Evaluating Innovative Multimedia Customer Handling Systems

—

George Barrett Smith

—

Thesis

submitted to the

University of Glasgow

Department of Psychology

for the degree of

Doctor of Philosophy

—

February 2000

Abstract

The subject of this thesis is tripartite communication between two human participants and a computer, where one human has access to the computer but the other does not. Its goal is to show that there is an intricate, triadic, relationship between the three.

Previous investigations of such communication which have indicated that there are triadic relationships have concentrated on in depth study of the dialogues between the two human participants but have been informal and typically ethnomethodological. Formal or controlled experimental investigations have not taken account of the human-human dialogue, and as a consequence have not recognised the communications' triadic nature.

This thesis redresses this dichotomy through extensive analysis of controlled experimental studies of simulated tripartite customer service situations. These analyses cover the three measures of perception, performance and process which have been used in the study of mediated human-human communication to build a complete model of the interactions. Particular attention was paid to the process of the human-human dialogues, using an adaptation of the Conversational Games Analysis (Kowtko, Isard and Doherty-Sneddon, 1992) coding scheme to gather in depth information about the ways in which participants used their utterances.

The analyses of these measures provided empirical evidence that the informal studies were correct in proposing the relationship between the three participants in these communications were triadic. It was shown that not only broad differences in the computer's interaction style but also apparently small variations could have major impact on the human-human dialogue, and that the conditions of visibility between the three participants interacted strongly with these variations effects.

Acknowledgements

The people who have contributed to the completion of this thesis are numerous.

My primary debt of gratitude has to go to Anne Anderson, who continued to show faith that what I was doing was worthwhile and would be finished as deadlines slipped and distractions increased. She could never have been so forceful unless she cared.

Not far behind, come my debts to Jim Mullin and Michael Gardner, my secondary supervisors. Jim was able to simplify the difficulties of planning a research programme with a few words and a spreadsheet; Michael provided links to everyone else at B.T., showing me how the work might fit in with their expectations.

My examiners, Angela Sasse and Paddy O'Donnell, come next. I cannot imagine anyone wanting to read this tome, if not coerced. Extra thanks go to Angela for forcing me to break out all the articles I had filed on contextual methods in H.C.I., and finally pay proper attention to what they had to say.

Alison Newlands, who trained me in the use of Conversational Games Analysis, provided invaluable guidance in the methodology of dialogue coding. It is from discussions with her, and the coding of the dialogues in her thesis that the idea for the addition of a task orientation dimension to Conversational Games Analysis arose.

My final academic acknowledgement goes to all the remaining staff and postgraduates in the Multimedia Communication Group. Discussions with them frequently helped reinforce my hunches, quell my cynicism, and remind me what was really important.

Next, thanks has to go to all the people at B.T. without whom this thesis would never have existed. The research was funded by B.T.'s Future Customer Handling Systems group, and the influence of that group's leaders, Mike Ward and Nicola Millard, should be clear to anyone. Thanks also go to the much over-worked Paul Buckley, who provided up to date information on the practical implementation of customer service interfaces, Paul White and Rob Davis, who showed me how agent-customer dialogues fitted into the larger plan of B.T.'s customer service, and the many call centre personnel who assisted in my observations.

The remaining acknowledgements go to my family:

My distractions, Alexei and Libby, who never understood how I could be at home and at work at the same time. Particular thanks go to the then 23 month old Alexei, who woke me on the morning of my viva by slamming his hand on the table by my slumbering head and ordering me:

Daddy. Read the book!

My wife, Kristen, who helped me through the hours, days and weeks when I thought this thesis would never be finished. I promise never to coerce her into reading it.

My parents-in-law, Beverley and Jeff, who reminded me more than once that nobody enjoyed writing their thesis.

Finally, my father, Gordon, who knew how to do everything else.

Table of Contents

—

Index of Examples

Index of Figures

Index of Tables

—

—

Chapter 1

—

Literature Review

—

Tripartite Communication, Evaluation Methods in Human Computer Interaction and Models of Human-Human Dialogue and Mediated Human-Human Dialogue

1.0 Introduction

The aim of the thesis is to show that the design of the computer interface used in a tripartite computer-agent-customer interaction can impact on the dialogue between the agent and customer and that methods other than traditional HCI's will be required to inform the design of such interfaces.

Because there has been relatively little investigation of such tripartite communication, the role of the literature review is slightly unusual. Rather than showing a progression in the literature which leads to the current research, the review will cover the handful of studies of tripartite communication, noting gaps in their methodologies, and the remainder of the review will provide a background from diverse literature with which these gaps can be filled.

The review of studies of tripartite communication (Section 1.1) will show that those studies which have recognised its unique nature have lacked formality in their methodologies and those studies which have used formal methods and controlled studies have ignored its unique nature. Concurrent with this dichotomy, those studies which have recognised its unique nature have dealt with the dialogues between the two human participants whereas those which have not recognised its unique nature have also not considered the human-human dialogues.

To find means to rectify this, reviews will be made of current methods in HCI and of what is known about human-human dialogue.

The section on current methods in HCI (Section 1.2) will supplement what is seen of HCI methods in Section 1.1. It will have three subsections, one of which takes consistency as a topic of investigation to show how HCI uses formal and experimental methods (Section 1.2.1), a second which reviews the study of computers in the context of their use (Section 1.2.2), and the third of which looks at how HCI has used human language as an evaluative tool (Section 1.2.3). Consistency was picked as a focus for the former because it has been recognised that interfaces must be consistent with their users' expectations, and if the presence of two users alters these expectations current models may need to be suitably adjusted. The aim is to evaluate how HCI's current methods can be adopted or developed to aid in the study of tripartite communication.

The section on human-human dialogue (Section 1.3) will first cover basic findings about the structure of human-human dialogue gathered in studies of natural, face-to-face, communication (Section 1.3.1) and then review studies in which human-human dialogues have been mediated with technology (Section 1.3.2). The studies of mediated dialogue both provide examples of how technology can interact with human-human communication and also of how dialogue can be used in the evaluation.

It will be seen that the methods developed in the sociological and psycholinguistic study of dialogue offer more opportunity for direct application in the formal investigation of tripartite communication than HCI's methods.

1.1 Tripartite Communication

Although tripartite communication involving two humans and a computer is ubiquitous, occurring every day in travel agents' and doctors' offices and in an ever increasing amount of cold sales calls by telephone, the unique properties of such dialogues have received scant attention.

Whalen (1995a) and his co-workers have looked at how operators' software influences the structure of calls to emergency services from a sociological point of view. Fitter and Cruikshank (1982) and Greatbatch, Luff, Heath and Campion (1993) have studied the use of computers within general practice doctor-patient consultations, while Scott and Purves (1996) have produced a triadic model of the computer-doctor-patient relationship. Finally, Gray, John and Atwood (Gray, John and Atwood, 1992, 1993; Gray, John, Stuart, Lawrence and Atwood, 1990) studied the use of telephone operator legacy and redesigned workstations using formal human-computer interaction evaluative techniques, while their colleagues Lawrence, Atwood, Dews and Turner (1995) and Muller, Carr, Ashworth, Diekmann, Wharton, Eickstaedt, and Clonts (1995) have supplemented this work with analysis of the dialogues between the operators and customers.

1.1.1 Computers in Emergency Dispatch Facilities

Sociologists have studied the integration of computers into control rooms and emergency dispatch facilities on a number of occasions (e.g. Heath and Luff, 1992a; Martin, Bowers and Wastell, 1997; Whalen, 1995a, 1995b[1]) but only Jack Whalen (1995a) has directly considered the interaction between callers, call takers and their interfaces.

He spent 15 months working at an emergency dispatch facility, and performed detailed single-case analyses[2] of video-taped calls to dispatchers. As a result of this observation and analysis he argued that the call-taker's "Computer-aided dispatch" (CAD) software was less a neutral computer-based form whose standard slots the call-taker had to fill in and more a guiding feature of the call-takers' dialogues with their callers; what he called a "preeminent contingency" on the dialogue and the way subsequent work progressed.

In particular, Whalen argued that the structure of the computerised form which the call-takers used, called a "face sheet", affected the procedure of the dialogue between the call-taker and caller.

Firstly, the face sheet required certain slots to be filled before the request for emergency services would be transmitted to dispatchers. This essentially specifies the minimum information which needs to be gathered during the call; and could influence the way the caller's requests are directed by the call-taker.

Secondly, the layout of slots within the face sheet could order the information in a series of sequential steps intended to fill one slot at a time and in an order lead by the relative location of slots. For instance, it was expected that the arrangement of

---

1 Heath and Luff (1992a) observed work patterns in a London Underground control room. Martin et al. (1997) observed the "working division of labour" between individual dispatchers, their tasks, their supervisors and their computer and hard copy support systems in an ambulance control centre; concluding that decisions involve interaction between all parties and that the use of particular features of their decision support systems depended on more than the systems' "cosmetic" advantages.
Whalen (1995b) observed fire engine dispatchers' use of old and new decision support systems, concluding that decision support systems should be "less prosthesis and more evaluation tool" (p. 179). The fundamental conjecture was that these expert systems should suggest arguments for making decisions rather than appearing to provide fait accompli recommendations.

2 This is the micro analysis of one instance of a behaviour, grounded in already acquired knowledge and systematic but less detailed, perhaps observational, analysis of similar instances.

fields from top to bottom of the screen would reflect the order in which information was processed during the calls.

Thirdly, the slots required information to be entered in specific formats; for instance addresses had to be absolute ("52 Hillhead Street") rather than relative to the caller ("Across the street"). This could have implications for the way the call-taker questioned the caller to convert their subjective, and occasionally hysterical, descriptions into the CAD's required form[3].

All three ways in which Whalen suggested the form of the call taker's interface might influence their dialogue with callers found considerable support in the call which Whalen submitted to single-case analysis.

This call was a request for police evidence from a member of the public who had confronted two men hiding something on his land, but had left them when he realised they had a concealable firearm.

The call was received as the call taker was discussing a less urgent issue with another caller, and this in itself lead to a major indication of how the presence of the call taker's form influences their dialogue with customers. Immediately, the call taker asked the almost terse question "Nine one one, what is your emergency." This not only served to remind callers that the number should only be used to deal with emergencies, but also indicated that the central issue in the dialogue would be the classification of the event which had lead to the call. The field in which the code for the type of incident was at the top of the screen, apparently the one the cursor appeared in when a new face sheet was generated, so it is also noticeable that the first item on the sheet was the one the call taker guided the caller towards answering.

