Experimenting with Clarification in Dialogue

Patrick G.T. Healey[1] **(ph@dcs.qmul.ac.uk),**
Matthew Purver[2] **(matthew.purver@kcl.ac.uk),**
James King[1] **(jking@dcs.qmul.ac.uk),**
Jonathan Ginzburg[2] **(jonathan.ginzburg@kcl.ac.uk),**
Greg J. Mills[1] **(gj@dcs.qmul.ac.uk)**

[1]Department of Computer Science; Queen Mary, University of London,
London E1 4NS.

[2]Department of Computer Science; Kings College London
London WC2R 2LS.

Abstract

A new technique for integrating experimental manipulations into text-based, synchronous dialogue is introduced. This method supports fine-grained, systematic transformation of conversational turns and the introduction of 'artificial' probe turns and turn sequences. It can be used to introduce manipulations that are sensitive to aspects of the local linguistic and conversational context for any task or dialogue type. The use of this technique is illustrated by an experimental investigation of the effect of word category and level of grounding on the interpretation of reprise clarifications. The results show that these factors affect both the type and likelihood of response to reprise fragment clarifications.

Introduction

Empirical analyses of dialogue phenomena have been limited by a lack of techniques that provide adequate experimental control. The most detailed analyses of dialogue have focused on descriptive analyses of corpora of natural conversations (e.g. Schegloff, 1987). Corpus studies are limited in that they provide only retrospective, correlational data that make it difficult to resolve conflicting interpretations of the phenomena. Experimental techniques have been limited to the manipulation of relatively coarse-grained parameters of interaction such as task type, level of participation, or communicative modality (for overviews see Pickering and Garrod, 2003; Clark, 1992).

Development and testing of hypotheses about detailed mechanisms and procedures that sustain dialogue co-ordination has consequently been limited by the indirect nature of the available evidence. Psycholinguistic techniques do not provide general, systematic and fine-grained ways to integrate experimental manipulations into unfolding interactions.

This paper introduces a new technique for carrying out experiments on text-based dialogue which addresses these limitations. The rationale for the approach is set out together with some of its practical limitations. An experiment is reported which uses this approach to investigate the interpretation of clarification requests in dialogue: in particular, the influence of word type (content vs. function) and level of grounding (first vs. second mention) on interpretation. Further potential applications of the approach are discussed in the context of the experimental results.

Manipulating Chat Interactions

The experimental technique presented in this paper draws on two general developments. Firstly, the increasing use of text-based forms of synchronous conversational interaction, for example: chat rooms (MUD's, MOO's etc.) and instant messaging. Secondly, advances in natural language processing technology which make some forms of parsing and transformation fast enough to be performed on a time scale consistent with exchanges of turns in synchronous text chat.

The basic paradigm involves pairs of subjects, seated in different rooms, communicating using a synchronous text chat tool (see figure 1 for an example). However, instead of passing turns directly to the appropriate chat clients, each turn is routed via a server. The server is used to systematically modify turns in a variety of ways determined by the goals of the experiment. For example, simple forms of miscommunication can be introduced into an interaction by transforming the order of characters in some of the input words or by substituting words with plausible non-words. Importantly, the server controls which modifications are transmitted to which participant. So, if participant A types the word "table" the sever can echo back `A: table` to participant A and a transformed version, say, "blate" to participant B who sees `A: blate`. The ability to set up controlled asymmetries of this kind between the participants in a interaction creates a powerful range of experimental possibilities. Here, we describe an application of this technique to the investigation of reprise clarification requests (CR's).

Request for Clarification

Requests for clarification are critical for maintaining mutual-understanding in dialogue and have received attention from both the formal semantic (e.g. Ginzburg and Cooper, 2001, 2003) and conversation analytic traditions (e.g. Schegloff, 1987). Clarification requests (CRs) can take a variety of forms.

