Beyond Repair – Testing the Limits of the Conversational Repair System

David Schlangen *and* Raquel Fernández
Institute for Linguistics
University of Potsdam, Germany
{das|raquel}@ling.uni-potsdam.de

Abstract

We report on an experiment on the effects of inducing acoustic understanding problems in task-oriented dialogue. We found that despite causing real problems w.r.t. task performance, many instances of induced problems were not explicitly repaired by the dialogue participants. Almost all repairs referred to the immediately preceding utterance, with problems in prior utterances left unacknowledged. Clarification requests of certain forms were in this corpus more likely to trigger reformulations than repetitions, unlike in different settings.

1 Introduction

Clarification requests (CRs), i.e., utterances that request repair of understanding problems, are typically studied on corpora of transcribed conversations (see, *inter alia*, (Purver, 2004; Rodríguez and Schlangen, 2004)). While much knowledge about the use of this utterance type has been gathered this way, there are principled limitations to this approach:

• If there is a CR, the problem that caused it must be inferred from its form and the original speaker's reply, as it cannot be directly observed.

• As it is not obvious for the annotator whether there has been a problem or not, strategies for *avoiding* to ask for clarification cannot be studied straightforwardly.

• The effectiveness of the repair system can only indirectly be studied.

In this paper, we present the results of an experiment where we addressed these limitations through the controlled induction of understanding problems.

The remainder of the paper is structured as follows. In Section 2 we describe the method used in our experiment, the results of which are then presented in Section 3. A general discussion and conclusions close the paper.[1]

2 The Noisy Channel Experiment: Method

2.1 Overview

The experiment consisted in a voice-only cooperative task with two participants: an instruction giver (IG) had to describe in order of the numbering the placement of pieces on a puzzle (see Figure 1) to an instruction follower (IF), who only had access to the unsolved puzzle with unnumbered pieces.

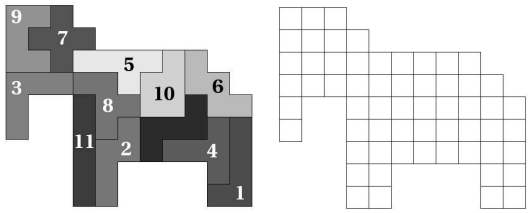

Figure 1: Solution and Outline

In half of the runs we manipulated one audio-channel by replacing (in real-time, at random points) all signal with noise, effectively blocking out the speech for the hearer. Around 10% of one speaker's

[1]The work described here is the second part of an experiment whose first part has been described in (Schlangen and Fernández, 2007). The part described here shares the general set up with that other work (i.e., introduction of noise in one channel), but uses different materials (a different task) and coding, and has a different focus for the analysis.

signal was removed in this way. The random, automatic placement of noise meant that we had no control over which part of the utterance exactly was masked, but we judged this preferable over more controlled manual placement of noise, which necessarily would have disabled real-time interactivity. The design is related to (Skantze, 2005), where distortion was introduced through a simulated ASR, although not in real-time.

We expected the manipulation to have an effect on the effort needed to complete the task and each of its steps (placing individual puzzle pieces). Further, and more specifically, given previously observed correlations between CR forms and problem types, we expected an increase in use of CR forms previously connected to clarifying acoustic problems. As our design tells us exactly which part of the stimulus was problematic, we also wanted to explore relations between this and whether, and if so, how clarification was requested.

2.2 Procedure

26 subjects (13 pairs) participated in the experiment. All were native English speakers (from a variety of native countries) that responded to a public call for participation. Half of them where college students while the other half had a range of different occupations. The age range was from 20 to over 40. None of the subjects reported any hearing difficulties.

The pairs of subjects were split into IG and IF and placed in different sound-proof rooms, connected by an audio-line via headsets. They were then separately briefed on the task. IG's solution was displayed on a computer screen, IF's puzzle board was implemented in a computer program. All audio was recorded; in the runs with the manipulation, both the audio before adding noise and after adding noise was recorded. IF's computer screen was video-taped.

2.2.1 Data Analysis

For analysis, the recordings were transcribed using Praat (Boersma, 2001) and annotated using MMAX (Müller and Strube, 2001); the annotators had access to both the textual transcripts and the audio material.

