Speaking through a noisy channel – Experiments on inducing clarification behaviour in human-human dialogue

David Schlangen and *Raquel Fernández*

Department of Linguistics
University of Potsdam, Germany
{das|raquel}@ling.uni-potsdam.de

Abstract

We report results of an experiment on inducing communication problems in human-human dialogue. We set up a voice-only cooperative task where we manipulated one channel by replacing (in real-time, at random points) all signal with noise. Altogether around 10% of the speaker's signal was thus removed. We found an increase in clarification requests of a form that has previously been hypothesised to be used mainly for clarifying acoustic problems. We also found a correlation between the percentage of an utterance being manipulated and the use of devices for pointing out error locations. From our findings, we derive a gold-standard policy for clarification behaviour.
Index Terms: dialogue, clarification requests, error-handling

1. Introduction

There recently has been a number of studies [1, 2, 3, 4, 5, 6, 7, 8] of Clarification Requests (CRs), *i.e.* utterances like B's in the following examples:

(1)	A: Did you talk to Peter?	‖	A: I brought a 3-5 torx.
	B: Peter Miller?	‖	B: What's that?

The interest in these constructions is well motivated, as understanding their use has eminent practical relevance (implementing similar clarification behaviour could improve the way spoken dialogue systems deal with understanding problems) as well as theoretical importance (their semantics has to be defined in terms of previous utterances, not propositions).

Most of these previous studies were done on corpora of dialogue recordings (see below for exceptions), and as we will argue in more detail below, this imposes certain limits:

- the causes for asking the CRs must be guessed *post factum* by the annotator;
- there is no control over the problem source;
- strategies for *avoiding* to ask for clarification cannot be studied straightforwardly.

In this paper, we present an experiment where we controlled the problem (by replacing signal with noise) and hence were able to overcome these limitations. Our findings further corroborate, by being achieved with a different methodology, previous results as well as offering more detailed insight into human clarification policies.

The remainder of the paper is structured as follows. In the next section we extend the discussion of previous work, both corpus-based and experimental. We then describe the method used in our experiment (Section 3), present the results (Section 4), and then close with a general discussion and conclusions.

2. Empirical Work on Clarification

2.1. Previous Corpus Studies

The CR investigation reported in [1] was based on text transcripts from the British National Corpus. They found a rate of 4% of all utterances being CRs. Their annotation scheme has good coverage, but was more motivated by theoretical concerns than by use as annotation scheme. (See detailed review in [4].) In this paper, we use the more fine-grained scheme introduced in [4] (more on this below). Using both transcripts and audio, [4] annotated 5.8% of all utterances in their corpus of task-oriented German dialogue as CRs. The same scheme (with some modifications) has been used in [6] on an English task-oriented corpus, with 4.6% CRs.

One finding of [4] that we will take as a starting point here is that intonation disambiguated between fragmental (non-interrogative) fragments used for clearing up acoustic problems and those used for clearing up references, as illustrated in the left hand side of following example:

(2)	A: I saw Peter.	‖	A: I saw Peter	*problem utt.*
	B: Peter ↗ | Peter ↘	‖	B: Who?	*CR*
	A: Yes. | My cousin.	‖	A: My cousin.	*CR reply*
		‖	B: Ah. OK.	*follow-up*

What these studies have in common is that for annotating CR functions, they rely on the CR reply and the follow-up to clear up ambiguities. This does not always work; [4] e.g. labelled the high number of 14.3% cases as ambiguous. For this there is a systematic reason, namely that "over-answering" is always an option. In the right hand side example in (2) the CR clears up both reference resolution problems (by giving a different description of the intended referent) and acoustic problems (by repeating reference, albeit with a different form); in such cases there is little objective reason for preferring one annotation over the other. This shortcoming suggests that experimental work with more control over the CR-process could supplement the corpus work.

2.2. Experiments

There are two possible directions for such experimental work. Healey and colleagues used a paradigm in a number of studies [2, 8] where a modified text-based chat tool is used which can, based on patterns in the input, temporarily take one participant off line (without him noticing), in order to send a "fake" (i.e., not actually typed by the real dialogue participant) CR to the other participant. After the CR-sequence is done, the original participant is seamlessly put back on-line, and the conversation continues without the participants being aware of any modifica-

tion. With this setting, *CRs* can be controlled and hence interpretations (and other effects) can be better studied.