Some CRs explicitly identify the clarification required, e.g., "What did you say?" or "What do you mean?". Others are more elliptical and involve repetition of only parts of the problem utterance. The most elliptical forms of CR are reprise fragments (RFs) which occur where part of the problem utterance, possibly a single word, is repeated without modification as in Excerpts 1[1] and 2[2] (taken from the British National Corpus (BNC) (Burnard, 2000)).

Lara:	There's only two people in the class.
Matthew:	**Two people?**
Unknown:	For cookery, yeah.

Excerpt 1: Example Reprise Fragment CR

Laura:	Can I have some toast please?
Jan:	**Some?**
Laura:	Toast.

Excerpt 2: Example Reprise Fragment CR

RFs account for approximately 30% of CRs in natural conversation (Purver et al., 2002) and are interesting partly because of their ambiguity. Although they can efficiently localise where a problem occurs they do not explicitly signal what problem the recipient has encountered. Purver, et al. (2002) distinguish between three main readings of RFs: Clausal, Constituent, and Lexical. A clausal reading treats a CR as asking about the content of the conversational move that prompted the CR. It can be roughly paraphrased as *"Is it X about which you are asking/asserting Y?"*. The constituent reading queries the content of a constituent of the problem turn and can be paraphrased as *"What/Who is X?"* or *"What/who do you mean by X?"*. The lexical reading is similar to the clausal reading except that it is an aspect of the surface form, not the content of the conversational move, that is is queried. This corresponds to *"Did you utter X?"*.

Purver et al. (2002) carried out an analysis of the BNC corpus of conversations to investigate the relationship between the different forms of CR and the readings they are given. Their findings indicate that, in contrast to other forms of reprise clarification, RFs can receive each of the possible readings. However, the corpus data show a strong preference for a Clausal reading (87% of cases) over the other forms.

As noted above, although corpus studies of this kind provide valuable information about the distribution of different CR forms and readings, they do not provide tests of the conditions which prompt particular readings. Intuitively, at least two factors would be expected to affect the type of reading assigned to a RF; word category and level of grounding. The linguistic category of the reprised word should influence expectations about what is being clarified. For example, reprise of a content word (e.g. noun or verb) should be more likely to signal a 'constituent' problem than a reprise of a function word (e.g. preposition or determiner). Dialogue participants would normally assume that the meaning of function words is well known in a particular linguistic community and that, as a result, a reprise of a function word is more likely to signal clausal or lexical problems. The interpretation of a RF should also depend on whether a reprised fragment is already considered to have been grounded by the participants in a conversation. For example, a reprise of a proper noun would be more likely to be read as signalling a constituent problem if it occurs on the first mention than on second mention. All things being equal, the content of a constituent is already considered to be established by the time a second mention occurs.

Reprise Fragment Experiment

A chat-tool experiment was designed to test the following hypotheses:

1. RFs for function words will normally receive clausal readings, whereas both clausal and constituent readings will be available for content words.

2. RFs for content words will receive more constituent readings on first mention than on second mention.

3. No difference is predicted for RFs for function words on first vs. second mention.

Method

Two tasks were used to elicit dialogue, a balloon debate and a story-telling task. In the balloon debate subjects are presented with a fictional scenario in which a balloon is losing altitude and about to crash. The only way for any of three passengers to survive is for one of them to jump to a certain death. The three passengers are; Dr. Nick Riviera, a cancer scientist, Mrs. Susie Derkins, a pregnant primary school teacher, and Mr. Tom Derkins, the balloon pilot and Susie's husband. Subjects are asked to decide who should jump. The advantages of this task are that it is effective at generating debates between subjects and involves repeated references to particular individuals.

The second dialogue task, from Bavelas et al. (1992), is the story-telling task. In this case subjects are asked to relate a 'near-miss' story about

[1] BNC file KPP, sentences 352–354

[2] BNC file KD7, sentences 392–394

some experience in which something bad almost happened but in the end everything was okay. The advantage of this task is the topic of the exchange is unrestricted, in effect a random factor, and the interaction relates to real events.

