-to-talk* policy on the user, where the turn can be taken when the user gives the explicit signal (releases a button); to simplify the decision when to yield, systems either simply complete all planned utterances and only then yield, or they allow barge-in [5], i.e. active turn-taking of the user (either through voice or through initiating *push-to-talk*).

The immediate suspicion, however, is that such deviations from the way people naturally handle these tasks come at a cost, as expressed in the following quote from [6]: *"[Such unnaturalness] will ultimately interfere with the user's ability to focus on the problem [the system is supposed to help with] itself rather than on making the interaction [with the system] work."*

In this paper we report on our study of one of those simplification strategies, namely *push-to-talk*, and show that, for the task under investigation, the observation in the quote above does *not* seem to hold. We collected experimentally a corpus of human–human task oriented dialogue, keeping the task constant but varying turn-taking conditions between *free* (free turn-taking, FTT), and *restricted* (push to talk, PTT). We found that while the PTT dialogues showed less interactivity (fewer, but longer turns; longer silences), the restriction did not slow down task completion. Hence, the PTT dialogues were actually *more* efficient (on a "per word" basis: reaching the same goal with fewer exchanged words). As we shall see, this is achieved by a different macro-pattern of dialogue acts, which in PTT by definition excludes concurrent feedback. Our findings seem to support the idea that the often stated goal of 'naturalness' in SDS design may not always be optimal in terms of efficiency.

The remainder of this paper is structured as follows. We first describe the experimental setup (task, setting and conditions), and present the measures we used to characterise the collected dialogues. In Section 4, we report our statistical analysis of the data with respect to turn-taking and discuss the results obtained. We close with some conclusions and further work.

2. EXPERIMENTAL SETUP

In this section we describe the task and the data collection experiment, which provided the corpus of human-human dialogues used in this investigation.

2.1. Task

In order to study the issues described above, we chose a simple task involving two participants with distinct roles (*player* and *executor*) who collaborate on reconstructing a puzzle. The player has access to the solution of the puzzle, while the executor is given the outline and the pieces. This setting, which is similar to the one used in the classic Tangram experiments of [7], gives rise to interactions common in instructional dialogue: the player holds the initiative and gives instructions to the executor, in order to reconstruct the puzzle. Figure 1 shows the solution of the puzzle and the target outline that is given to the executor.

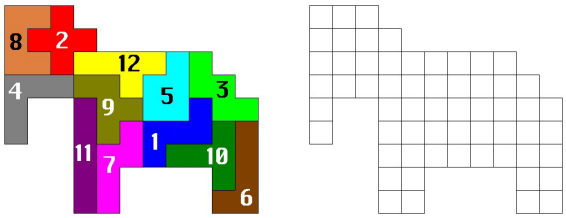

Fig. 1. Solution and Outline

As shown in the figure, the puzzle pieces in the solution were numbered. The player was told to follow the order given by these numbers when guiding the executor through the reconstruction process. The aim behind this was two-fold: first, we wanted to avoid excessive reference to previously positioned pieces in order to increase potential for clarification (as we thought that this might be a source of differences between turn-taking conditions); second, we wanted to have a common reconstruction process for all collected dialogues to allow for more systematic comparisons. The pieces are also shown in different colours here for easier identification; however, the pieces that the executor manipulated were all the same colour and were not numbered. Beside the set of pieces, the executor was given an empty gridded outline, as shown in the right hand side of Figure 1. Both player and executor were aware of the information available to their partners.

2.2. Setting

The experiments involved 20 subjects, 11 females and 9 males, grouped in 10 player-executor pairs: four female-female pairs, three male-male pairs, and three female-male pairs. All subjects were German native speakers between 20 and 45 years old, and the recorded conversations were in German. During the experiment the player and the executor were in different rooms and communication between them was strictly 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).

2.3. Turn Taking Conditions

We investigated two different turn-taking conditions: *free turn-taking* (FTT) and *push-to-talk* (PTT). In the FTT condition, player and executor communicate by means of microphones and headsets. The channel is continuously open and therefore turn-taking is *natural*, as it would be for instance in a telephone conversation. In the PTT condition, on the other hand, subjects communicate using walkie-talkies that only offer a half-duplex channel. Here 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)—i.e., subjects cannot freely take the turn nor barge in at will.

Five pairs of subjects were assigned to each of these 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.

3. DATA ANALYSIS AND METHODS

We collected a total of 10 dialogues (5 for each turn-taking condition), which make up a total of 194.54 minutes of recorded conversation. The recordings were transcribed and segmented using the software Praat [8]. The transcribed corpus contains a total of 2,262 turns and 28,969 words.

