Answering Clarification Questions

Matthew Purver[1], **Patrick G.T. Healey**[2], **James King**[2], **Jonathan Ginzburg**[1] and **Greg J. Mills**[2]

[1]**Department of Computer Science**
King's College, London
London WC2R 2LS, UK

[2]**Department of Computer Science**
Queen Mary, University of London
London E1 4NS, UK

Abstract

This paper describes the results of corpus and experimental investigation into the factors that affect the way clarification questions in dialogue are interpreted, and the way they are responded to. We present some results from an investigation using the BNC which show some general correlations between clarification request type, likelihood of answering, answer type and distance between question and answer. We then describe a new experimental technique for integrating manipulations into text-based synchronous dialogue, and give more specific results concerning the effect of word category and level of grounding on interpretation and response type.

1 Introduction

Requesting clarification is a vital part of the communicative process and has received attention from both the formal semantic (Ginzburg and Cooper, 2001; Ginzburg and Cooper, forthcoming) and conversation analytic traditions (Schegloff, 1987), but little in the computational dialogue system community. In theory, a perfect dialogue system should be able to interpret and deal with clarification requests (CRs) made by the user in order to elicit clarification of some part of a system utterance, and be able to request clarification itelf of some part of a user utterance. This is no easy task – CRs may take many different *forms* (often highly elliptical), and can be intended to be interpreted with many different *readings* which query different aspects of the original utterance. As a result, dialogue system design has traditionally attempted to avoid the necessity for CR interpretation by making system utterances as clear and precise as possible, and avoid having to generate all but the most simple CRs by using robust shallow methods of interpretation or by relying on highly domain-dependent lexicons and grammars. However, as systems become more human-like, it seems likely that we will have to cope with user CRs at some stage; and the ability to generate system CRs can be useful in order to repair misunderstanding, disambiguate other utterances, and learn new words – see (Knight, 1996; Dusan and Flanagan, 2002; Purver, 2002).

The investigations presented here had two main aims: to examine (a) how CRs are interpreted, and (b) how they are responded to. The two are clearly dependent – the response must depend on the interpretation – but there are many other influencing factors such as CR form, context and level of grounding. Answers to (a) should help us with the following questions:

- What factors can help us disambiguate and correctly interpret user CRs?
- What factors should govern generation of system CRs such that they are correctly interpreted by the user?

Answers to (b) should help with the following related questions:

- How (and when) should we answer user CRs?
- How (and when) should we expect users to respond to system CRs?

The paper is organised as follows. The next section gives a brief overview of CRs in general and some previous corpus work. Section 3 describes further corpus work which gives some general results concerning response type. Section 4 then describes a text-based dialogue experiment examining the detailed effects on interpretation and response of part-of-speech (PoS) type and level of grounding for one particular CR form, and section 5 then draws some general conclusions.

2 Clarification Requests

Purver et al. (2001; 2002) presented a taxonomy of CR forms and readings derived from a corpus study using the British National Corpus (BNC) – see (Burnard, 2000). This showed that some forms showed a high correlation with certain readings, but that some were highly ambiguous.

Purver et al. (2002)'s taxonomy of CR forms is given in table 1 and CR readings in table 2[1]. Some CRs (the non-reprise class) explicitly identify the clarification required, e.g. *"What did you say?"* or *"What do you mean?"*, and some forms (e.g. literal reprises) appear to favour a particular reading almost exclusively, but most are more ambiguous. Indeed, they found that the two most common forms (the conventional and reprise fragment form) could take any reading.

Although this corpus study provided information about the distribution of different CR forms and readings, it did not provide any information about the specific conditions which prompt particular readings and affect how the CR is answered. In this paper we concentrate mostly on the reprise fragment (RF) form, where only a single part of the problem utterance, possibly a single word, is reprised[2] as in example (1). This form is not only common (approximately 30% of CRs in the previous study) and can appear with many readings (although biased towards a clausal reading – 87% of occurrences), but specifies the problematic element that it clarifies quite precisely, and therefore should give us scope for examining the effect of features of that element.