Subjects

Twenty-eight subjects were recruited, 20 male and 8 female, average age 19 years, from computer science and IT undergraduate students. They were recruited in pairs to ensure that the members of a pair were familiar with one another and only subjects who had experience with some form of text chat such as chat rooms, IRC, ICQ or other messaging systems were used. Each subject was paid at a rate of £7.50 per hour for participating in the experiment.

Materials

A custom experimental chat tool, written in Java and Perl, was used for the experiment. The user interface is similar to instant messaging applications: a lower window is used to enter text, and the conversation is displayed in the main upper window as it emerges (see Figure 1). The chat clients were run on two Fujitsu LCD tablet computers with text input via standard external keyboards, with the server running on a standard PC in a separate room.

User Interface The Chattool client user interface is written in Java. The application window is split into two panes: a lower pane for text entry and an upper pane in which the conversation is displayed.

A status display between the two panes shows whether the other participant is active (typing) at any time. This can be manipulated during the generation of artificial turns to make it appear as if they are generated by the other participant. The client also has the ability to display an error message and prevent text entry: this can be used to delay one participant while the other is engaged in an artificially-generated turn sequence.

Server Each turn is submitted to a server (also written in Java) on a separate machine when a ‘Send’ button or the ‘Return’ key is pressed. This server passes the text to a NLP component for processing and possible transformation, and then displays the original version to the originator client, and the processed (or artificially generated) version to the other client. The server records all turns, together with each key press from both clients, for later analysis. This data is also used to dynamically control the speed and capitalisation of artificially generated turns, to be as realistic a simulation of the relevant subject as possible.

NLP Component The NLP component consists of a Perl text-processing module which communicates with various external NLP modules as required: part-of-speech tagging can be performed using LTPOS (Mikheev, 1997), word rarity/frequency tagging using a custom tagger based on the BNC (Kilgarriff, 1997), and synonym generation using WordNet (Fellbaum, 1998).

Experimental parameters are specified as a set of rules which are applied to each word in turn. Preconditions for the application of the rule can be specified in terms of part-of-speech, word frequency and the word itself, together with contextual factors such as the time since the last artificial turn was generated, and a probability threshold to prevent behaviour appearing too regular. The effect of the rule can be to transform the word in question (by substitution with another word, a synonym or a randomly generated non-word, or by letter order scrambling) or to trigger an artificially generated turn sequence (currently a reprise fragment, followed by an acknowledgement, although other turn types are possible).

The current experimental setup consists of rules which generate pairs of RFs and subsequent acknowledgements[3], for proper nouns, common nouns, verbs, determiners and prepositions, with probabilities determined during a pilot experiment to give reasonable numbers of RFs per subject. No use is made of word rarity or synonyms.

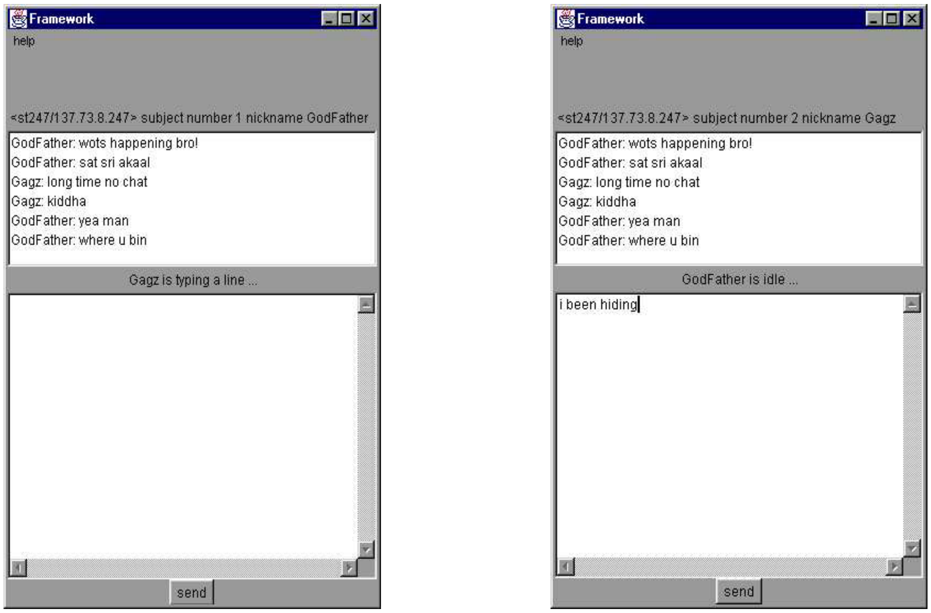
The turn sequences are carried out by (a) presenting the artificially-generated RF to the relevant client only; (b) waiting for a response from that client, preventing the other client from getting too far ahead by locking the interface if necessary; (c) presenting an acknowledgement to that response; and (d) presenting any text typed by the other client during the sequence.