This shows the call taker trying to control the dialogues, a feature which is even more evident when they interrupt early attempts by the caller to provide a narrative rather than relevant details, in order to bring it in line with their software's interface geography and to a certain extent how the call taker pushed the caller towards describing the event in a way which could be entered into their software.

---

3 In previous work, the researchers had shown how just such requirements had lead to a well reported argument between a call-taker and caller, subsequent delay in the dispatch of emergency vehicles and the death of a patient (Whalen, Zimmerman and Whalen; 1988).

Later in the call, the call taker had to fill in a row of short fields, and even though some were skipped over because the call taker could fill them with information they already had or knew they could fill them once the important details had been sent to dispatchers, the order in which pressing the <RETURN> key moved through these fields was also the order in which the call taker asked for them.

More subtly, the call taker would return to the first field throughout the call, as more accurate or pertinent representations of the event were gathered. The way in which the presence of the form altered the call taker's dialogue with the caller was particularly evident at these times; for instance the call taker interrupted one question to request clarification of a code they had mistyped and later used such clarification to cover extended backtracking through the form's fields.

This shows that the software is influencing the progress of the dialogues when otherwise the order in which information was revealed was unimportant and otherwise could have been arbitrary.

The constraint that the software required certain fields to be completed was apparently less rigourously applied, since the call taker would enter dummy information in fields to expedite the passing on of the sheet to the dispatchers. However, in the final analysis they too played an important role, both because they were returned to once the call had been actioned and, Whalen reported because dispatchers receiving face sheets with dummy entries would perceive them as more important than complete ones.

Altogether, the coordination of the call taker's use of their software and their dialogue with their caller is striking, as the call taker moves the cursor to fields entering probable answers even as they are asking questions, and reacts to what the caller is saying to bring the dialogue back to one which will facilitate the filling of the form

From this Whalen suggests that the call taker's face sheet, which is intended to standardise the work and should be indifferent to the details of each call, in its application in real calls becomes a leading feature in the progress of the calls even if the independent nature of the simultaneous dialogues between the caller and call taker ensures the consultations are more than the fulfilment of the software's mandates.

Smith and Whalen (1995) reinterpret Whalen's (1995a) observations, contending that the face sheet should not only be considered an artefact which influences the dialogue between the call taker and caller, but also a third party which filtered information before it was sent to the dispatcher.

1.1.2 Computers in Medical Consultations

One of the other areas in which the impact of computer on the conversation between two humans has been studied is in doctors' consultation, and one of the most important such studies is Fitter and Cruikshank (1982).

They studied the use of diagnostic expert systems emplaced in a hospital outpatient clinic and during general practice training, analysing video recordings of the interactions and post-consultation questionnaires.

Although they were unable to find strong evidence for the consultations being lead by the interface's structure, for instance doctors asking about topics in the order they were presented by the interface, asking questions or using the systems' terminology rather than their own, there was ample evidence that the consultations had become more "computer-oriented".

The doctors were seen, for instance, to enter responses which they could infer, but about which they had not asked questions, to improve the systems' abilities to make diagnoses.

More importantly, the way in which the system was included in the consultation varied as well, with broadly three patterns of interaction being present. The first involved the gathering of information without reference to the computer and then entering it all while the customer was occupied with other things; for instance getting undressed for a physical examination. The second they called a "conversational pattern" and involved entering each symptom details as soon as they were gathered. The third they called a "burst" pattern and involved collecting a few symptoms then leaving the patient unoccupied while they were entered.

During the preliminary observations made in this study, the patterns' use seemed to be largely based on doctors' personal choice, with pros and cons of all three being recognised. One interesting observation, in the light of later work, was that doctors using the conversational pattern employed various strategies to prevent their interruption, such as asking the patient to confirm entries, staring at their screen, staring into space as if in thought or pursuing alternative topics.

Christian Heath contributed to Fitter and Cruikshank's (1982) work by providing a preliminary micro-analysis of the interaction of the doctor-patient dialogue with their computer use.

Doctors' attention being focussed on their computer appeared to be indicated by gaze, but transition of attention might also be signalled verbally by doctors telling their patients what they were about to do. This was most common when the "burst" pattern was being used, and when a "conversational" pattern was used, they tended just to switch attention without explicit warning. The latter appeared to disturb patients' dialogue because there was no natural break.

A second aspect Heath noticed was the disruption introduced by the system's printer. Not only did the noise make audition difficult, but also the rhythm of speech seemed to match the printer's.

The questionnaires found that patients were generally more favourable to the use of computers in the consultation room once they had experienced them, and that, while there is generally a decrease in patient stress during consultations, those in which computers were used lead to a smaller decrease than those without. At the same time, a multiple regression test showed that the strongest predictor of attitude towards computers was post consultation stress, limiting the importance of both the other findings.

Fitter and Cruikshank's principle conclusions were that the doctor was an "information filter" who used their knowledge of each of their partner's interactive demands to translate between them, and that the overall effect of the computer on the doctor's dialogue with the patient was to formalise it, both by providing particular requirements for information and because their response times interrupted fluent conversation.

During his analysis of the emergency dispatch facility conversations, Whalen briefly commented on how callers, whom he presumes can hear the call taker's keystrokes, coordinate their utterances with the call taker's typing.

For instance, in one call Whalen analyses, a period of typing following the donation of information and a soft "okay" used as a minimal acknowledgement is treated as evidence that the call taker is involved with their computer, and so the caller waits until directly asked a question before continuing. Elsewhere, however, he reports occasions of callers talking over call takers' typing seeing it as an opportunity to elaborate on information they have just given.

The latter behaviour, overlapping elaborations, are considered by Whalen to be intended to be helpful attempts to help the call taker's recording of the information they have just been given, for all that the overlapping speech and typing may make it difficult for the call taker to follow the elaborations so that sometimes they ask callers to remain silent while they enter information into their software. In the light of his own examples, it could also be argued that the callers are trying to fill in apparent periods of inactivity in the call taker's gathering of information with their own narrative of the events – which would more easily be construed as disruptive.

The former behaviour, awaiting explicit requests to continue, has influenced a more adventurous interpretation of the computer's role in tripartite communication in doctor's consultations.

Greatbatch, Luff, Heath and Campion (1993) studied a series of video-recorded doctor-patient consultations using the methods of conversation analysis[4] to determine the extent to which the technology shaped and mediated the communication between the doctor and patient. Although little control was made of the physical relationship between the computers in the doctors' offices and the customer, they found striking patterns in the communication between the three which lends support to the proposal that the computer is being treated as a partner.

The most basic discovery was that the pattern of interaction between the doctor and patient changed when the computers were introduced. Fitter and Cruikshank's (1982) "conversational" pattern of taking one piece of information and recording it was more common when doctors were taking hand-written notes, whereas their "burst" pattern of taking several pieces and recording them as a group was more common when they used the computer. This would also appear to have been confirmed by Urkin et al. (1993, reported in Scott and Purves, 1996).

The most striking relationship they observed, however, was that initiation of patients' unsolicited turns in the dialogue recurrently followed doctors' keystrokes which completed tasks in their interaction with the computer. The customers

---

4 Conversation analysis is defined by Luff (1990) as a field which "pays attention to the social context in which interactions take place, and it aims by rigourous and detailed examination of naturally-occurring data aims to show that there is an organised structure to those interactions" (Luff, 1990, p.1). The work of Sacks, Schegloff and Jefferson (1974) was seminal in its development (see Section 1.3.1.1.1)

appear to monitor the doctors' use of their keyboard to predict when the doctor had completed an interaction with the the system, and initiate their turns accordingly.

For example, Greatbatch et al. (1993) observed that the patients looked for particular patterns in the use of the right hand to indicate when they could take their turns. Movement of the hand across the keyboard towards the carriage return key, lifting the hand higher than usual, a more pronounced thrust at the key and consequently a louder report from the keyboard were all taken to be evidence that the entry was complete, and the doctor was available to be talked to.

Essentially, they were looking for patterns of intonation in the way the doctor typed which would indicate that a turn transition point (see Section 1.3.1.1.1) was about to be reached.

This can be contrasted with consultations in which the doctor takes hand-written notes, because in these the patients do not co-ordinate their speech with the doctor's writing, and the doctors interleave their taking of notes into the conversation with the patient rather than obviously concentrating their attention on the computer.

The conclusion that Greatbatch et al. (1993) drew about the relationship between doctor, patient and computer was that, when computers are present, the consultations are structured around their use and operation. In a more wide ranging review Scott and Purves (1996) reinterpreted the relationships to include the computer as a partner in the consultation.

They recognised that there were obvious benefits of computers for patient record keeping and decision support, but cautioned that any detrimental effect on the doctor-patient relationship had to be weighed against these — particularly in general practice where relationships can be long and delicate.

The majority of the work they reviewed was about patients' attitudes towards the computers used in consultations and towards the doctors who were using them.

Although early studies from the 1970s and 1980s that they reviewed showed some patients to be opposed to the presence of computers in general practice, by the 1990s considerably more patients appeared to recognise the benefits of computers to themselves; opposition had also become untenable since near 100% of practices are computerised. Patients seemed to think that the computers might improve diagnoses, and there was only a limited fear about loss of confidentiality, but there appeared to be more worry about the doctor's communication with patients being damaged. For instance, in Cruikshank's (1984) study, 52% of patients agreed that a general practitioner's personal touch would be lost.

Continuing their reviews with Fitter and Cruikshank (1982) and Greatbatch et al. (1993), and recognising such elegant findings as Bright's (1991) that general practice patients tended to be more comfortable with computers when they were closer rather than farther away, Scott and Purves (1996) built a triadic model of the relationship between doctor, computer and patient which they called the DCP model.

In this model, each participant not only interacts with each other participant dyadically, but also has an effect on the relationship between the other two. They argue that consideration of each dyadic relationship within the triad is no longer valid because it is not possible to distinguish effects which should actually be ascribed to the third participant.

They note that shifts in perception of the triad will reflect different viewpoints; for instance the doctor could be considered an information filter between the computer and patient, the computer an intrusion on the doctor-patient dialogue, the computer a co-participant which leads both through the consultation or the patient could be seen as central to the consultation and the doctor and computer combined as their service provider.

1.1.3 Computers in Customer Service

Although there has been an increasing interest in the use of customer facing computers, for instance the design of web sites intended to provide a sales and information source for customers even for synchronous communication (e.g. Kobayashi, Shinozaki, Sakairi, Touma, Daijavad and Wolf, 1998; Zhang, Wolf, Daijavad and Touma, 1998[5]), there has been very little study of the asymmetric use of computers by human customer service agents.