(1)[3]

Gary:	Aye, but <pause> you know <pause> like you se- she mentioned one in particular, like
Jake:	What?
Gary:	the word skeilth
Jake:	**Skeilth?**
Lilias:	Mm.
Gary:	Aha.
Jake:	Aye, yeah, yeah, take skeilth.

Intuitively, at least two such features would be expected to affect the type of reading assigned to a RF: PoS category and level of grounding.[4] The PoS 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. RF interpretation 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 have been established by the time a second mention occurs.

3 Corpus Investigation

Accordingly we have re-examined the corpus from the above study in order to add information about

[1] They also give a *correction* reading, which we have excluded here: such CRs are almost exclusively self-corrections and as such do not fit well with our discussion here. They are also very rare compared with the other classes, making up only about 2% of CRs.

[2] Such reprises need not be verbatim repeats: users may use anaphoric terms or use a clearer expression in order to clarify the fragment in question.

[3] BNC file KPD, sentences 578–584

[4] Another is intonation. However, there is no intonational information in the BNC. In the future we hope to investigate this using other corpora and experimental methods.

Class	Description	Example
non	Non-Reprise	*"What did you say?"*
wot	Conventional	*"Pardon?"*
frg	Reprise Fragment	*"Paris?"*
slu	Reprise Sluice	*"Where?"*
lit	Literal Reprise	*"You want to go to Paris?"*
sub	Wh-Subsituted Reprise	*"You want to go where?"*
gap	Gap	*"You want to go to . . . ?"*
fil	Gap Filler	*". . . Paris?"*
oth	Other	Other

Table 1: CR forms

Class	Description	Paraphrase
cla	Clausal	*"Are you asking/telling me that . . . X . . . ?"*
con	Constituent	*"What/who do you mean by 'X'?"*
lex	Lexical	*"Did you utter 'X'?"*
oth	Other	Other

Table 2: CR readings

category, grounding and method of answering.

3.1 Method

The same corpus was re-marked for four attributes: response type and CR-answer distance, and the PoS and last mention of the original source element.

The markup scheme used for response type evolved during the study and is shown in table 3: it includes classification of apparently unanswered CRs into those that may have been answered, but the sentence possibly containing an answer was transcribed in the BNC as `<unclear>`; those that appear to have remained unanswered because the CR initiator continued their turn without pause; and those that are not answered at all (or at least where we have no indication of an answer – eye contact, head movement etc. are not recorded in the BNC but could function as answers). In cases where the initial response was followed by further information, both were recorded, but the results here are presented only for the initial response. Further work later may take both into account, along the lines of (Hockey et al., 1997) who showed this to be important for questions in general.

CR-answer distance was marked in terms of the sentence numbering scheme in the BNC – in these cases it corresponds very closely to distance in speaker turns, although the correspondence is not exact.

PoS category and time of last mention of the source element were marked, but have not currently been used due to lack of useful data (see below).

Reliability of the markup has not yet been examined. However, the method is close to that of (Purver et al., 2002) (and the corpus is identical), where reliability was examined and found to be acceptable. We then examined the correlation between CR type and response type, between reading and response type, and the spread of CR-answer distance.

3.2 Results

3.2.1 Response Type

Results for response type are shown in table 4 as raw numbers, and also in table 5 as percentages for each CR type, with the *none*, *cont*, *uncl* and *qury* classes conflated as one "unanswered" class, and only the most common 4 CR forms shown.

The most striking result is perhaps the high overall number of CRs that do not receive an answer: 39% of all CRs do not appear to be answered overall, although this reduces to 17% when taking account of those marked *uncl* (possible answers transcribed

none	No answer
cont	CR initiator continues immediately
uncl	Possible answer but transcribed as <unclear>
qury	CR explicitly queried
frg	Answered with parallel fragment
sent	Answered with full sentence
yn	Answered with polar particle

Table 3: CR response types

as <unclear>) and *cont* (the CR-raiser continues without waiting). The most common forms (conventional and RF) appear to be answered least – around 45% go unanswered for both. The form which appears to be most likely to be answered overall is the explicit non-conventional form.