The other possibility is to control the *problems* in the dialogue, by creating them. [5] conducted an experiment where one channel in a dyadic conversation was filtered through an automatic speech recogniser, and the participant (the wizard) had to base her reactions solely on this (highly defective) input. One interesting finding of this study is that the wizard preferred to ask task-related questions rather than direct CRs, from which one can infer that the understanding of the actual utterance was in this setting less important than understanding the task-related intentions.

Our experiment reported here also follows this route. We control the problem by replacing some of the signal of one speaker with noise, hence creating acoustic understanding problems. By targeting a different level (acoustic understanding), our experiment can contribute evidence that is complementary to the findings described above. We also think, however, that it has certain advantages over [5], by being more clearly modelled on "natural" situations (e.g., transient noise through environment events like passing cars, etc.). In contrast, it is not clear how human reactions to the highly untypical ASR output are revealing of natural behaviour (but see discussion in [5]).

3. The Noisy Channel Experiment: Method

3.1. Overview; Hypotheses

The experiment consisted in a voice-only cooperative task between two participants where in one condition we manipulated one channel by replacing (in real-time, at random points) all signal with noise. Altogether around 10% of the speaker's signal was thus removed. The roles of the participants were asymmetric: the instruction giver (IG) read items from a screen and dictated those to the instruction follower (IF).[1] Only the channel from IG to IF was manipulated in the experiment group.

We expected the manipulation to have an effect on the effort needed to complete one dictation item, with different item types (see below) being vulnerable to different degrees. 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 the specificity of the CR.

3.2. Experimental Design

3.2.1. Subjects & Materials

A total of 32 subjects, arranged in 16 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 were college students while the other half had a range of different occupations (including web designers, teachers, musicians and waiters). 21 of them were in their twenties, 7 in their thirties and 4 were over 40 years old. None of them reported any hearing difficulties.

As mentioned above, the task consisted in the IG dictating items to the IF. There were 44 items altogether, of 4 types: a) num*bers*: strings of numbers; b) sent*ences*: "normal" sentences; c) id*ioms*: conventional sentences or phrases like "a stitch in time saves nine"; d) *modified idioms* (`fid`): idioms where we exchanged one word, e.g. "All doors lead to Rome".

These different types systematically vary the amount of available context (or mutual information): numbers offer none at all; sentences prime normal syntactic expectations and collocations; idioms set up very strong expectations; which in the manipulated idioms are misleading.

3.2.2. Procedure

IG and IF were placed in different sound-proof rooms, connected by an audio-line (via headsets; 22kHz frequency range). The subjects were then individually briefed on the task. The IG had in front of him a computer program that displayed the dictation items, one at a time, with the IG being able to skip to the next item (but not back). The IF used a computer to type the dictated items into a text editor. During the run, 2 (control group) or 3 (noise group) channels of audio were recorded (IG w/o and (if appropriate) w/ noise; + IF), as well as a screen capture video of the text editor. Logging messages of noise program on the duration of noise event were also kept (for synchronisation with the recordings).

The noise-insertion program is purpose-built. It operates in (near) real-time (with $\sim$5ms response time); when a signal over a certain threshold is detected, user-changeable parameters determine the likelihood that the signal is replaced by noise or respectively that noise is switched off again. We used brown noise, as it is less unpleasant to the listener than white noise.

3.2.3. Data Analysis

For analysis, the recordings were transcribed using Praat [9] and annotated using MMAX [10]; the annotators had access to both the textual transcripts and the audio material.

• *utterance:* We followed the utterance segmentation conventions from [11]; roughly, independent syntactic units are grouped as one utterance.

• *move:* We grouped together as one move all utterances belonging to the dictation of one item, beginning with the first task-related utterance (such as "OK, now numbers again.") and ending with the final confirmation by IF of task-completion.

• *effort:* We used 'average repetition rate' (`arr`) to measure the effort spent per dictation item. `arr` is calculated per move, as the number of words from the current item spoken by IG, normalised by the number of words in this item. *I.e.*, if `arr` = 1, then every word from the dictation item was spoken once by IG (obviously, this is the minimum if the task is done correctly). A value of 1.5 means that some words have been repeated, etc.