One place it has been investigated is US West Technologies. For instance, in Project Ernestine (Gray, John and Atwood, 1992, 1993; Gray, John Stuart, Lawrence and Atwood, 1990) the use of a new ergonomically designed workstation for Toll and Assistance operators, who assist the customer in completing calls and recording the correct billing, was compared with that of the line and character-based system it replaced.

In Toll and Assistance (TA) calls time really is money, and Gray and his collaborators estimated that at the time an average saving of 1 second per call would save the company 3 million US dollars a year. With this in mind, a new workstation was built using ergonomic principles both in the creation of a graphic user interface to replace the old line and character oriented one, eliminating keystrokes, and also in a keyboard specifically designed for the TA operator's tasks. It was predicted that this design would cut up to 2.5 seconds from the average call.

In an extensive field trial which lasted around 4 months, sampled 78240 calls, and involved 48 experienced TA matched operators of which 24 used the new workstation and 24 used the old, their expectations were completely contradicted.

Although the users of the new workstation mastered its use within the first week, their calls consistently and statistically significantly lasted an average 4% (roughly 1 second) longer than those using the older systems, and this did not even seem to hide a benefit to particular call types.

Anecdotal evidence suggested that its users liked the new workstation and wanted to "beat" their times with the older system, but were being slowed by the system.

---

5 Kobayashi et al. (1998) investigated two applications for synchronous sharing of banking information over the web; one designed for use in kiosks and one for use with home customers. They discussed the requirements, both similar and independent for each application.
Zhang et al. (1998) investigated three options for voice communication during synchronous auto insurance purchase over the web; internet telephony, a proprietary protocol for simultaneous voice and data transmission and a second telephone line. The delays from internet telephony were found to be particularly damaging. Although the second line lead to perceptibly better voice transmission than the proprietary protocol, it did not seriously affect the task performance.

To find an explanation for this failure calls were analysed using a variation on GOMS analysis (Card, Moran and Newell, 1983).

Whereas GOMS represents every perceptual, cognitive and motor operation as well as the dependencies between them, the variety used by Gray et al. (1990, 1992, 1993) introduced the concept of the critical path. This represents "the sequence of of tasks that determines the soonest the project can finish" (p. 6 CLARIS Corp, 1987; quoted in Gray et al., 1990, p. 636), and in GOMS terms means that only the longest of activities which are being performed in parallel contributes to the total time for the task. For instance the movement of a hand to expected keys during the elicitation of a response would be ignored.

The GOMS analysis showed that considerable changes had been made in the tasks the agents were performing with the introduction of the new workstation; and this had lead to the increased call times.

For instance, although the new workstation had removed two keystrokes during the mandatory opening of the call, requiring 10 GOMS operators, none of these were on the critical path because the work time was defined by the speed of the conversation and the required keystroke was reintroduced later in the call – but this time on the critical path. The GOMS analysis would have predicted that the new workstation would lead to calls which were 3% slower than the old systems.

It was also found that aspects of the ergonomic design were flawed. For instance, the new workstations grouped certain function keys which were necessarily on the critical path close together on the basis of their relationships, but this encouraged the use of only one hand and consequently slower input times.

These findings are all relevant to customer service, but in concentrating so heavily on a human computer interaction style analysis Gray, Johns, Atwood and their collaborators miss the relevance of the unique nature of the tripartite customer service relationship. This is a missed opportunity, because for all that the GOMS analysis apparently concentrates on the dyadic interaction between agent and computer, it inherently models the communication between agent and customer within the agent's role.

Although the reason for this is probably the researchers' background in HCI and a recognised goal of validating GOMS in a large real world situation, it is also worth noting that the subject of the research could have limited them.

Because the technical changes in the interface recognised by the GOMS analysis could be estimated to account for the majority of the performance difference, this suggests that the role of the computer systems in these TA calls is much more pervasive than in the emergency service or general practice calls reviewed above. From the limited descriptions and abbreviated PERT charts provided in Gray et al.'s (1990, 1992, 1993) papers it appears that their TAO calls are largely scripted, which means each step through the calls is controlled by the computer, often to the extent that operators are advised or commanded what to say. This may explain why the GOMS predicted effect of software matches the observed effect so well, but suggests that the TA calls they study are an unusual form of tripartite communication because the agent is acting as little more than an intelligent natural language interface in a dyadic customer-computer interaction.

On top of that, the finding that it is a change in task which contributes to the majority of the effect of the new workstation means that it is not so much a comparison of technology as of tasks, and so the role of all three participants may have been changed. The finding is more that care must be taken when customer service situations are to have an order imposed on them, than that the design of the agent's support system can impact on their dealings with the third party client.

Lawrence and Atwood moved on to consider the role of the agent within the customer service calls more closely. Lawrence, Atwood and Dews (1994) analysed real dialogues between customer and Directory Assistance (DA) operators, who assist callers to find unknown telephone numbers, as well as the interaction between the dialogues and the operators' keystrokes.

Their conclusion was that the operator was a "surrogate user" who mediated between the customer and their systems. They were involved in everyday dialogues with their customer and technical dialogues with their systems, with practice performing the two in parallel, but essentially only translating between the two and making simple inferences.

Muller et al. (1995) picked up where they left off, with the intention of showing that the DA operators are in fact expert workers who help customers navigate unknown knowledge domains – such as government hierarchies – and deal with changes in the customers' worlds. For them, this distinction was important because it could affect relationships between workers and management within the company.

They managed this through qualitative modelling of the tasks DAs perform, which was verified by DA trainers and quantitative analysis of videotaped DA calls, with participatory analysis provided by the DA operators.

The modelling of the DA operators' tasks suggested that they were "knowledge workers", who used their own expertise during the formulation of database requests in 53% of their calls. They used this, for instance, to enter the correct form of request to maximise their software's chance of finding a number, to select the government department which matched callers' requirements, and even to volunteer work-around solutions to contacting businesses or agencies they knew to be difficult to reach.

As a supplement to their work on DA calls they performed a case study of a single TA operator, to test Gray et al.'s (1990, 1992, 1993) comments implying that TA operators did not perform knowledge work. Real time scoring found that 38% of calls required considerable knowledge work, but note that much of this occurred in infrequent call types which Gray et al. may have excluded[6].

Overall, they conclude that, whilst operators clearly are mediators between customers and computers, their main role is to make sense of customers' poorly formulated requests. In doing so they showed that there was frequently more to the nature of tripartite customer service communication than two, or even one if Lawrence et al.'s (1994) interpretations were followed, dyadic interactions. The operator has to build a query for their software which accurately reflects the customer's requirements, and this requires that they consider both customer's and computer's demands as well as drawing on their own knowledge.

---

6 Gray et al. (1990, 1992, 1993) intended to use 20 call categories which would have accounted for 88.3% of all completed calls, but in the final analysis rejected 5 of these because of infrequency. Muller et al. (1995) comment that 46% of their calls were "call completion" but the knowledge work in these accounted for only 6% of the total.
From the limited description and abbreviated PERT charts provided in Gray et al.'s papers it appears that their TA calls were very strongly lead by their interfaces' structure, suggesting that in these calls the TA operator really was acting as little more than an intelligent natural language interface, and explaining the close relationship between Gray et al.'s GOMS based predictions and the observed effects.

1.1.4 Summary

The conclusion to draw from the reviewed work on tripartite communication between two humans and a computer is that it it should be considered triadic and not as a set of dyads.

The work of Gray et al. (1990, 1992, 1993), which applied formal HCI methods in evaluating the interaction between the agent and computer in a tripartite customer service call, found little more than that, when developing new systems, the tasks they support should not also be changed.

All the other researchers, using less formal methods, discovered at the very least that the role of the agent was more complex than being a simple "surrogate user" and more commonly that the interaction between all three participants was linked.

Muller et al. (1995) showed that the agents in DA calls were experts who provided advice outside their immediate task domain as well as using their experience to help formulate the requests they really intended.

Whalen (1995a) showed that the design of an interface could influence the order in which call-takers at an emergency dispatch facility gathered information from their callers.

Greatbatch et al. (1993) showed that a general practice doctor's interaction with their computer intricately affected both the way they dealt with their patients and the way the customers structured their dialogues with the doctor; essentially being treated as a third party in the communication.

Scott and Purves (1996) built a model of the doctor-computer-patient communication which showed how inter-related each party and their dyadic interactions were; explicitly recognising the whole as triadic.

Methodologically, it is worth noting two things in comparison of the studies which have made claims for the tripartite communications being triadic, and those which have not. Those which did not make the claims used formal evaluative techniques and considered only the interaction between the agent and their computer. Those which did make the claims depended on informal sociological techniques and concentrated on the dialogues between the two human participants.

The study of dialogue has lacked formality and the formal study has lacked dialogue.

The remainder of this chapter will complete a review of formal and experimental methods in HCI, showing how language has been included within them (Section 1.2), and a review of what is known about human-human dialogue, showing how it is known to be affected by technology (Section 1.3).

The primary aim will be to extrapolate how language has been used in previous formal studies, with the intention that this may inform its use in the formal study of tripartite communication.

1.2 Human-computer interaction

As well as showing what is currently known about tripartite communication with computers, the work described in Section 1.1 also provides a fairly comprehensive overview of the methods currently employed in HCI, and the varied influences which have contributed to it[7].

Sociology has influenced participant observation studies such as Whalen's (1995a), where researchers perform field studies observing the actual use of computers in day-to-day work, typically assessing changes in the way companies function when they are computerised, and single-case analyses to demonstrate the behaviours they consider important. It was also at the heart of the work by Greatbatch et al. (1993) who used the methods of conversation analysis, and Muller et al. (1996) who applied participatory techniques in novel fashions.

Psychology, Software Engineering and Formal Linguistics has influenced more formal work such as Gray et al.'s (1990, 1992, 1993) in which formal models are built of the interaction between human and computer, and predictions from them evaluated through experiment.

On the other hand, even though some of the techniques applied were intended to be used during the design of software, all the work in Section 1.1 was linked by being applied after software had been designed and built, and was therefore evaluative.

7 These methods and influences are attested to in such histories of HCI as Gaines and Shaw (1986), Baecker, Grudin, Buxton and Greenberg (1995), Shackel (1997) and Carroll (1997)

Early HCI drew from the human factors traditions, and therefore concentrated on evaluation, but since the publications of Card, Moran and Newell (1983) and Winograd and Flores (1986) it has been recognised that "*Design is where the action is, not evaluation.*" (Newell and Card, 1985, p. 214).