Some forms appear to have high correlations with particular response types. As might be expected, sluices (which are wh-questions) are generally answered with fragments, and never with a polar yes/no answer. Yes/no answers also seem to be unsuitable for the conventional CR form, which is generally answered with a full sentence. RFs, conversely, are not often answered with full sentences, but can be responded to either by fragments or yes/no answers.

Similarly, from tables 6 and 7 (again, percentages given for each CR reading, with "unanswered" response types conflated and only the most common 3 readings shown) we can see that there is a correlation between reading and response type, but that this correlation is also not as simple as a direct reading-answer correspondence. Clausal CRs are unlikely to be answered with full sentences, but can get either fragment or yes/no responses. Constituent CRs are less likely to get yes/no responses but could get either other type. Interestingly, constituent CRs seem to be roughly twice as likely to get a response as clausal or lexical CRs (even though there are fewer examples of constituent CRs than the others, this difference is statistically significant, with a $\chi^2_{(1)}$ test showing $<$0.5% probability of independence).

3.2.2 Answer Distance

Results for CR-answer distance are shown in table 8. It is clear that the vast majority (94%) of CRs that are answered are answered in the immediately following sentence, and that none are left longer than 3 sentences. While we do not yet have concrete equivalent figures for non-clarificational questions, a study is in progress and initial indications are that in general, answers are less immediate: only about 70% have distance 1, with some up to distance 6.[5]

We therefore expect that (a) answering user CRs must be done immediately, and that any dialogue management scheme must take this into account, and (b) we should expect answers to any system CRs to come immediately – interpretation routines (we are thinking especially of any ellipsis resolution routines here) should not assume that later turns are

	unans	frg	sent	yn	
wot	45.6	8.7	44.8	0.8	(100)
frg	43.2	21.1	3.4	32.2	(100)
slu	37.0	50.0	12.9	0	(100)
non	13.4	26.9	26.9	32.6	(100)

Table 5: BNC results: Response type as percentages for each CR form

	unans	frg	sent	yn	
cla	39.8	22.2	7.8	30.0	(100)
con	20.0	35.0	33.3	11.6	(100)
lex	42.7	17.2	36.5	3.4	(100)

Table 7: BNC results: Response type as percentages for each CR reading

	1	2	3	$>$3	Total
Distance	273	14	2	0	289

Table 8: CR-answer distance (sentences)

[5]Thanks to Raquel Fernández for providing us with these preliminary figures.

	none	cont	uncl	qury	frg	sent	yn	Total
wot	21	13	24	0	11	57	1	127
frg	23	22	6	0	25	4	38	118
slu	8	6	5	1	27	7	0	54
non	4	2	1	0	14	14	17	52
lit	5	2	1	0	1	1	10	20
fil	3	0	1	0	7	1	4	16
sub	4	0	3	0	4	4	0	15
gap	1	0	0	0	1	0	0	2
oth	0	0	0	1	0	1	0	2
Total	69	45	41	2	90	89	70	406

Table 4: BNC results: Response type vs. CR form

	none	cont	uncl	qury	frg	sent	yn	Total
cla	33	31	11	2	43	15	58	193
con	9	3	0	0	21	20	7	60
lex	21	11	30	0	25	53	5	145
oth	5	0	0	0	0	1	0	6
Total	69	45	41	2	90	89	70	406

Table 6: BNC results: Response type vs. CR reading

relevant to the CR.

3.2.3 Further Details

While interesting, we would like to know more detail than the general trends described above: in particular we would like to know the effect of the factors we have mentioned (word category and grounding) for particular forms. As stated above, we concentrate here on the reprise fragment form.