This measure is rather coarse-grained, as it anchors the effort spent on one item only on the repetitions of IG and disregards for example check-questions or repetitions by IF, but it has the advantage of being relatively straightforward to code, and, more importantly, unlike time-based measures like 'length of move', it is robust against individual differences w.r.t. typing speed, delivery in installments, off-topic talk, etc.

• *noise classification:* Noise events were classified with the following features. `dictated` marks whether noise removed something that was part of a dictation item or not; `noise_what` gives information about what was in the noise, it has the values a) *part_of_word*; b) *whole_word*; c) *whole_phrase*; d) *everything*. `noise_extent` classifies the extent of the noise relative to the utterance length, as *<10%*; *10-33%* etc. up to *everything*.

• *CRs:* We coded CRs with [4]'s scheme (see there for details). - The possible values of the attribute `mood` are a) *declarative*: default word-order (not interrogative or imperative), modified

[1] At the same session, and before the dictation task, the subjects also tackled a different task (reconstructing a puzzle). We will report elsewhere the findings on this task.

# of CRs	0	1	2	3	4	5	6
ns mv	44.71	40.00	8.82	4.12	1.18	0.59	0.59
no-ns mv	98.90	1.10	0	0	0	0	0

Table 1: % of moves containing 0,1,..,6 CRs; by condition

by `f` or `r` for falling and rising boundary tone, respectively; b) *polar question*: fully realised syntactic polar interrogatives; c) *alternative question*; d) *wh-question*; e) *imperative*; f) *other*. We added g) *gap*, for CRs that are characterised by a lengthening of the final vowel and a mid-level tone, e.g. as in "A: It was hard for *noise* B: It was hard for _ ?".

- Values for the attribute `form` are a) *particle*: or conventional phrase, e.g. "pardon?"; b) *partial*: a syntactic fragment; c) *complete*: a syntactically 'complete' sentence.

- Values for the attribute `rel-antec` are a) *repetition*: parts of the problematic utterance are repeated *literally*; b) *add-wh*: a *wh-word* is added ("you saw what?"); c) *addition*: other additions; d) *reformulation*: a phrase is uttered that is co-referent to elements of the original utterance, but is not a literal repetition; e) *independent*: no elements of the problematic utterance are repeated or reformulated.

- Finally, The attribute `antecedent` holds the ID of the utterance that triggered the CR; `extent` codes whether CR points out an element in the problem utterance as having caused the problem (*yes/no*); and `severity` marks whether the CR does present a hypothesis for confirmation, or not.

Note that with the exception of the CR features, the features listed here are relatively straightforward to annotate (and could perhaps even be automated), which is why we didn't systematically test for annotator agreement. The CR scheme has been used repeatedly by different groups and hence can be regarded as validated.

4. Results

4.1. Recordings

The 16 experimental runs resulted in 12 usable recordings (two runs had to be excluded for technical reasons (equipment failure) and two because subjects were not suitable for the task due to e.g. dyslexia). The usable recordings make up a total length of 205 minutes (avg. length per dialogue: 17.14 min). The transcriptions were segmented into 7469 utterances (in average 622 per dialogue).

4.2. Analysis of Moves

To give an impression of the effect of the manipulation, the following example shows a typical exchange in the noise-group. (Material that was not audible to IF is marked by square brackets; punctuation represents boundary tones, '.' low, '_' mid, '?' high; selected annotation is shown on CRs.) The example nicely illustrates many different means for requesting clarification.

(3) IG: Cris , without an h , had the [right word] on the base of his [tongue] .
IF: Cris ha:d _ */gap, partial, repetition, y-ext, no-hyp/*
IG: The right w[ord] on the ba[s:]e . of his tongue .
IF: Word or verbs ? */alt-q, partial, repetition, y-ext, y-hyp/*
IG: Word .
IF: Word on the base of his tongue ? */r-dec, part, rep, y-ext, y-hyp/*
IG: Mhm .
IF: OK . Next .