Section 1.2 will review some of the methods which have developed in the design side of HCI. Section 1.2.1 will cover the work which grew directly out of Card, Moran and Newell's (1983) work, which has drawn primarily from psychology and artificial intelligence. Section 1.2.2 will cover the work which has been more strongly influenced by Winograd and Flores (1986), which draws from ethnography and its related subjects. Section 1.2.3 reviews studies which have adapted what is known about human-human dialogue to HCI research.

1.2.1 Formal and Experimental Methods in HCI

The purpose of this section is to review what has been refered to within the thesis as formal HCI; it is the application of the methods and findings of psychology and cognitive science to the design of user interfaces. This work is defined by the testing of hypothetical changes in interfaces through rigorous laboratory experiments and the formal modelling of ensuing human-computer interactions.

Rather than take snapshots from throughout the literature, this will be accomplished by showing how HCI has investigated one relevant issue in interface design; consistency.

Consistency in interfaces is widely regarded as one of the most important features in the design of human-computer interfaces. The first of Shneiderman's (1998) Eight Golden Rules of Dialogue Design is to "Strive for Consistency"[8], Reisner (1981, p.234) was able to write "It is almost a truism that consistency of an interface will make it easier to use." and the instruction to design for consistency is ubiquitous in interface design guide-lines (e.g. Smith and Mosier, 1986).

[8] The prescriptivism implicit in the building of guide-lines was a common feature of early HCI, which depended on interface "gurus" sharing their practically gained experience (see Baecker et al. (1995), Shackel (1997) and Carroll (1997) for reviews)

At the same time, Greatbatch et al. (1993) noted that HCI's definitions of consistency and its inferences from empirical work may need to be reformulated to handle the design of interfaces in tripartite communications with computers. It therefore provides an obvious tributary through which this thesis could feed into mainstream HCI.

During the description of the work on consistency, the concentration will be on showing what HCI has discovered about consistency, but once this is complete, the work will be returned to with a more selective eye and the intention of deducing the common techniques of formal and experimental HCI.

1.2.1.1 Formal Modelling of HCI Consistency

1.2.1.1.1 Backus-Naur Form

One of the earliest studies of consistency was made by Reisner (1981) who tried to formally model consistency within interfaces by applying the Backus-Naur Form (BNF)[9], which she inherited from the formal specification of computer languages but which is strongly related to the tradition of production rule grammars associated with Chomsky (1964).

Her contention was that the descriptions of individual interfaces' "action languages" would permit the assessment of those interfaces on the three criteria of **density**, **string simplicity** and **structural consistency**. Density referred to how many actions were possible in an interface, and was represented by the total number of terminal symbols in the language. String simplicity referred to how complex strings of actions needed to be to achieve goals, and was represented by the length of terminal strings for particular tasks. Structural consistency referred to the similarity between different tasks in the way they were completed, and was represented by the number of rules necessary to describe a set of terminal strings.

This was the first recognition that there could be a consistency beyond the same terms meaning the same thing and the same single actions achieving the same results in different contexts; which might be measured by density and string simplicity. Instead, her structural consistency measured the degree to which similar sequences of actions occurred in systems.

---

9 She later named the notation and methodology she had used Grammatical Representation of Action Language (GRAL).

She then made and tested predictions about how the string simplicity and consistency would affect the user's learning and expectations of how a system would respond.

Ten office workers learned two drawing systems over a two hour period and were then observed during a memory test in which they had to perform a set of tasks without using the manuals and filling out questionnaires comparing the systems.

The interface which was more structurally consistent according to her BNF models was found to take 1/3 less time to learn, and informal testing showed that this could have been for several reasons.

Firstly, when the system which had inconsistent rules for the selection of certain tools, and a consistent rule for other actions had been learned, users would make "expectation responses" in which they searched for a consistent means to perform an action.

Secondly, when there were separate rules for performing similar actions, for instance the drawing of continuous and discrete shapes in one system which needed only the learning of a single rule in the other, over 70% of her users erroneously misapplied the incorrect rule in the system with two during the memory test.

Finally, although one of the systems had a rule for joining lines which was intended to be consistent at certain points of the task, this was inconsistent with rules applied at other times in the task and lead to destructive actions when the more general rules were over applied.

In both the last two cases, the structurally inconsistent interface had been designed to minimise physical action, but a conflict between ease of use and learning had been found.

Her main aim had been to show that her BNF forms could predict the usability of systems' designs before they were built, consequently adding rigour to a field which was considered soft and therefore marginalised by its parent software engineering, but she had also seeded an interest in how formal models of interfaces could be used to assess interface consistency.

1.2.1.1.2 Task Action Grammar

This seed was cultivated by a number of researchers, for instance Polson (1988) who applied a variant of the GOMS model of task execution, but the most influential model has been Payne and Green's (1986, 1989) Task-action Grammar (TAG).

This grammar, they argued, improved on the earlier models by paying more attention to the user's model of their tasks, by recognising that the task world could be categorised, thus adding more recognition of the user's mental representation of their tasks and increasing the psychological validity of the models.

As well as this, they also added several other types of consistency to Reisner's structural consistency, which they termed syntactic, so they were also capable of capturing relationships between families of rules which other schemes could not.

Lexical consistency which is a metaphorical concept, not present in grammars of natural languages which occurs in computer interaction because the command words and individual actions through which human-computer interaction takes place benefit from being *congruent* with the actions which will be performed by the system (Carroll, 1982; see below).

Semantic consistency was a form of completeness whereby, if actions are permissible in one task domain, they should be permissible in others.

Semantic-syntactic alignment referred to consistency between actions performed and results; such that a consistent interface would use the same element to perform the same function in all cases. In one counter-example they provide, <CONTROL> and <ESCAPE> keys are used to toggle between small and large units in some cases, but direction of cursor keys in others.

With this model they were able to predict earlier findings about consistency. Lexical consistency meant they could predict Carroll's (1982) finding that a command language with names which were congruent, i.e. commands which performed opposite actions used opposite words so that RETREAT was in opposition to ADVANCE, was easier to learn than one which was incongruent, i.e. commands bore no relationship to each other so that BACK was in opposition to GO. Syntactic consistency and semantic-syntactic alignment meant they could not only support Reisner's (1981) findings but also capture relationships between the semantic and syntactic aspects of her systems.

To show that the inclusion of categories within task domains improved the model, they set up their own experiment in which subjects had to learn to perform three types of activity, each of which contained a number of distinct tasks, within a "lost property office" system.

Three task-action grammars were employed. The first, which TAG considered the most consistent, maintained a single syntax between all three categories of activity. The second, which TAG considered the next best, used a different syntax for each activity. The third, which TAG considered the least consistent, used only two rules but they were both used within each category of activity. A number of other task-action mapping systems, including Reisner's, and less formal arguments about consistency which they review, would have predicted other orderings.

The subjects were trained in the use of the lost property office system then asked to perform a number of tasks through written "questions" on their screens. Their performance was measured through measuring the time it took them to complete their tasks and how many errors they committed. A memory task was used to assess the mental workload each grammar imposed. Finally, they were given questionnaires intended to capture their subjective feelings about the grammars they used.

The predictions based on the TAG models of these grammars were supported. The least consistent grammar lead to more syntax errors than the medium and that to more errors than the easiest. Similarly, the least consistent lead to longer mean performance times than the medium, and again that to longer times than the easiest. The questionnaires generally suggested that subjects were aware that the hardest language was less well designed than the other two, but there was little awareness of the medium being less consistent than the easiest.

The experiment had shown that the TAG scheme, by improved modelling of user's grammar of the tasks they were performing, could more closely model the consistency of the systems with user's expectations.

1.2.1.2 Discussion of Consistency

The growth in these two models, from Reisner's BNF– which hoped to have some psychological validity– to Payne and Green's TAG– which explicitly represented the user's mental representation of their tasks– reflected a growing recognition in more discursive HCI work that consistency was not just a concept which was internal to systems, but should also consider the users.

At a workshop on Coordinating User Interfaces for Consistency at CHI '88 (Nielsen, 1989), this dichotomy was clear. No unique definition of consistency could be arrived at, so instead the "gurus" set out to define the dimensions of consistency; the range over which consistency could apply, how consistency could be maintained and its benefits and dangers.

The majority of the participants at the workshop (12/15) were from companies involved in software and system design and, perhaps as a consequence, the majority of the recognised benefits (14/18) were to the system designers and their companies; for instance reducing development costs and giving better product definition in the marketplace, rather than to users.

Although it was briefly noted that systems should be consistent with users' expectations, it was also implied that these expectations could be created by consistent design within software, rather than by depending on users' non-computer experience.

Perhaps the most important result to be generated by the workshop was the impetus in Grudin to write a discussion paper on the potential failings of the consistency (Grudin, 1989).

He began by defining three classifications of consistency; internal consistency, consistency with other software known by the user and consistency with users' non-computer experience.

The gist of his arguments against too strong an application of all three forms of consistency was that, for all that consistency with users' experience could assist their learning of interfaces, a dedicated design could improve their long term performance[10]. He backs this up with several examples of studies of ergonomics and character-based computer systems which showed heuristically designed

---

10 It is symptomatic of the complexity of HCI research that whilst it is recognised that experience can greatly alter the nature of interaction between a user and their software, interfaces are typically evaluated by their ease of use for novices and experimental work almost universally involves the learning of new interfaces.
In defence of studies of novices, it can be argued that experience may mask the work required to understand interfaces so novice studies can indicate what is involved in mastering interfaces (Suchman, 1987).

systems rapidly becoming inefficient because of actual patterns of use or physiological constraints.

In reponse to Grudin, Reisner (1993) returned to the study of consistency to show that some of his examples of beneficial inconsistency, particularly an analogy based on the placement of knives within a house, were in fact consistent because they recognised the user's expectations and usage even though they were inconsistent with the hypothesized knife placement rule. Grudin had in effect being arguing against too strong an application of designer's prescriptions which neglected the system's user.

Reisner argued that what was really needed was a method for finding mismatches between the designer's point of view and the user's[11]. With this in mind, she extended her early BNF forms to include both user and designer judgements of the system's constituents, with the intention that a system would only be judged consistent when these were the same.