Examination of original CR source fragment PoS category, in order to test the effect of the content/function distinction, showed that almost all RFs were of content words or whole phrases: only 6 of 118 RFs were of function words, all of which were determiners (mostly numbers). This is interesting in itself: perhaps RFs are unlikely to be used to clarify uses of e.g. prepositions. However, the effect may be due to lack of data, and does not provide us with any way of testing the distinction between clausal and constituent reading that we expect.

Markup of last mention of the original source fragment has also not given results in which we can be confident. For RFs, we have seen that all constituent readings occur on the first mention of the fragment (as expected) – but there are too few of these examples to draw any firm conclusions. It is also impossible to know whether first mention in the transcription is really the first mention between the participants: we do not know what happened before the tape was turned on, what their shared history is, or what is said during the frequent portions marked as `<unclear>`.

So we need more information than our current corpus can provide. In order to examine these effects properly we have therefore designed an experimental technique to allow dialogues to be manipulated directly, with reprises with the desired properties automatically introduced into the conversation. The next section describes this technique and the experiment performed.

4 Experimental Work

Empirical analyses of dialogue phenomena have typically focused either on detailed descriptive analyses of corpora of conversations (Schegloff, 1987) or on the experimental manipulation of relatively global parameters of interaction such as task type or communicative modality (Clark and Wilkes-Gibbs,

1986), (Garrod and Doherty, 1994). These studies have been used to to motivate a variety of proposals about turn-level mechanisms and procedures that sustain dialogue co-ordination. Further development and testing of these proposals has, however, been limited by the indirect nature of the available evidence. Corpus studies provide, retrospective, correlational data which is susceptible to challenge and re-interpretation. Current psycholinguistic techniques do not provide ways of integrating experimental manipulations into interactions in a manner that is sensitive to the linguistic and conversational context. This section introduces a technique for carrying out experiments in which text-based interactions can be directly manipulated at the turn level, and gives the results of an experiment which uses this approach to investigate the effects of the factors mentioned above on interpretation and reponse to RFs. We also briefly discuss the range of potential applications and some of the practical limitations of the approach in the context of the experimental results.

4.1 Manipulating 'Chat' Interactions

The experimental technique presented here 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.), instant messaging, and some online conferencing tools. Secondly, advances in natural language processing technology which make some forms of text processing 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 each completed turn directly to the appropriate chat clients, each turn is routed via a server. Depending on the specific goals of the experiment, the server can be used to systematically modify turns in a variety of ways. For example, some simple forms of mis-communication 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 broadcast 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).

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.

4.2 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.

Following (Bavelas et al., 1992), the second dialogue task used was the story-telling task. In this case subjects are asked to relate a 'near-miss' story about some experience in which something bad almost happened but in the end everything was okay. This was chosen because, unlike the balloon task, the topic of the exchange is unrestricted, in effect a random factor, and the interaction relates to real events.

4.2.1 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.

4.2.2 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 and is designed to be familiar to subjects experienced with instant messaging/chat applications. The application window is split into two panes: a lower pane for text entry and an upper pane in which the conversation is displayed (see figure 1). A status display between the two panes shows whether the other participant is active (typing) at any time. This can be artificially controlled 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 on the fly to 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: PoS tagging can be performed using LT-POS (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 PoS, 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[6], 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.

4.2.3 Procedure

Prior to taking part subjects were informed that the experimenters were carrying out a study of the effects of a network-based chat tool on the way peo-

[6] Acknowledgements are randomly chosen amongst: "ah", "oh", "oh ok", "right", "oh right", "uh huh", "i see", "sure".

ple interact with one another. They were told that their interaction would be logged, anonymously, and kept for subsequent analysis. Subjects were advised that they could also request the log to be deleted after completion of the interaction. They were not informed of the artificial interventions until afterwards (see below).

At the start of the experiment subjects were 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 effects the order of 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 PoS category of the RF and for the first/second mention clarification. PoS 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.