Procedure

Subjects were informed that the experiment was investigating the effects of a network-based chat tool on the way people interact with one another. They were infomed that their interaction would be logged, anonymously, and kept for subsequent analysis. Subjects were advised that they could request the log to be deleted after completion of the interaction and that they were free to leave at any time. They were then given a brief demonstration of the operation of the chat tool.

To prevent concurrent verbal or gestural interaction subjects were seated in separate rooms. Each pair performed both dialogue tasks and were given written instructions in each case. The balloon task was carried out once and the story-telling task twice; one story for each participant. To control for order

[3] Acknowledgements are randomly chosen amongst: “ah”, “oh”, “oh ok”, “right”, “oh right”, “uh huh”, “i see”, “sure”.



<st247/137.73.8.247> subject number 1 nickname GodFather

GodFather: wots happening bro!
GodFather: sat sri akaal
Gagz: long time no chat
Gagz: kiddha
GodFather: yea man
GodFather: where u bin

Gagz is typing a line ...

<st247/137.73.8.247> subject number 2 nickname Gagz

GodFather: wots happening bro!
GodFather: sat sri akaal
Gagz: long time no chat
Gagz: kiddha
GodFather: yea man
GodFather: where u bin

GodFather is idle ...

i been hiding

Figure 1: Chattool Client Interface

Table 1: Story Telling Task Excerpt, Noun Clarification, Subjects 1 & 2

Subject A's View		Subject B's View
B: `Obviously the relatives were coming around like they do to see me`		B: `Obviously the relatives were coming around like they do to see me`
	Probe →	A: `relatives?`
	Block	B: `Yeah just unts and uncles`
	Ack →	A: `ah`
A: `yeah`		A: `yeah`

effects presentation of the two tasks was counterbalanced across pairs. A 10-minute time limit was imposed on both tasks. At the end of the experiment subjects were fully debriefed and the intervention using 'artificial' clarifications was explained to them. This resulted in a within-subjects design with two factors; category of reprise fragment and level of grounding (first vs. second mention).

After the experiment, the logs were manually corrected for the part-of-speech category of the RF and for the first/second mention clarification. Part-of-speech required correction as the tagger produced incorrect word categories in approximately 30% of cases. In some instances this was due to typing errors or text-specific conventions, such as "k" for "okay", that were not recognised. Detection and classification of proper nouns was also sensitive to capitalisation. Subjects were not consistent or conventional in their capitalisation of words and this caused some misclassifications. In addition a small proportion of erroneous tags were found. Each system-generated CR was checked and, where appropriate, corrected. Because pairs completed both tasks together CRs classified as 'first mentions' were checked to ensure that they hadn't already occured in a previous dialogue.