We segmented the recordings into *utterances* (following the guidelines in (Meteer and Taylor, 1995)) and *moves*, which we defined as all utterances belonging to the placement of one piece. We then annotated the *transition status* at move boundaries, split into `grounding state`, where a) the participants can be explicitly *confident* about their placement ("OK, I've got it. Next one!"); b) rather *unconfident* ("Well, I'll put it there. Let's see what happens."); c) they can put the current sub-task *on hold* and go back to a previous piece; d) which in turn then can be moved and placed with any of these previous grounding outcomes, or can be *reconfirmed*; and `success`, which we checked on the video recordings. Values for this feature are: *success*, *failure*, *not moved* (for moves that revisited previously placed pieces, but did not move them), and *on hold* for moves that are on hold while a previous piece is repaired.

Within the moves, we marked regions belonging together thematically, and annotated them with the following categories: a) identification of the *piece* that is to be placed; b) specifying its *orientation* and c) *location* on the grid; other common dialogue actions were d) talking about the *task setup* ("I am supposed to do these in order"); e) the *grounding status* ("well, let's see what happens"); f) noting *problems* ("This doesn't work. Something must be wrong."); g) giving a *description of the state* of the board ("To the left I have the Swiss cross, and next to it..."). Everything else was coded as h) *other*.

Finally, we identified utterances that were CRs and coded them with (Rodríguez and Schlangen, 2004)'s scheme; for reasons of space, we refer to that paper or to (Schlangen and Fernández, 2007) for a description of the values.

3 Results

3.1 Recordings

The 13 experimental runs resulted in 9 usable recordings, as two runs had to be excluded because of equipment failure and two because subjects aborted the task or didn't follow instructions.

3.2 Dialogue-based Analysis

The pairs in the noise condition finished the task in an average 1130 seconds, producing in average 653 utterances; the pairs in the control group needed 618 seconds and 422 utterances. These differences are statistically significant (Welch's t-test; t=2.7, df=4.7,

	success	failure	not_moved	on_hold
noise	57.14%	17.86%	10.71%	14.29%
no-noise	89.19%	5.40%	2.70%	2.70%

	confid	unconf	on_hold	reconf
noise	61.90%	9.52%	21.43%	7.14%
no-noise	94.60%	0%	5.40%	0%

Table 1: Success of Moves, in Percent of all Moves (top) and Grounding Status at Move-Transitions

$p<0.05$ for length in seconds; t=2.8, df=7.0, $p<0.05$ for utterances). There are however no significant differences between the groups (χ^2) w.r.t. how much time was spent on different sub-tasks like identifying pieces or placements: the pairs in the noise condition don't do different things, they just do the same things for longer / more often.

3.3 Move-based Analysis

Table 1 shows the distributions of move outcomes. The majority of moves in the no-noise condition end with confident and successful placement. In contrast, in the noise condition only just over half of the moves are actually successful, and consequently there are more moves that are repairs of previous mistakes. The differences between the groups are significant (χ^2, for both $p<0.01$).

The mean length of moves in terms of utterances is very similar for both groups (28.5 for noise group, 30.81 for control group), and indeed the difference is not significant: there seems to be a constant upper limit on how much time is spent on each move before the players move on, confidently or not.

Table 2 shows the ratio of contributions by IG and IF within each move, averaged over all moves and separated according to *grounding status* and *description of state*; e.g., the "54/46" in the second line means that 54% of contributions in moves in the noise group that ended in a wrong placement came from IG and 46% from IF. Problems in a move that lead to an unsuccessful conclusion and/or not-confident grounding only in the control group had an effect on the contribution ratio, leading to more contributions by IG. (The differences are significant, χ^2 tested, * $p< 0.05$, *** $p<0.001$.)

	noise	no noise	signf.
all	55 / 45	57 / 43	
wrong	54 / 46	68 / 32	*
corr.	56 / 44	56 / 44	
!conf	54 / 46	74 / 26	***
conf	57 / 43	57 / 43	

Table 2: Ratio IG/IF contributions, by move success

3.4 Utterance-based Analysis

The recordings of the noise group have been segmented into 3249 utterances, those of the control group into 1607. In the noise group, there were 561 utterances that contained noise, i.e., 30.1% of all IG utterances (only those can contain noise). Only 28 of those (= 5.0%) triggered a clarification request (that is, were coded as being the antecedent of one). In the noise group, there was only one CR that was not triggered by a noise utterance; in the control group there were 8 CRs altogether.

The majority of turns (both of IG and IF; turn defined as sequence of utterances before speaker change), was one utterance long, this tendency being stronger in the control group (61.8% compared to 55.6% in the noise group; difference in length distribution is significant, χ^2, $p<0.001$). However, there were turns of length up to 13 utterances.