4.3 Results

The readings attributed to each RF were classified in the same way as the original BNC-based corpus, with the addition of one further category: non-clarificational, referring to situations in which the fragment is treated as something other than a CR (this did not apply when building the original corpus, as only utterances treated as CRs were considered). In the experimental results, 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., 2002)'s observations for the BNC. For statistical analysis these three categories of reading were grouped together 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 may be due in part to the medium: unlike verbal exchanges, participants in text-chat can produce their turns simultaneously. This can result in turns getting out of sequence since users may still be responding to a prior turn when a new turn arrives. Users must then trade off the cost of undoing their turn in progress to respond to the new one, against going ahead anyway and responding to the new turn later if it seems necessary. Thus 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. However, comparison with the BNC results from section 3 above show similar figures: only 56% of the `frg` class received a clear answer. Although the true figure will be higher (of the 56%, 5% may have been answered, but the next turn was transcribed as `<unclear>`, and we cannot know in how many cases the reprise may have been answered using non-verbal signals), it seems likely that a significant proportion may simply be ignored.

	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	0	1	9

Table 9: Frequency of Reading Types By RF Category and Mention

The distribution of reading types according to word category 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).[7] As table 9 shows, RFs of Function words were almost exclusively interpreted as Other, i.e. either 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 pattern of interpretation for Content and Function reprises is 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 RF's of a first mention than a second mention and, conversely, Other readings are less likely for RF's to a first mention than a second mention.

The effect of grounding on the relative frequency with which a clarification received a response was also tested. This indicated a strong effect of mention ($\chi^2_{(1)}$ = 12.01, p = 0.00); 58% of reprise clarifications of first mentions recieved a response whereas only 33% of second mention clarifications did.

4.4 Discussion

The experimental results support two basic conclusions. Firstly, people's interpretation of the type of CR a reprise fragment is intended to make is influenced both by the category of the reprise fragment and its level of grounding. Secondly, reprise fragment CRs to first mentions are much more likely to be responded to than reprise fragment CRs for second mentions.

Text-based and verbal interaction have different properties as communicative media. Amongst other things, in text-chat turns take longer to produce, are normally produced in overlap, and they persist for longer. However, even given these differences, the general pattern of clarifications observed in the experimental task is similar to that noted in verbal dialogue. In particular, Lexical, Gap and Non-clarificational readings are infrequent and reprise fragment clarifications are ignored with surprising frequency. In the present data, the clearest contrast between text-based and verbal interaction is in the relative frequency of Constituent and Clausal readings. In the BNC reprise fragments receive Clausal readings in 87% of cases, and constituent readings in 6% of cases. In the experimental corpus they receive Clausal readings in 48% of cases and Constituent readings in 34% of cases.

These findings demonstrate the viability, and some limitations, of investigating dialogue coordination through the manipulation of chat-tool based interactions. The chat tool was successful in producing plausible clarification sequences. Although in some cases participants had difficulty making sense of the artificial clarifications this did not make them distinguishable from other, real, but equally problematic turns from other participants. The clarifications were mostly successful in creating realistic exchanges such as those illustrated in figures 2 and 3. When questioned during debriefing, no participants reported any suspicions about the experimental manipulation.

The main practical difficulty encountered in the present study related to text-chat conventions such as novel spellings, abbreviations, and use of 'smileys'. This created specific problems for the PoS tagger which assumes a more standard form of English. These problems were also compounded by the noise introduced by typing errors and inconsistency in spelling and capitalisation.

The experiment presented here exploits only one possibility for the use of this technique. Other prossible manipulations include; manipulation of 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

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

it carries, particularly in comparison with corpus-based techniques, is in the investigation of dialogue phenomena which for various reasons are infrequent in existing corpora.