To analyse our FTT and PTT sub-corpora, we used the measures listed below. They were calculated globally for each dialogue in each condition, and the five last measures were also calculated independently for player and executor.

- **min/dial**: length of dialogue in minutes
- **wrds/dial**: average # of words per dialogue
- **trn/dial**: average # of turns per dialogue
- **sec/trn**: average length of turns in seconds
- **wrds/trn**: average # of words per turn
- **wrds/sec**: average # of words per second

Higher-level annotations were done in MMAX [9]. Here we annotated each turn with one or more dialogue acts (DA). The set of DA tags we used is partially inspired by the DAMSL scheme [10], but adapted to the the needs of our task. We make a first distinction between task and grounding acts. Task acts are further classified into task-execution and task-management acts. Grounding acts include different types of feedback acts, as well as clarification requests (CR). A summary of the DA tag set used is shown in Table 1. CRs were further

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

Table 1. DA Tag Set

classified with a simplified version of the scheme developed by [11].

After each experiment subjects completed an online questionnaire, where they were asked to evaluate the setting, the difficulty of the task, and the collaboration with their partners.[1]

4. FTT *VS.* PTT: RESULTS

One of the obvious differences between FTT and PTT is that, by design, the latter prevents overlap—a speaker can either send or receive at a time—, while FTT allows simultaneous speech. Besides the presence/absence of overlap, however, we found that the different conditions had a clear effect on the sequencing of the interaction and the degree of interactivity. Our FTT dialogues contain roughly twice as many turns as PTT dialogues; even if turns which are in complete overlap are not counted, the number of turns (which we take as indicator of interaction) is still significantly higher in FTT ($p < 0.025$).[2] PTT turns, on the other hand, are on average twice as long as FTT turns. The two different interaction patters are visualised in Figure 2.[3]

Table 2 summarises the results of the analysis of our FTT and PTT sub-corpora for each of the measures listed above. The third row shows the statistical significance of the differences between the FTT and the PTT values, calculated globally

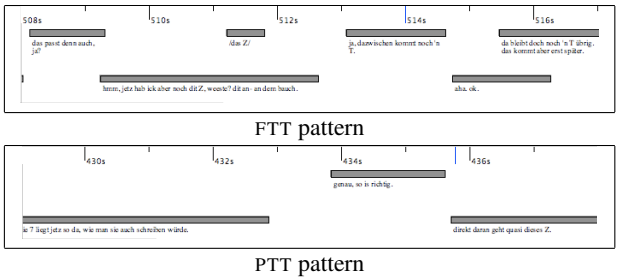

Fig. 2. Sequencing and Interaction Patters

(G) and for the player (P) and the executor (E) when appropriate. The stars indicate degrees of significance: one star signifies $0.025 \leq p \leq 0.05$; two stars $0.01 \leq p < 0.025$; three stars $p < 0.01$. We use a dash when differences are not statistically significant.

A somehow striking result is that, *for the task at hand*, the different interaction patterns enforced by the two different turn-taking policies do not lead to a significant difference in task completion times. Coupled with the observation that the number of words uttered in PTT dialogues is significantly lower than in FTT dialogues, and the speaking time (excluding inter-turn pauses) also tends to be lower in PTT, this allows the conclusion that the PTT dialogues are actually *more* efficient than the unrestricted, more "natural" FTT dialogues.

The DA annotation indicates that PTT dialogues are more focussed on the task while FTT dialogues devote a substantial effort to grounding behaviour and the management of interaction. Table 3 shows the average percentages of DA tags from the total of acts in each turn-taking condition. PTT dialogues tend to contain more task-related acts (45.2% on average) than FTT (38.1%), although this difference is not highly significant ($p = 0.055$). The most significant difference is found in the amount of grounding acts, in particular in the number of positive feedback acts, like backchannels and acknowledgements, which is consistently higher in FTT ($p = 0.01$).

DA Tags	FTT	PTT	t-test
Task-related acts	38.1	45.2	–
Positive feedback	34.3	26.7	***
Other grounding acts	23.9	22.6	–
Other acts	3.7	5.5	–

Table 3. Average % of DA Tags from Total of Acts

38% of positive feedback acts in FTT were uttered in complete overlap. As Clark [12] points out, one of the functions of overlapping backchannels is to signal that the utterer of the backchannel does not intend to take the turn, but instead encourages the other dialogue participant to go on. This kind of interaction management actions are necessary in unrestricted conversation, but significantly decrease in frequency in PTT, where subjects are freed from the pressure of managing turn-taking. The absence of this pressure seems to balance the lack of constant grounding behaviour (which should in prin-

[1] The questionnaire given to the subjects is available online at http://www.ling.uni-potsdam.de/DEAWU/Questionnaire/Fragebogen.html (in German).