In all utterances within IG turns in the noise group (i.e., at all distances from the speaker transition), noise events were equally likely to occur. However, a noise event in an utterance *at* the transition point—that is, in either the last utterance of a longer turn or in a single utterance turn—had a chance of 8.33% of triggering a CR. A noise event one utterance away from the transition point only has a 0.87% likelihood of triggering a CR. There are no CRs in the corpus whose antecedent is further away.

Lastly, we turn to a more fine-grained analysis of the clarification requests that occurred. We compared the distributions of CR-features in this corpus with that resulting from the the other task done in the same setting, where items like strings of numbers and sentences were read from a screen by IG for the IF to write them down (see (Schlangen and Fernández, 2007)).

What is interesting here is that despite the manipulation being the same, there were significant differences in the CRs that occurred: in the puzzle task

of the present paper, there were significantly more CRs that did not point at the exact problem location (`extent`), more CRs that did not present a hypothesis (`severity`), fewer CRs constructed through repetition of material (`rel-antec`), and fewer replies to CRs that were repetitions, and more reformulations or elaborations (`answer`). (All differences were tested with a χ^2 test, $p<0.01$.)

4 Discussion and Conclusions

We now briefly summarise these observations: Pairs in the noise condition needed significantly longer to finish the task, and this was not due to higher effort for repairing understanding problems, but rather to higher effort needed for repairing task-level problems, i.e. wrong placements. In fact, while there were more repairs in the noise condition than in the control condition, most induced problems went unacknowledged – and as the performance differences show, it seems to be valuable information that they miss.

That CRs typically clarify the immediately preceding utterance has been observed before (Purver, 2004; Rodríguez and Schlangen, 2004). Our setting allows us to see the strength of this constraint: even if there are problems with earlier utterances within a turn—and we know that they are there, as we produced them—, they are a lot less likely to be repaired than those in the last utterance of a turn. We speculate that IF judged the information gain they would achieve by clarifying too low to take the step to interrupt IG's turns. They rather settled on a more independent strategy with more reliance on tentative placements (as shown by the grounding status), which for this task turned out to be less successful than understanding IG's commands. It seems that there needs to be a baseline of understanding before utterance-level clarification is even attempted.

Another interesting observation is that while the *forms* of the CRs that are present are not significantly different from those in comparable conditions but with different task (see previous section), the CRs are interpreted differently: significantly often, forms that trigger verbatim responses in that other corpus trigger reformulations or elaborations here. There are two possible explanations (not mutually exclusive): the CR addressees are more primed to expect clarification requests that target the meaning level (Clark, 1996) and hence treat the CRs as being such. Or, given the spontaneous, rather unplanned nature of these also often rather long description utterances, there are memory limitations that make verbatim responses harder.

To summarise, our results show that a) clarification is not *automatic*, but underlies complex considerations about the value of the missing information; b) CR forms are interpreted in a (task-)context-dependent way.

In future work, we will look in more detail at the dialogue acts of the utterances at turn-boundaries. We also plan to test task-performance in the same setting, but with the IF instructed to follow a clarification policy of 'always interrupt and clarify if there is noise'.[2]

References

Paul Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Marie Meteer and Ann Taylor. 1995. Dysfluency annotation stylebook for the switchboard corpus. http://www.cis.upenn.edu/∼bies/manuals/DFL-book.pdf.

Christoph Müller and Michael 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*, pages 45–50, Seattle, USA, August.

Matthew Purver. 2004. *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, King's College, University of London, London, UK.

Kepa Joseba Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in german task-oriented spoken dialogues. In *Proceedings of Catalog (SemDial04)*, pages 101–108, Barcelona, Spain, July.

David Schlangen and Raquel Fernández. 2007. Speaking through a noisy channel - experiments on inducing clarification behaviour in human-human dialogue. In *Proceedings of Interspeech 2007*, Antwerp, Belgium, August.

Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325–341.

[2]**Acknowledgements:** Thanks to: S. Bachmann, A. Steinhilber, H. Bohle (transcription and annotation); M. Waeltermann (noise program); J. Dreyer (ZAS Berlin), B. Pompino-Marshall (HU Berlin), P. Healey and G. Mills (QMU London) (lab use); M. Stede and A. Corradini (discussions of set-up). This work was supported by EU (Marie Curie Programme) and DFG (Emmy Noether Programme).