For analysis, we grouped all "noise-moves" (moves with at least one noise-event in them) together and all "no-noise-moves" (no noise events) Note that the latter can also occur in dialogues from the noise-group, if by chance there was no noise event for one whole move. CRs occurred overwhelmingly in noise-moves. Table 1 plots the percentage of (no)noise-moves that have various numbers of CRs in them. For noise-moves, it shows a power-law distribution: most have only one or two CRs, a few have 4 and more. It also shows the relatively high resistance against noise: even of the noise-moves, 45% were without CRs. There is a strong correlation between the number of noise events per move and the number of CRs (r = 0.61; $p = 0.001$), however.

	all	`num`	`id`	`sent`	`fid`
noise	1.47	1.15	1.51	1.63	1.54
no-ns	1.30	1.23	1.21	1.30	1.54
sig	p=0.01	–	p=0.01	p=0.001	–

Table 2: Average Repetition Rate, Conditions / Item Types

`noise what`	*part_of_word*	*whole_word*	*whole_phrase*
	18.08%	55.07%	63.0%
`noise extent`	*less than 10%*	*a quarter*	*a third*
	13.76%	20.91%	47.30%
	half	*two thirds*	
	42.59%	70.00%	

Table 3: Likelihood of triggering CR, by "damage rate"

Altogether, 71% of noise events did not elicit a CR; interestingly, the rates vary for the item types: `num` 78.0%; `sent` 61.6%; `id` 55.6%; `fid` 59.3%. The average effort (as measured by `arr`) is shown in Table 2; again, there is a difference between item types, with numbers and modified idioms showing no significant change in effort due to introduction of noise (Wilcoxon signed rank test). The difference in average effort between the item types is only significant within the noise moves (ANOVA, $p = 0.001$).

4.3. Analysis of Stimulus-Response Pairs

We now turn to an analysis of the utterances (especially pairs of problem-utterance / CR as identified by `antecedent`) within the noise condition. Of the 3048 utterances in the noise-group, 406 contained noise, and 135 were identified as CRs. The vast majority of CRs were uttered by IF (98%). For the majority of CRs, an utterance containing noise was annotated as antecedent (89.6%).

As discussed above already, the majority of noise-events (71%) did not result in CRs (at least not directly). This robustness differs with respect to what was said: noise in dictation instructions (as coded with `dictated`) is more likely to lead to CRs than noise in other utterances (30% vs. 1.2%). As Table 3 shows, the more material is missing, the more likely the utterance is to trigger a CR. Looking in more detail into the CRs that are triggered by noise, it can be seen that the greater the "damage" to the utterance, the *less* likely `cr_form`:*rising_declarative* becomes (from 61% to 28%), and the *more* likely *gap* becomes (23% to 57%); similarly, greater damage leads to more CRs that do not offer a hypothesis (`cr_sev`).

Figure 1 finally plots the distribution of CR-types found in our corpus (deawu) and in [4] (baufix).[2] It shows that in our corpus there are relatively much more rising declaratives and

[2]Due to the low number of CRs in no-noise moves, comparisons between conditions were not meaningful and hence only this cross-corpora (and cross-language) comparison is shown. See [4] for a discussion of cross-linguistic CR similarities.

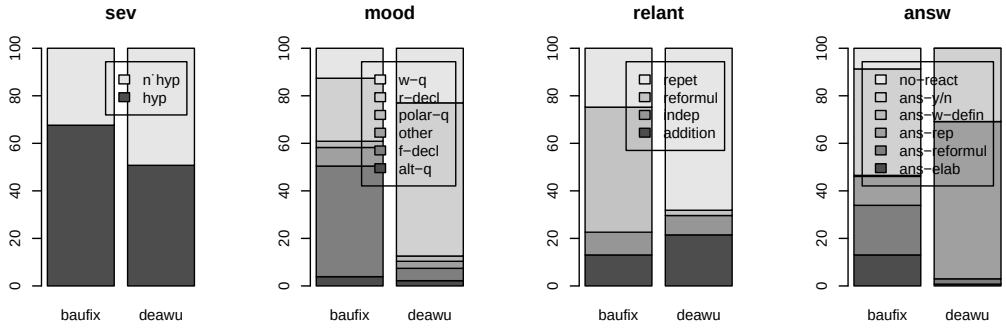

Figure 1: Comparison of CR distributions, [4] and our corpus

wh-questions, and much fewer falling declaratives. Our CRs are much more often repetitions of their antecedents, and much more often is the CR reply a repetition; conversely, in our corpus there were very few reformulations (in either CR or reply). The differences in the shown features are all significant (χ^2, $p = 0.001$). Differences in `extent` (hypothesis presented or not) were *not* significant.