Results

In addition to the Clausal, Constituent and Lexical readings introduced above, Purver et al. (2002) identify three other possible interpretations of reprise fragment clarifications: 'Gap', 'Correction' and 'Non-clarificational'. Gaps occur where the fragment reprised is not the one about which clarification is actually being requested, but the one immediately preceding it. For example, in Excerpt 2, the reprised word is "some" but the clarification is of the following word – "toast". Corrections occur where the fragment is offered as a correction and can be paraphrased as *"Did you mean to say X?"*. 'Non-clarificational' refers to situations in which the fragment is treated as something other than a CR. In the

Table 2: Balloon Task Excerpt, Verb Clarification, Subjects 3 & 4

Subject A's View		Subject B's View
A: `so we agree`		A: `so we agree`
B: `agree?`	← Probe	
A: `yeah to chuck out Susie derkins`	Block	
B: `uh huh`	← Ack	
B: `yes`		B: `yes`

present corpus, gap, lexical and non-clarificational readings were low frequency events (4, 1 and 8 instances respectively) and no instances of correction readings were noted. These figures are comparable with Purver et al.'s (*ibid.*) observations for the BNC. For statistical analysis these three catergories together with explicit requests for clarification of the CR were were grouped as 'Other'.

Across the corpus as a whole a total of 215 system-generated RFs were produced. In 50% of cases the system-generated clarification received no response from the target participant. This is discussed below.

Table 3: Frequency of Reading Types By RF Category and Mention

	Response Category			
Category	None	Con	Cla	Other
Cont (1st)	29	14	23	4
Cont (2nd)	43	7	16	9
Func (1st)	6	0	0	6
Func (2nd)	20	1	0	9

The distribution of reading types according to word categrory was tested firstly by comparing the frequency of Clausal, Constituent, and Other readings for content words and function words. This proved to be reliably different ($\chi^2_{(2)}$ = 35.3, p = 0.00).[4] As Table 3 shows, RFs of Function words were almost exclusively given Other readings i.e., either they were explicitly queried indicating they could not be interpreted, or they were interpreted as Gap, Lexical or Non-clarificational. By contrast Content word reprises were interpreted as Clausal CRs 53% of the time, as Constituent CRs 29% of the time and as Other 18% of the time.

Content word and Function word clarifications were also compared for the the frequency with which they received a response. This showed no reliable difference ($\chi^2_{(1)}$ = 1.95, p = 0.16) indicating that although the interpretations given to Content and Function CR's are different they are equally likely to receive some kind of response.

The influence of grounding on reading type was assessed firstly by comparing the relative frequency of Constituent, Clausal and Other readings on first and second mention. This was reliably different ($\chi^2_{(2)}$ = 6.28, p = 0.04) indicating that level of grounding affects the reading assigned. A focussed comparison of Constituent and Clausal readings on first and second mention shows no reliable difference ($\chi^2_{(1)}$ = 0.0, p = 0.92). Together these findings indicate that, across all word categories, Constituent and Clausal readings are more likely for CR's of a first mention than a second mention and, conversely, Other readings are less likely for CR's to a first mention than a second mention.

The effect of grounding on the relative frequency with which a CR received a response was also tested. This showed an effect of mention ($\chi^2_{(1)}$ = 3.87, p = 0.05); 56% of reprise clarifications of first mentions received a response whereas only 43% of second mention clarifications did.

Discussion

The experimental results indicate that people's interpretation of reprise fragment CR's is influenced both by the category of the reprise fragment and its level of grounding.

One concern that arises with these results is whether they represent an artifact of differences between text and utterances as media or whether they bear on more basic aspects of the use of CR's in interaction. In contrast to utterances, text-chat turns have no intonation, they take longer to produce, are normally produced in overlap, and persist for longer. Turns can also get out of sequence since users may still be responding to a prior turn when a new turn arrives. In some cases we observed that the response to a clarification was displaced to the end of the turn in progress or to a subsequent turn.