1.2.1.3 Consistency in the Thesis

The reason consistency was used as an example of a topic in HCI is that since it is central to good interface design, as guide-lines universally recommend (Shneiderman, 1998; Smith and Mosier, 1986), and it depends on a matching between the structure of the software's task language, as imposed by its designers and engineers, and the structure the user expects of their interaction (Payne and Green, 1986, 1989; Reisner, 1993), then in tripartite communication it will also have to account for the influence of the third party on the user's expectations. This essentially means that it will have to both account for consistency with an experienced user's demands for parsimonious and rapid interaction as well as a novice third party's demands to give and acquire information in the order they expect.

For example, in many current customer service calls addresses are used to confirm customers' identity, and the most parsimonious way to identify a customer's address is by the postcode. Therefore the software may explicitly, by dialogue boxes, or implicitly, by the order of form fields, request the customer's postcode before their county, town, street, or house name or number; and the

---

[11] This has also been recognised in observational studies of software engineers by Kellogg (1988).

agent will appreciate the parsimony of this because they only have to type in 6 or 7 characters if the address is already in their database. The customer, on the other hand, may be pulled up short by an initial request for a postcode because it is more common to identify addresses by street, suburb, town or county when conversing with other humans.

This represents one avenue whereby this thesis may feed into HCI, but a more important reason for looking at it was to complete the overview of HCI begun in Section 1.1; capturing the methods employed by formal HCI.

Both Reisner (1981) and Payne and Green (1989) provide good examples of these.

Their goal of validating models is not ubiquitous, but marks a conscious effort to turn the softly regarded HCI into a hard science, as does the use of hypothesis testing experiments, when early HCI depended to a large degree on self reports. Although they may have suggested that these models could be used in the development and evaluation of interfaces, within their own work they were used to find the basic building blocks of interface design and that represents one aspect of formal HCI which is only subsidiary in evaluative HCI.

The measures they use are however common; employing the three P's of perception, performance and process. Subjects' perception of the ease of use of software is gathered through questionnaires, often supplemented by open ended questions. The effects of different designs on performance is measured through analysis of subjects accuracy and the time they take to complete tasks. The way in which different designs interact and conflict with subjects' expectations of the structure of tasks is assessed by collecting details of the subjects' process through tasks, typically by classifying the errors they make, often looking for over-generalisation, or by measuring cognitive load with means such as memory tests.

One duality within HCI is evident in the design of the experiments Reisner (1981) and Payne and Green (1989) used. Although they are trying to discover basic interactions between computer interfaces and human cognitive resources, they designed experiments which resembled "real world" tasks. Payne and Green (1989) in particular went to some pains to ensure their tasks reflected a realistic scenario and included secondary tasks to make cognitive load similar to a potential user's. Even when performing controlled experiments, researchers must be aware that HCI is intended to be applied in the design of real software and wary of criticism their findings will falter when applied[12].

That criticism can always be applied to the majority of HCI's controlled experiments, however, because they almost universally involve one user interacting with one computer; a situation which is rare in the real world even in the use of standard desktop applications such as word processors because employees are always surrounded by co-workers or at the beck and call of their telephone.

Given the current thesis' interest in tripartite communication it is particularly noticeable in Payne and Green's (1989) study that, for all their attempts to make the task realistic, they ignore the fact that the lost property office task they designed would typically involve interaction with a third party.

Rather than having their subjects interact with "customers" to determine what should be asked of their databases, the requests were presented through written "questions" on their screens. This does not affect the reliability of their findings about consistency, but neither would have the use of a less realistic task. Instead, the implication is that researchers should consider whether it is useful to make tasks realistic and, reflecting on the work on tripartite communication with computers described in Section 1.1, should be aware of the total context of the tasks they model.

1.2.2 Contextual Methods in HCI

Contextual methods in HCI are those which emphasize the importance of the context in which interfaces and systems will be used. These methods stress that human-computer interactions should be studied in the field, because it is only there – in the context of the activities which surround it – that the way software is used can be understood.

The generally quoted impetus to the use of contextual methods was Winograd and Flores (1986), but equally important in its promotion in HCI seems to have been John Whiteside at Digital Equipment Corporation (e.g. Whiteside, Bennett

---

12 This was noted by Baecker et al. (1995, p. 571) who also commented that research could never cover the wide range of approaches to interface design which resourceful designers might propose.

and Holtzblatt, 1988). Two of the methods, whose history in HCI can be traced back directly to these works are Activity Theory (Section 1.2.2.1) and Contextual Inquiry (Section 1.2.2.2).

1.2.2.1 Activity Theory

Activity Theory was first introduced to Human Computer Interaction by Bødker (1991), but has been evangelised more recently by Nardi (1996a). It originated in the philosophical framework formulated by the psychologist Vygotsky (1978) for the cultural-historical study of language, but credit for much of its current form can be given to his colleague Leont'ev (1974, 1978).

It is recognised that the theory can be difficult to understand (Nardi, 1996b), and there are several inter-related reasons for this.

One simple reason is its vocabulary, which has often been inelegantly or misleadingly translated from the Russian. Phrases employed in Activity Theory may have very different meanings outside it, for instance the term "object-oriented", and confusion is inevitable.

The second reason for the difficulty is that the theory provides points of interest to many different fields[13]. Each field has approached the theory with its own preoccupations, and has interpreted the important aspects of the theory to be different. This means that reading multiple researchers' appreciations of Activity Theory can be more confusing than informing.

A third reason, which is related to both the proceeding reasons, is that the theory is still evolving, and as individual researchers have traced the implications of particular aspects of the theory, it has been found that the theory is under-specified (Nardi, 1996b). Learners of the theory may encounter such occasions and wonder how they may be overcome, or discover other researchers' bug fixes which may seem to bear apparently loose relationships with the core theory[14].

---

[13] The chapters in Nardi (1996a) and skeletons in Ryder (1999) demonstrate the wide range of sciences which have shown interest in Activity Theory, including anthropologists, educators, work scientists, psychologists, linguists and philosophers.

[14] When an undergraduate, the author had trouble unifying Vygotsky's (1978) own Zone of Proximal Development with the rest of the theory.

A final reason is that Activity Theory provides principles which should influence research, but does not provide direct guidance in operationalising those principles. Researchers are instead left to develop methods to capture the data which can then be related back to the theory (Kaptelinin, 1996a, 1996b). Consequently, it can be difficult to perceive how the theory can be applied in practice.

Nevertheless, there are many researchers who believe that if human-computer interaction is to incorporate the context of computer use, Activity Theory provides a suitable framework for achieving that incorporation[15].

To avoid as much confusion as possible, this review will be limited to the core concept of the **activity**, but will perforce consider the role of artifacts in activity, because this appears to offer the most practical benefit to Human-Computer Interaction[16]. It is based primarily on Kuuti (1996), but has been influenced by Bødker (1991) and reading of several other chapters in Nardi (1996a); other potentially important concepts are the notion of consciousness and the asymmetry between humans and the things in their environment (Nardi, 1996b).

1.2.2.1.1 "You are what you do"

The central tenet of Activity Theory is that "*you are what you do*" (Nardi, 1996a); that is to say that it is the things we do, and the way the world influences the way they are accomplished, which defines our nature. Even the mental planning of physical actions – called **orientation**, and the closest Activity Theory comes to the concept of cognition – takes place in the context of mental **reflections** of the real world.

---

[15] Kaptelinin (1996a, 1996b) goes so far as to suggest that human-computer interaction would be better reconstituted as "computer-mediated activity"; indicating to his readers that computers are not partners in a communication but the tools which mediate our manipulation of objects (see Section 1.2.2.1.1.1), and also that the use of computers cannot be considered without an awareness of the context provided by the activity (see Section 1.2.2.1.1.2) in which they are used.

[16] Bødker (1991) and several authors in Nardi (1996a) suggest that changes will need to be made to Activity Theory for it to serve the purposes of HCI – for instance Kuuti (1996) suggests the replacement of the orientation phase.

1.2.2.1.1.1 Action

The unit of analysis which Activity Theory uses is called the **activity**[17], but central to any activity are the **actions** which take place within it, and it is with these that the review will start.

Any action involves two primary elements; the **intention** which the human **subject** brings, and the **object** on which they are acting. The purpose of the action is to transform the object in some way, and the result is known as the **outcome**. The intentions are the subject's idealised version of what will happen to the object. The object is the item – physical (e.g. a piece of paper), a mental representation of the physical (e.g. a map) or abstract (e.g. the style of a map) – which will be transformed during the action. In general, the object defines the action (Bødker, 1991; Kuuti, 1996; Nardi, 199b).

Within the original conception of the theory, a typical animal's actions include only intentions and objects. Human actions, on the other hand, are mediated by **tools**, which may also be called **artifacts**. These tools restrict the way in which the object can be manipulated, but these restrictions are not purely a function of the tool. Because their development is part of the historic development of the action itself, they provide a way in which humans can control their own behaviour. Through the tools they develop – not internally, but externally – humans determine their own future actions; they represent the collective experience of the acting human's culture. Furthermore, within Activity Theory, tools are inseperable parts of human behaviour.

With these three terms[18] defined, it is possible to explain one further strength of Activity Theory, which may also be another source of confusion. What is the object in one action may be the tool in another and vice versa. For instance, the

---

[17] The three level model of human conduct presented here is not an exact model, but an explicatory approximation. It is "an orienting device to handle the necessary focus changes in the course of ongoing research" (Raeithel and Velichovsky, 1996).

[18] The Scandinavian Tradition, or at least Engestrøm (e.g. 1987, cited by Kuuti, 1996), incorporated three more concepts within actions; these are the **community** which is given similar status to the human's intentions and the object, and social **rules** and **division of labour** which mediate between the community and the human and the community and the object respectively.
These appear to have been added as a direct consequence of the Scandinavian Tradition's application of Activity Theory in work science. Although intuitively they have value in contextual HCI research, they will not be covered in detail because they are not part of the original core theory.

map which is an object being drawn with a pen and paper in one action, may be the tool through which object directions are given in another; and similarly the pen and paper which are a mediating tool in the drawing of the map in the first action, may be the object of a third, repairing, action if they cease working[19]. In some cases it may be difficult to distinguish the constituent parts of the action; for instance, in the action of drawing a map, the ink and paper tools become part of the physical representation of the map object.

1.2.2.1.1.2 Activity

Although actions are, arguably, the most easily observable level of conscious behaviour recognised by Activity Theory, the theory also postulates that these actions take place in the context of larger tasks, called **activities**, and that it is only within the context provided by these activities, that the actions can be fully understood.

To demonstrate this, Kuuti (1996) refers to Leont'ev's (1978) example of hunting in which bush beaters frighten the game towards the catchers. If it was not known that the bush beaters were involved in a collective hunting activity, their actions would not interpretable – they would never catch their prey just by frightening it away.