5. Discussion and Conclusions

As our results show, different item types differ in robustness against damage. However, it is not quite "the more context, the more robust"; it's more like those items where the IG expects problems in any case (numbers and modified idioms) have *a higher base-line level of effort and hence are more robust*. The relevance of the damage also seems to depend on the relevance of the damaged item; i.e., *asking for clarification is not an automatic process*, but rather depends on a judgement on the importance of the missing information. *If possible, CRs are produced that point out the location of the problem and present a hypothesis* (as the correlation between size of the damage and locating devices shows). The comparison to the baufix corpus finally confirmed the hypothesis that *CRs that repeat material, with rising intonation, are predominantly used for clarification of acoustic problems*. (Witness the significant increase in this type due to introduction of noise.)

To boil these observations down to a policy on when and how to clarify: a) decide if material is important; b) if at all possible, present hypothesis, and c) locate problem. If d) the problem was an acoustic one, repeat what was understood, with rising intonation.

In future work we will compare this with other task done in the same session, which has much higher contextual constraints and places much less "load" on individual utterances. It is there that we assume more global strategies for dealing with noise can be found. Also, we are planning to do another set of recordings with the noise level raised from the 10% here to a more disruptive 50%, to study behaviour under extreme limitations.

6. Acknowledgements

Thanks to our students for help with transcription and annotation, to M. Waeltermann for work on the noise program, to J. Dreyer at ZAS Berlin & P. Healey and Gregory Mills at Queen Mary U London for letting us use their labs; and to M. Stede and A. Corradini for early discussions of the set-up. This work was supported by the EU (Marie Curie Programme) and DFG (Emmy Noether Programme).

7. References

[1] M. Purver, J. Ginzburg, and P. Healey, "On the means for clarification in dialogue," in *Proc. of the 2nd SIGdial Workshop on Discourse and Dialogue*, 2001.

[2] M. Purver, P. G. Healey, J. King, J. Ginzburg, and G. J. Mills, "Answering clarification questions," in *Proc. of the 4th SIGdial Workshop on Discourse and Dialogue*, 2003.

[3] D. Schlangen, "Causes and strategies for requesting clarification in dialogue," in *Proc. of the 5th Workshop of the ACL SIG on Discourse and Dialogue*, 2004.

[4] K. J. Rodríguez and D. Schlangen, "Form, intonation and function of clarification requests in german task-oriented spoken dialogues," in *Proc. of Catalog (the 8th workshop on the semantics and pragmatics of dialogue; SemDial04)*, E. Vallduví, Ed., 2004, pp. 101–108.

[5] G. Skantze, "Exploring human error recovery strategies: Implications for spoken dialogue systems," *Speech Communication*, vol. 45, no. 3, pp. 325–341, 2005.

[6] V. Rieser and J. Moore, "Implications for generating clarification requests in task-oriented dialogues," in *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 2005, pp. 239–246.

[7] J. Edlund, D. House, and G. Skantze, "The effects of prosodic features on the interpretation of clarification ellipses," in *Proc. of the 9th European Conference on Speech and Communication Technology*, 2005.

[8] P. G. T. Healey and G. Mills, "Clarifying spatial descriptions: Local and global effects on semantic coordination," in *Proc. of brandial'06, the 10th International Workshop on the Semantics and Pragmatics of Dialogue (SemDial10)*, D. Schlangen and R. Fernández, Eds. Universitätsverlag Potsdam, 2006.

[9] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9–10, pp. 341–345, 2001.

[10] C. Müller and M. Strube, "MMAX: A Tool for the Annotation of Multi-modal Corpora," in *Proc. of the 2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 2001, pp. 45–50.

[11] M. Meteer and A. Taylor, "Dysfluency annotation stylebook for the switchboard corpus," 1995, http://www.cis.upenn.edu/ bies/manuals/DFL-book.pdf.