[2] This notation indicates that the probability p that the observed difference between conditions is due to chance is lower than 0.025.

[3] Figure 2 shows segments of two dialogues in our corpus as displayed by the tool ZeitWort. More information on this tool can be found at http://www.ling.uni-potsdam.de/∼das/potsdiallab.html

Para-	FTT			PTT			t-test significance		
meters	G	P	E	G	P	E	G	P	E
min/dial	20.1	n/a	n/a	18.5	n/a	n/a	–	n/a	n/a
wrds/dial	3540	2127	1413	2254	1551	702	*	–	–
trns/dial	328	150	178	115	58	56.4	***	***	***
sec/trn	3.71	5.5	2.56	7.21	10.3	4.04	***	**	**
wrds/trn	11.3	15.2	7.97	20.2	27.9	11.8	*	–	–
wrds/sec	3.03	3.03	3.15	2.75	2.73	2.9	–	–	–

Table 2. Comparison of Quantitative Measures in FTT and PTT

ciple be a downside according to grounding models like that of [12]), leading to no loss in efficiency.

Interestingly, the dialogue participants in the role of the player in the FTT dialogues were significantly less sure ($p < 0.01$) about what the other dialogue participant (the executor) wanted, (lower score on the question "Did you always understand what the executor wanted from you?"), which suggests that the executors in PTT took more care designing their contributions.

5. CONCLUSIONS

We have presented an empirical study based on a corpus of task-oriented human-human dialogues, which we have collected under controlled conditions. In particular we have compared free turn-taking with a restricted turn-taking policy based on push-to-talk, which resembles turn-taking simplifications commonly adopted in spoken dialogue systems. We found that the different turn-taking conditions lead to different interaction patterns: push-to-talk dialogues show less interactivity but, surprisingly, this does not slow down task completion.

Although this was an exploratory investigation, these results suggest that, for some tasks, a loss in 'naturalness' need not lead to a loss in efficiency. In the future we plan to investigate in more detail the conditions under which this holds by experimenting with other tasks. We also plan to explore and test other turn-taking simplifications, like the pause threshold method, where the system waits for pauses of a certain length before taking the turn.

At the moment we do not have a satisfactory explanation of how exactly the loss of interactivity and continuous grounding is counterbalanced in the push-to-talk dialogues. We are investigating several alternative hypotheses, which however require a more detailed content-analysis of the turns, and especially, we believe, of the referring expressions used in the task acts.

6. ACKNOWLEDGEMENTS

The experiments reported in this paper were run in the Phonetics Lab at the Zentrum für Allgemeine Sprachwissenschaft (ZAS) in Berlin. This work has been supported by the DEAWU project, Marie Curie Transfer of Knowledge grant EU #FP6-2002-Mobility 3014491. Thanks to Andrea Corradini for collaboration on specifying the task.

7. REFERENCES

[1] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematic for the organization of turn-taking in conversation," *Language*, vol. 50, 1974.

[2] C. Ford, B. Fox, and S. Thompson, "Practices in the construction of turns: The "TCU" revisited," *Pragmatics*, vol. 6, no. 3, 1996.

[3] R. Sato, R. Higashinaka, M. Tamoto, M. Nakano, and K. Aikawa, "Learning decision trees to determine turn-taking by spoken dialogue systems," in *Proceedings of ICSLP-02*, 2002.

[4] L. Ferrer, E. Shriberg, and A. Stolcke, "A prosody-based approach to end-of-utterance detection that does not require speech recognition," in *Proceedings of ICASSP'03*, 2003.

[5] N. Ström and S. Seneff, "Intelligent barge-in in conversational systems," in *Proceedings of ICSLP-00*, 2000.

[6] J. Allen, G. Ferguson, and A. Stent, "An architecture for more realistic conversational systems," in *Proceedings of the Conference on Intelligent User Interfaces*, 2001.

[7] H. Clark and Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, vol. 22, 1986.

[8] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9–10, 2001.

[9] C. Müller and M. Strube, "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*, 2001.

[10] J. Allen and M. Core, *DAMSL: Dialogue Act Markup in Several Layers*, 1997.

[11] K. Rodríguez and D. Schlangen, "Form, intonation and function of clarification requests in German task-oriented spoken dialogues," in *Proceedings of Catalog'04*, 2004.

[12] H. Clark, *Using Language*, Cambridge University Press, 1996.