Leont'ev's example also raises the issue of **individual** activities and **collective** activities. The hunting group are engaged in a collective activity, which involves **communication** between multiple humans, each with their own individual activities, which imply certain actions. The example demonstrates that studying activities at the individual level may not be enough to permit understanding, because some of the involved humans' intentions – and possibly tools and objects – must be inherited from the collective level of activity[20].

---

[19] When mediating tools become the objects of actions, this may be termed **breakdown** (Bødker, 1991); a concept which will be returned to in Section 4.1.2.2.

[20] This implies a recursion problem, similar to that with the grounding criterion (see Section 1.3.1.2.2), because there may be multiple levels of collective activity, and each one may depend for its understanding on those higher up the chain. For instance, the collective hunting activity may inherit from a higher-level collective subsistence activity, which also includes gathering activity; and that subsistence activity may inherit from higher still activities.
Anyone using Activity Theory will have to question what level in the chain of collective activities their analysis needs to extend.

1.2.2.1.1.3 Operations

Below the action level is the operation level, in which exist automatic, sensori-motor, reactions. These are the building blocks of actions, and are performed without conscious awareness, being triggered by material conditions during the process of an action. As with other concepts within the theory, in some circumstances operations can become actions and actions may become operations. If, when drawing a line on a map, the artist's pen fails to leave a mark, that otherwise automatic operation may be repeated with more conscious awareness as the artist seeks to remedy the fault – in which case every individual movement within the line drawing act become the operations – and once the reason for the fault has been found, the action may return to being an operation.

1.2.2.1.2 Application

As well as being recognised as difficult to understand, it is also recognised that Activity Theory may appear difficult to apply (Nardi, 1996a).

In her book, Bødker (1991) does not demonstrate the use of Activity Theory, nor does she provide general recommendations about the design of interfaces based on Activity Theory. Instead, she makes recommendations about how interfaces can be classified using Activity Theory and how Activity Theory could influence the design process.

An early recommendation is that the novice - expert dimension often used to define users is defined in terms of the level of activity at which their use of the computer takes place – for the novice, every step will be a conscious action whereas for the expert even the planning of the instance of the use of a piece of software may be operationalised.

Of particular importance to her were occasions in which the artifacts in an instance of human-computer interaction forced shifts in the object of actions. For instance, if the computer tool becomes the object of attention (a breakdown), there is some failure in the application. Focus shifts, more deliberate shifts in the subject-tool-object relationships than breakdowns, are also of interest. In her book, she shows how the nature of the interface to two different word processing packages alters the relationship between the subject and the object.

In her search for a way in which Activity Theory could be practically applied, she looked to immediate predecessors who had either considered computers to be tools which mediate individual actions or media through which communication to coordinate collective activities takes place. In the first case, the new interfaces must be designed in the light of how work is already practiced, considering the historical design of the task, and that the tool must be tested in the context of real work. In the second case, the important issue becomes how the interface alters the understanding of messages.

For designers, Bødker (1991) recommended a synthesis of these two alternative approaches, which expanded the use of computers as tools in such a way as to enable it to handle communication. This emphasizes prototyping methods, such as mock-ups and rapid application development tools, rather than formal description methods, because "we must be aware that the user interface only reveals itself, fully, in use" (Bødker, 1991, p. 142).

She claimed to be applying both the redefinition of the user interface inherent in the application of Activity Theory, and the design recommendations in the development of a tool for the development of computer systems to support both individual and cooperative work.

Later – in a section of Nardi's (1996a) book which was intended to overcome the criticism that Activity Theory was difficult to apply by providing concrete examples of its application – she (Bødker, 1996) showed how activity theory could be used to guide designers' search for relevant issues in the qualitative analysis of human-computer interaction.

The analysis included two phases, one the placing of the application in its context, and the second studying direct use of the application.

During the first phase, the researchers identified who was using the application, what the objects of their use were, and how they manipulated those objects. This ranged from managers who wanted summary statistics which were produced by their secretaries and field workers, to data entry secretaries who worked directly on individual records in a database.

The aim was to place the work historically, in the web of activities in the workplace, characterising the tools used, considering what support there was for individual activity and actions from the surroundings, identifying the objects, and considering the different roles the software must play.

During the actual analysis, they concentrate on situations in which users seemed uncertain about what was happening, and noted particularly the focus shifts during those situations.

When users were focussed on an activity or action, they asked what the purpose of the activity or action was, what the object of the action is, and what tool was being used to manipulate the object. If the activity involved cooperation, it was also questioned whether all the cooperators considered the purposes, objects and tools were to be the same.

When the focus shifted, Bødker (1996) noted the objects between which the shift took place, whether the shift was a breakdown or deliberate, and what had caused the shift.

Using these observations, Bødker (1996) was able to show how inconsistency in the content of menus in a word processing caused confusion, and how a second piece of software obstructed its users. She was able to recommend training or redesign which would overcome the specific problems.

Nardi (1996c) showed how using Activity Theory to frame the analysis of interviews of slide producers, which focussed on the users preference for specific slide-producing software or generic drawing packages, could turn the findings from simple ethonography into generalisable findings.

The study had found that different tools were used by different users, but more importantly that some users used both. The implication, supported by statements in the interviews, was that there were different types of slide-building tasks. Their initial analysis had drawn distinctions between levels of task which the slide-builders performed, but in retrospect she had realised that the definition of these levels of task were unprincipled and served no real purpose in defining what it was about the tasks which influenced the slide-builders' choice of software. The analysis of the slide-building process in Activity Theory's terms – taking account of the context inherited from higher levels, and the relationships between the human subjects and their objects – clarified the analysis.

Although Nardi (1996c) recognises that a simple transliteration of "task" into "object" and "subtask" into "action" would be undesirable, it can be seen that in broad terms, the task of building a particular type of slide (object) implies a context, which in turn influences the software (tool) which would be best suited to performing the subtasks (actions) of building the slides. Furthermore, the type of slide to be built depends for its context in the collective activity of the slide-builders' jobs.

Nardi (1996c) suggests that such a principled means of distinguishing between tasks and subtasks would benefit HCI and cognitive science in general.

1.2.2.1.3 Activity Theory in HCI

There are several potential benefits which Activity Theory has been argued to have for HCI.

Not least of these is Nardi's (1996a) contention that it could provide a shared, and specified, vocabulary to help in the comparison of studies and the propogation of results. A recurrent criticism of HCI studies is that results are not generalisable or applicable, and this might be remedied if the matter of the studies could be fitted into a larger framework of human-computer activity. Currently, HCI vocabulary is often used ideosyncratically in each study, and the concepts which underlie it, being underspecified, may constrain or confuse the analysis of the studies.

On the other hand, the Activity Theory vocabulary in English encapsulates complex concepts in poorly translated forms. This means that the force of the terms is difficult to understand, and even once they are understood, because the roles of the items which are being termed – subject, object and tool – change according to the context of their observation, it can be difficult to define those items within the an instance of the theory.

Another potential benefit of the theory, related to the choice of its vocabulary, is the redefinition of the role of the computer in the study of human-computer interaction. If Activity Theory were adopted, the field might be better redesignated "computer-mediated activity" (Kaptelinin, 1996a, 1996b), implying both that the role of the perfect computer should be that of mediating tool, not communicative partner as is implied by the term "interaction", and that the larger scale activities within which computer use takes place deserves attention.

As attractive as the designation of the computer as a tool is, however, it rests on several assumptions. The first is that computer interfaces should be "transparent" (Winograd and Flores, 1986); because the tool should be something whose use is more likely to shift towards being operationalised, than towards being the object of an action. Although it is generally accepted that interfaces which separate the user from the object on which they are working as little as possible are beneficial, an interpretative, and relevantly interactive, interface could also be beneficial[21].

It also assumes that computers are always involved at the action level and not at the activity level. Although in stereotypical computer use they may well be the tool, they may also be involved in the planning of activity – Bødker (1991) considers this before proposing that the tool model can be generalised to include communication at the activity level[22].

The clearest benefit which Activity Theory brings is the enforcement of a consideration of computer use's context, implied in the Kaptelinin's (1996a, 1996b) use of the term "activity". Not only does this remind researchers and designers that the context of computer use is important – for there have already been ample arguments for this (e.g. Winograd and Flores, 1986; Whiteside, Bennett and Holtzblatt, 1988) – but also provides a specification of the ways in which the context can influence actions.

Simplified, the design of the tools encapsulates the historical context of the development of the work within which the subjects of an action are involved, and the collective activity within which the individual's activity takes place influences the objects of the actions.

1.2.2.2 Contextual Inquiry

Contextual Inquiry is a set of guidelines intended to improve the gathering of customer's requirements. As such, it is primarily used in the design of new software, particularly as part of the process of Participatory Design[23] (e.g.

---

21 The Vellum software, as interpreted by Tognazzini (1992), provides an example.

22 The definition of expert systems may need particular attention.

23 Participatory Design is typically regarded as the involvement of the eventual user of a system in its design as much and as early as possible, and contrasts with the prevaling view of traditional system design, in which users were consulted late in the design process and little heed was paid to their views.
In fact Participatory Design, and by extension Contextual Inquiry, has several central tenets which take it beyond designs whose developers made a few visits to their users sites. These are summarised by Blomberg and Henderson (1990) as being

- The goal is to improve the quality of work life
- The orientation is collaborative
- The process is iterative

Too often, these tenets are ignored (e.g. www.infodesign.com.au), even when designers have the best intentions (e.g. Blomberg and Henderson, 1990).

Blomberg and Henderson, 1990; Good, 1992); rather than for building theories about human-computer interaction. On the other hand, to achieve its improvements over earlier methods of requirements gathering, it draws heavily from well respected academic techniques in ethnography, field research and participant observation.

The major exponent of Contextual Inquiry has been Holtzblatt and one of its best respected descriptions is Holtzblatt and Jones (1993), in which the principles behind its techniques and concrete guidelines on their practice are laid out[24].

1.2.2.2.1 Principles

Two principles are central to the application of Contextual Inquiry, understanding the **context** in which systems will be used, and that the design process is a **partnership** between the designers and the users[25].

1.2.2.2.1.1 Context

Understanding of the context in which systems will be used is important because without it a system cannot support a user's work. To achieve this understanding, during Contextual Inquiry, users are consulted while engaged in their daily activities.

The reason for this is the belief that people do not reflect on the process of their work while performing it, and retrospective recollections will produce abstractions and summaries which, without the touchstones of actual experience, will not capture the real failings and successes of current systems.