One respect in which this feeds into the present study is that persistence makes a Lexical reading of a CR less plausible since participants can still see what word was used in a previous turn. In the BNC corpus Lexical readings of reprise fragment CR's account for 3% of the sample analysed by Purver et. al. (2002). In the present experimental corpus we found only one instance of a lexical reading (0.004%). Me-

[4] A criterion level of p < 0.05 was adopted for all statistical tests.

dia differences may also contribute to differences in the distribution of clausal and constituent readings. In the BNC reprise fragments content words receive Clausal readings in 81% of cases, and constituent readings in 6% of cases. In the experimental corpus they receive Clausal readings in 53% of cases and Constituent readings in 29% of cases.

However, the finding that almost 50% of the experimental CRs are ignored does not seem to be attributable to differences between text-based and verbal interaction. Reprise fragments are also often ignored in verbal exchanges. For the sample of reprise fragments analysed by Purver et. al (2002) only 56% receive a clear answer. For 5% of those that do not receive a clear answer the transcription of the next turn is not clear enough to determine whether a response occurred or not. For the remainder no response is recorded in the transcript. In some of these cases there will have been a non-verbal response and these are not transcribed in the BNC. However, not all CR's could be resolved in this way. On balance it seems that even in face-to-face interaction a significant proportion of reprise fragment CR's receive no direct response.

Perhaps more importantly, the experimental results show a reliable difference in the frequency of responses to CR's for first and second mentions of a word. This indicates that the CR's are not just being missed, the recipients of the CR's sometimes choose not to address them. In addition, there is a reliable difference in the profile of reading types for CR's on first and second mention with a shift away from clausal and constituent readings toward Other readings. CR's for second mentions are thus more likely to be either ignored, explicitly queried or treated as doing something other than just clarifying the reprised fragment.

It appears that participant's responses to reprise fragment CR's reflect a trade-off between the effort required to diagnose a problem and the risk to mutual-understanding of carrying on without addressing it.

This work demonstrates the viability of investigating dialogue co-ordination through the manipulation of chat-tool based interactions. This technique supports task independent, systematic and fine grained experimental interventions in interaction. It was successful in producing plausible clarification sequences and although some artificial clarifications were difficult to interpret this is also true of genuine CR's from other participants. When questioned during debriefing, no participants reported any suspicions about the experimental manipulation. The main practical difficulties encountered in the present study related to `txt` conventions such as novel spellings, abbreviations, and use of 'smileys' and typing errors and inconsistency in spelling and capitalisation.

The experiment presented here exploits only one possible use of this technique. Amongst other things potential manipulations include; distance, in turns or time, between target and probe, substitution of synonyms, hyponyms and hypernyms, introduction of artifical turns, blocking of certain forms of response. The important potential it carries, in comparison with existing techniques, is in the direct testing of claims about the fine-grained mechanisms of dialogue co-ordination.

Acknowledgments

This work was supported by the EPSRC under the project "ROSSINI: Role of Surface Structural Information in Dialogue" (GR/R04942/01).

References

Bavelas, J., Chovil, N., Lawrie, D., and Wade, L. (1992). Interactive gestures. *Discourse Processes*, 15:469–489.

Burnard, L. (2000). *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

Clark, H. H. (1992). *Arenas of Language Use*. University of Chicago Press & CSLI.

Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Ginzburg, J. and Cooper, R. (2001). Resolving ellipsis in clarification. In *Proceedings of the 39th Meeting of the ACL*, pages 236–243. Association for Computational Linguistics.

Ginzburg, J. and Cooper, R. (2003). Clarification, ellipsis, and the nature of contextual updates. *Linguistics and Philosophy*, forthcoming.

Kilgarriff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2):135–155.

Mikheev, A. (1997). Automatic rule induction for unknown word guessing. *Computational Linguistics*, 23(3):405–423.

Pickering, M. and Garrod, S. (2003). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, forthcoming.

Purver, M., Ginzburg, J., and Healey, P. (2002). On the means for clarification in dialogue. In Smith, R. and van Kuppevelt, J., editors, *Current and New Directions in Discourse & Dialogue*. Kluwer Academic Publishers.

Schegloff, E. (1987). Some sources of misunderstanding in talk-in-interaction. *Linguistics*, 25:201–218.