5 Conclusions

The main conclusions we draw from the results presented here are as follows:

- Reprise CRs appear to go without response far more often than might be expected, both in the BNC and in our experimental corpus. Both may be effects of the media (transcription in one case, turn sequencing overlap in the other), but the figures are large enough and similar enough to warrant further investigation.
- Corpus investigation shows some strong correlations between CR form and expected answer type. It also shows that responses to CRs, when they come, come immediately.
- Both word PoS category and first/second mention appear to be reliable indicators of RF reading. This can help us in disambiguating user CRs, and in choosing forms when generating system CRs.
- RFs generated on the first mention of a word have a higher likelihood of receiving a response than on second mention.
- We have presented a new experimental technique for manipulating dialogue, which we believe has many potential uses in dialogue research.

6 Acknowledgments

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

References

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

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

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

Sorin Dusan and James Flanagan. 2002. Adaptive dialog based upon multimodal language acquisition. In *Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces*, Pittsburgh, October.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Simon Garrod and Gwyneth Doherty. 1994. Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, 53:181–215.

Jonathan Ginzburg and Robin Cooper. 2001. Resolving ellipsis in clarification. In *Proceedings of the 39th Meeting of the ACL*, pages 236–243. Association for Computational Linguistics, July.

Jonathan Ginzburg and Robin Cooper. forthcoming. Clarification, ellipsis, and the nature of contextual updates. *Linguistics and Philosophy*.

Beth Ann Hockey, Deborah Rossen-Knill, Beverly Spejewski, Matthew Stone, and Stephen Isard. 1997. Can you predict answers to Yes/No questions? Yes, No and Stuff. In *Proceedings of Eurospeech '97*.

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

Kevin Knight. 1996. Learning word meanings by instruction. In *Proceedings of the Thirteenth National Conference on Artifical Intelligence*, pages 447–454. AAAI/IAAI.

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

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2001. On the means for clarification in dialogue. In *Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue*, pages 116–125. Association for Computational Linguistics, September.

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

Matthew Purver. 2002. Processing unknown words in a dialogue system. In *Proceedings of the 3rd ACL SIGdial Workshop on Discourse and Dialogue*, pages 174–183. Association for Computational Linguistics, July.

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

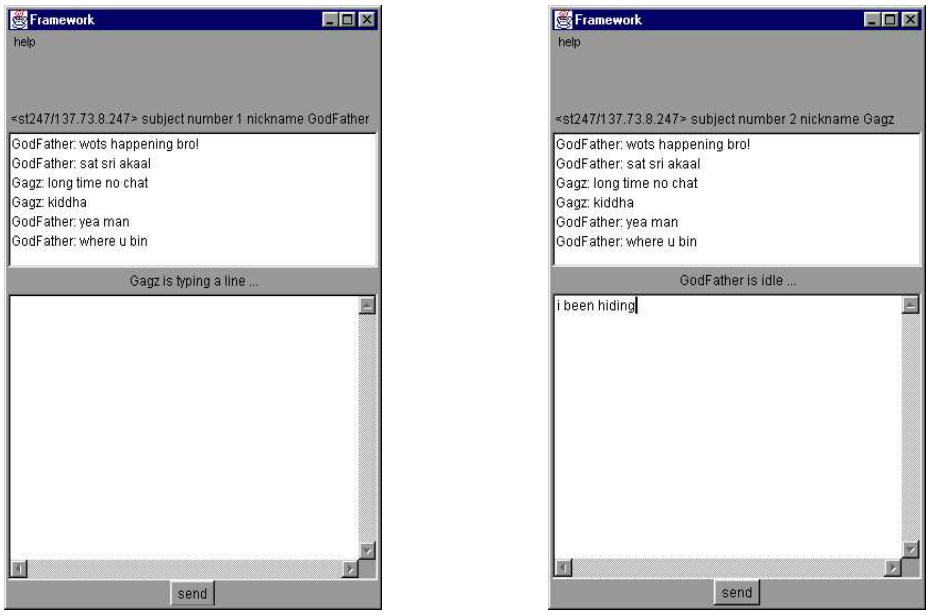

Figure 1: Chattool Client Interface

Subject A's View		Subject B's View
A: 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		B: yeah

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

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

Figure 3: Balloon Task Excerpt, Verb Clarification, Subjects 3 & 4