Holtzblatt and Jones (1993) contend that users' retrospective recollections will produce abstractions of their typical tasks, perhaps reflecting how they have been told tasks should be carried out rather than how they are actually carried out. Consulting users in their workplaces, designers can see how users' real work diverges from these abstractions; basing their own models on observations of particular, concrete, instances of the users' work.

---

24 More recently, Contextual Inquiry has been subsumed within Contextual Design (Beyer and Holtzblatt, 1998). The principles remain essentially the same.

25 Blomberg and Henderson (1990) even avoid the term "users" because it implies that those partners in the design process have little role other than using the systems, rather than the systems being one tool in the larger environment of their work.

Similarly, Holtzblatt and Jones (1993) contend, in retrospective recollections of their work, users will typically produce summaries which may ignore the actual features which make a difference to them. Users may not have a concrete awareness of many small faults which accrue to make their perception of a system negative, nor may they be aware of which aspects of features they consider beneficial are critical.

Interviewing users in their work context overcomes these issues by providing the users with touchstones which help them recall their work experiences; turning the process into one closer to recognition than recall. Things in the environment of the users' work provide the designer with hints about relevant questions to ask, and may remind the user of issues they had grown accustomed to working around.

Holtzblatt and Jones (1993) comment, for instance, on a detached and a contextual interview with a user of a text editor. In the detached interview, the user said that there were "little things that get in the way", but in the contextual interview the lack of automatically wrapping text became a "biggy" which the user wished the software engineers would fix.

1.2.2.2.1.2 Partnership

Partnership is important in Contextual Inquiry because its exponents believe that only through dialogue with users can their experience be gathered, and it is only if the users consider themselves partners that there will be effective dialogues. This is embodied in the Contextual Inquiry catchphrase:

The user is the expert.

The use of Contextual Inquiry implies background assumptions and techniques which are intended to foster these partnerships.

One such assumption which must be recognised is that that users are experts, and this has two implications. The first is that the designer is not there to provide answers or solve problems, but to learn from the user; and this liberates the designer from the belief that they should know everything before they start. A second is that it should prevent the designers from interpreting their observations in terms of their own experience; instead asking the users to interpret in their own.

A secondary implication of the recognition that the users are the experts is that they have all the information, even to the extent that the designers may not know the correct questions to ask, and the traditional interview structure must be avoided. To avoid interpreting their observations in their own terms, designers must not even come to the contextual inquiry with too direct questions to ask, in case those questions themselves influence the description of the work gathered.

Instead, Holtzblatt and Jones (1993) say, care must be taken to share control during the inquiry; to create equal dialogue. Users must be allowed to lead the conversation while the designer asks open questions, at the most intended to confine the users conversations to the areas of concern.

The central question to be answered during almost any inquiry is, "What is going on here?"; and it is one which has to be answered by the designer and user as a pair. Therefore, according to Holtzblatt and Jones (1993), one of the most important aspects of an inquiry is the building of shared meaning. This means that the designer must pay attention to what their user is saying, rather than trying too carefully to construct answers to their own questions.

The users answers must be acknowledged, registered, and if no points are immediately raised, the conversation can be stopped or the topic changed.

The aim of these techniques is to foster reflection on and engagement in the work experience both for the user and for the designer. While the users work, both they and the designer should be engaged in the work, and when they stop, both should feel the need to reflect on what has happened.

1.2.2.2.2 Practice

The core of Contextual Inquiry is the contextual interview during which information is gathered about a system's prospective users, and a considerable amount of any description of Contextual Inquiry is an overview of the practical techniques involved in conducting those interviews. Tutorials in these techniques are also common at major Human-Computer Interaction conferences.

These range from the reminder that thank-you notes be sent after a contextual interview, through identifying the people who expect to benefit from the system and determining what the focus of the interviews will be, to guidance in the sorts of questions to ask during the interview and how to analyse the interviews.

Details of this process will not be repeated here, because most are derived from the principles already described. The one additional point worth noting is that multiple designers are expected to participate in any instance of contextual inquiry.

This is claimed to have many benefits; from generating shared views about what is required during the post interview analysis, to espousing greater group ownership of the products.

1.2.2.2.3 Application

To date, as would be expected from the introduction in Section 1.2.2.2, Contextual Inquiry has been used almost entirely for designing real products rather than generating theory.

Techniques which borrow heavily from the principles of Contextual Inquiry are reported by several of the practitioners in Good (1989).

Blomberg and Henderson (1989), albeit in a self-admittedly flawed fashion, employed it in the design of a user interface development tool.

Good (1992) used it in the design of a tool for the direct manipulation of computer models of molecular structure in virtual reality.

Holtzblatt and Jones (1993) report its use in the design of a text editor and suggest that they have used it during the design of many pieces of software for Digital Equipment Corporation's customers.

1.2.2.3 Contextual Methods in the Thesis

The two methods reviewed in this section provide different reasons for studying computers in context. Both provide principled frameworks for the study of computer use in context, but the frameworks promote different uses of that context.

There are several differences between Activity Theory and Contextual Inquiry, which are not relevant here:

- Contextual Inquiry specifies its practice whilst Activity Theory does not;
- Contextual Inquiry is presented as a completed technique whereas Activity Theory is recognised to be still evolving;
- Contextual Inquiry the user is the centre of analysis whereas Activity Theory hinges on the tool;

More important is the different way of regarding the context which the two frameworks imply.

Activity Theory's framework draws the researcher back from the software, to see how it fits into the greater scheme of the work in which it is being used. In this way, it detaches the researcher from the specifics of the software and promotes an awareness of how the history and culture of the work which is being performed influences the software which is used.

Contextual Inquiry implicitly recognises the influence of the larger work within which computer use takes place, but rather than leading the researcher towards a principled specification of that context, it uses the context to guide the elucidation of specific details of the software from the user.

In effect, in Activity Theory, the researcher looks away from the computer, trying to capture how the context of the work influences the use of the software, whereas in Contextual Inquiry, the researcher looks over the user's shoulder to capture how specific issues in the design of the software influence the performance of their work.

For instance, when Nardi (1996c) applied Activity Theory in the analysis of slide-makers use of different software, she was not influenced to look closely at the differences in the software, but to note the differences in the activities which the software was employed within. On the other hand, Holtzblatt and Jones (1993) are drawn to analyse the benefits and disadvantages of the specific features in a text editor, with the context of use serving as a touchstone to prompt the user's own mention and interpretation of the features.

What both bring is a framework for the description of how computers are used, in the case of Activity Theory ensuring that the description fits into a larger model of human practice and in the case of Contextual Inquiry ensuring that the researcher does not impose their own prejudices on the description.

In concentrating on the building of descriptions, and for all that it has been proposed that both be used in a design process, their role in that process is essentially evaluative. Activity Theory has been used to discover how software which is already in place is being used (Bødker, 1996; Nardi, 1996c[26]), and Bødker (1991), who explicitly recommended its use in design, envisaged that use to be in an iterative cycle of prototype testing. Contextual Inquiry is typically use early in the design process to find out what works in the software users already use and what may be causing them problems (Blomberg and Henderson, 1989; Good, 1992).

This is far from a disadvantage, however. The study of computer use in context early in a design or research programme ensures that the design meets the users requirements or that the research investigates issues in design which are actually relevant.

In a sense, such contextual studies are actually the starting point of this thesis. It was observation – influenced by the author's basic training in sociological and ethnographic methods during the early part of his research – of British Telecom customer service agents, and discussion with their managers and the designers of their software, which lead to the decision to study tripartite communication with computers in the first place.

The discussions indicated that research was needed into the inter-relationships between the three parties, and this was supported by a review of the literature on tripartite communication (Section 1.1). Although there had been persuasive observations indicating that the design of the agents'[27] software could have effects which permeated to their customer, there was no basic evidence from controlled laboratory experiments to support it. One aim of the research described in this thesis was to provide that basic support.

In this way, Nardi's (1996b) recommendation that laboratory-based work be preceded by appropriate field studies was followed, and the social science paradigm of using controlled laboratory-based studies to support the analysis of observational studies (Richardson, 199) would be adhered to.

---

[26] Raeithel and Velichovsky's (1996) study, which could be seen to contradict this argument because Activity Theory apparently underlies predictions about improvements in software, does not in fact do so because the predictions are based on the importance of joint attentional focus which was admitted by the researchers to have parallels in Western psychology, and could have easily have been predicted from studies of gaze – particularly those of "Gaze Awareness" (Daly-Jones, Monk and Watts; 1998) – or research into the building of co-reference in dialogue (see Section 1.3.1.1.2).

[27] Or doctors', or call takers' ...

Furthermore, every experimental chapter begins with a description of similar observations and discussions within which the specific design issue that is studied within that chapter arose. In this case, the aim was to ensure that the development of a model of tripartite communication was being furthered through the study of issues which had their own practical relevance within tripartite customer service interactions.

1.2.3 Use of Language in HCI

Because there has been no use of language in the formal evaluation of interfaces used in tripartite communication nor in the couple of formal experimental HCI studies reported in the Section 1.2.1 nor explicitly in the studies mentioned in the last section, it could be inferred that there is no use of language as an evaluative tool in formal HCI research, but this would be a misrepresentation.

There are in fact a number of ways in which it is used including verbal protocols (Ericsson and Simon, 1980), naturally generated expressions of users' representations of systems (Suchman, 1987) and the analysis of the dialogues between users and their systems applying the findings and techniques of sociology's conversation analysis (Luff, Gilbert and Frohlich, 1990).

It is important to review these cases for the thesis because Section 1.1 indicated that formal methods of analysing the tripartite communications between two humans and a computer did not include the dialogues between the humans and methods which did include the dialogues lacked formality, for all that they were rigourous. The aim will be to find evidence of formal use of language in HCI.

1.2.3.1 Verbal Protocols

Verbal protocols are in Anders Ericsson and Simon's (1980) sense think-aloud reports, which are typically used to infer thought processes in problem solving tasks.

Anders Ericsson and Simon's (1980) main aim and success was to demonstrate that these could have more experimental validity than the introspection used in early psychological research. The way they achieved this was to build a taxonomy of verbal protocol gathering techniques in which instances of verbal protocol use can be classified on the three scales of concurrent-retrospective reporting, individual-group reporting and unaided-aided verbalisation.

The concurrent-retrospective scale distinguishes between the time at which reports are gathered relative to the task being reported. The individual-group scale divides between those reports which are generated by groups who perform tasks, and those who perform tasks and report individually. The unaided-aided verbalisation scale refers to the degree of intrusion of probing which is given by the person gathering the reports, from prompts such as "keep talking" to those such as "Why did you do that?".

From meta-analysis of research in cognitive psychology they could recommend the use of concurrent reports because retrospective ones allow additional processing, such as a procedure of tidying up quick-and-dirty mechanisms for achieving goals, but could not give specific advise as to the benefit of unaided verbalisations over aided ones, or individually produced ones over group ones.

In HCI, verbal protocols typically involve subjects talking aloud about the thoughts they are having while trying to achieve goals with an interface while they are using it, from which researchers infer how the actual design of the interface interacts with their subjects' mental representation of it.

They are in common use in practical evaluations, but direct reference to them in the academic literature is less frequent. One example is Henderson, Podd, Smith and Varela Alvarez (1995) in which four methods of software evaluation which involved user participation were compared; logged data, questionnaires, interviews and protocol analysis.

The researchers investigated the strengths and weaknesses of each method in the evaluation of three pieces of common desktop software (spreadsheet, word processor and database).

1.2.3.2 Naturally Generated Protocols

Suchman (1987) used a similar but distinguishable method when she studied interaction with a photocopier "expert help system" by inviting two people to collaborate in performing novel but relatively simple tasks.

Although either could have performed the actions alone, she argued and found that in the interest of collaboration, each participant would share their beliefs about the task they were engaged in; their current state, their goals and how to achieve them. Whereas the protocols Anders Ericsson and Simon (1980) referred to were explicitly requested by researchers and often prompted, Suchman's (1987) were spontaneous and so she consider them "naturally generated protocol".

In studying the total interaction, or the sense of "shared understanding" as she termed it, she took into account four resources; the users' plans and intentions as represented in the protocols, the actions the users made with the machine, the way the help system and machine reacted to these actions, and the designer's rationale behind these reactions. The discussion between the two subjects showed how the system's actions were interpreted and clarified the users' actions.

For instance, in the sequence reproduced in Figure 1/1 it can be seen how the exchange between the two humans shows their representations of how the interaction will proceed, when the system's actions do not match those representations.

THE USERS		THE MACHINE	
Not available to machine	Available to machine	Available to the user	Design Rationale
B: Okay, and then it'll tell us,			
okay, and:: It's got to come up with the little start thing soon. (pause)		1. Instruction to wait while the machine prepares for the photocopying of bound documents.	Selecting the procedure
Okay, we've done all that. We've made our bound copies. (pause)		2. Instructions in using Bound Document Aid and Recirculating Document Handler.	Instructions for copying a bound document: Accessing the Bound Document Aid.
A: It'll go on though I think. Won't it?			
B: I think it's gonna continue on, after it realises that we've done all that.			

Figure 1/1 Example of naturally generated protocol. Reproduced from Suchman (1987, p. 128[28])

[28] The entries in the "Available to User" column have been expanded for clarity -- in the original they are anchors to representations of screen displays.

The users are trying to make a two-sided copies of a bound document. Their actions on the help system's initial screen has triggered instruction in photocopying bound documents. The users completed those actions, and – as their final comments indicate – expect further instructions, but the system which has not kept a record the initial intentions for photocopying the bound document, awaits the users explicitly indicating they are ready to return to the initial screen.

The users have brought their expectations about the structure of human-human communication to their interaction with the photocopier, and these are not matched by the system's plans.

1.2.3.3 Conversation Analysis in HCI

In Luff, Gilbert and Frohlich (1990), the use of sociology's Conversation Analysis is recommended as an alternative to the prevailing cognitive science approach to human-computer interaction. The predominant proposal in the collection's articles is that it is possible to inform design decisions and remove some of the onus on software to second-guess its users by recognising what Conversation Analysis knows about the initiation, structuring, repair of breakdowns and ending of human-human conversation.

This has intuitive value based on the habitual consideration of human-computer interaction as conversation with a computer; embedded in such terms as "man-machine dialogue", "conversational metaphor" and indeed "human-computer interaction" itself (Greatbatch et al., 1983; Robinson, 1990).

The practical value was demonstrated by Norman and Thomas (1990).

Norman and Thomas (1990) proposed that Conversation Analysis would provide a set of findings and methods which could at the least inform HCI, but more progressively provide a principled approach which could overcome many of the failings in its current approaches; though not providing a panacea. They argued that HCI's multi-disciplinary nature was holding it back because it lead to pragmatic solutions dependent on specific technologies which the rapid development of interactive technology rendered redundant or results which were too general and could not be applied without considerable interpretative effort by designers.

One of their arguments for the use of conversation analysis was that, as Suchman (1987) had shown, people bring what they know about human-human communication to their interaction with machines; and this is the subject matter of conversation analysis. A second was that conversation analysis would allow a specification of the nature of human-computer interaction which was independent of technology.

Conversation analysis would permit the building of interactions which progressed in accordance with users' expectations, based on the principles of human-human communication derived from rigourous conversation analysis.

As an example, they suggested that what was known about the repair of misunderstandings in human-human communication could be used to improve the design of a sample file opening dialogue box. In video recordings of users interacting with a system they noticed that people would over-generalise the use of this dialogue box to open any class of file when it could only open files of a specific class.

They recognised that there were a number of ways to improve this situation, but concentrated on how findings from conversation analysis could prevent users repeated use of the dialogue box expecting that the repeat would unearth some failure in the first pass. The dialogue box made no response indicating what sort of files were available, merely feedback indicating that files could be opened. In human-human communication, when a request is made it is expected that there will be an appropriate response, and if not some indication will be made about why there was no appropriate response which would permit the request to be reformulated.

1.2.4 Summary

The first section in this review of HCI (Section 1.2.1) was intended to supplement what had been seen of HCI's techniques in Section 1.1, concentrating on the application of formal techniques.

It showed that the preoccupation in laboratory-based HCI studies had been with single users interacting with single computers, and that the methods of evaluating these had been fairly simple; analysing completion times and error rates and categorising errors to determine mismatches in user expectations and system design.

The study of completion times and error rates has already been used in the study of tripartite communication, by Gray et al. (1990, 1992, 1993), and did not lead to any insight into the unique nature of the tripartite communication. The categorisation of errors is only loosely formalised, and so grants no great benefit in the search for formality. Although the building of formal models of the communication – which was often the stated aim of these studies – might be beneficial, since the background of findings about tripartite communication is sparse such a development would be presumptuous.

The second section (Section 1.2.2) concentrated on contextual methods in HCI, whose proponents have argued that the use of computers cannot be understood without reference to the whole scheme of work, and the cultural and physical situations in which they are used[29].

Some of the principles which underlie the contextual methods were used in those studies of trpiartite communication which were reviewed in Section 1.1. More importantly within the thesis, they are also the principles which were used, in a less structured fashion, during the observation of real customer service work which preceded both the selection of the thesis topic and the topics of investigation in each experiment.

The third section (Section 1.2.3) reviewed how language has been used as an evaluative or experimental tool in HCI, searching for ways in which it has been used in formal studies or controlled experiments, to augment its currently informal use in the study of tripartite communication (see Section 1.1.4).

Verbal protocols could not be produced concurrently with the task the human participants are engaged in because that already involves conversation, retrospective reporting is not recommended because it allows subjects to "tidy" their memories too much.

On the other hand, naturally generated protocols such as those Suchman (1987) used can occur spontaneously, they can been recognised in the study of "breakdown" in which mediating technology is mentioned in the dialogues it

29 Activity Theory (Section 1.2.2.1) goes some way further than this, by providing a theoretical framework through which the influence of context is passed. In not specifying the means which are used to observe an instance of activity, however, it does not preclude laboratory experimentation; Nardi (1996b) notes that field observation made within an Activity Theory basis may compliment laboratory experiments.

mediates (Wright and Monk, 1990; see Section 4.1.2), but both the application of these and of Conversation Analysis in the ways suggested in Luff, Gilbert and Frohlich's collection (1990) suffer from being essentially no more formal than the sociological methods already in use.

Although all three methods are rigourous and promise to produce concrete guide-lines from soft data, they provide no benefit for the current purposes.

This chapter will now move on to a general review of what is known about human-human dialogue, with coverage of the effects of mediating technology on these dialogues providing several examples of how language can be used formally.

1.3 Human-human Dialogue

The search for means to study the unique nature of tripartite communication with formal means continues in this section with a review of what is known about the human-human leg of the communications when taken in isolation.

The first section (Section 1.3.1) and its subsections will describe what is known about the most basic features of human-human dialogue concentrating on how coherence is maintained. The second section (Section 1.3.2) and its subsections will describe the varied effects of mediation on human-human dialogues. This will cover such primary effects as people's perception of the media and effects on the results of the tasks they are involved in as well as more in depth effects on the process of the dialogues; some of which measure the way in which coherence is affected by mediation.

The conclusion will be that formal methods for investigating the effects of technological mediation on dialogue will be useful in the investigation of tripartite communication.

1.3.1 Natural Dialogue

Although there is some debate as to its precise definition (Anderson, 1992) and extended sequences of interaction may be required to reach it (Clark and Marshall, 1981) one of the most noticeable features of dialogue is its coherence. Any more than momentary breakdowns in this coherence become comic, as is exemplified by the following sequence from the film Duck Soup, quoted in Akmajian et al. (1984) (see Example 1/1).

Groucho:	Just for that, you don't get the job I was going to give you.
Chico:	What job?
Groucho:	Secretary of War.
Chico:	All right, I take it.
Groucho:	Sold!

Example 1/1 Example of comic incoherent dialogue.
Reproduced from Akmajian et al., 1984.

Although there are two processes taking place simultaneously in any dialogue, the specification of that information which is being shared, and the establishment of the belief that the collocutors share mutual knowledge to perceive that information in the same way (Clark and Schaefer, 1987), there is a background of sociological and psychological research which suggests that it is the development of mutual knowledge which is at the heart of coherence. Only with mutual knowledge, and therefore when any utterance is interpretable in the same way by all participants, can dialogue be considered coherent.

1.3.1.1 Features of Dialogue Coherence

Research in psychology and sociology has shown that there are a number of ways in which participants must co-ordinate in order for a conversation to be coherent, and the aim in this first section will be to describe them, so that later sections can describe how they have been explained. The most commonly referred to three are the co-ordination of turn-taking, the construction of shared references and the repair of errors in understanding. In terms of how these fit into the process of conversation, turn-taking is involved with how people decide when they can speak, reference is involved with how people decide what to say and repair is involved with how speaker and hearer react to what has been said.

1.3.1.1.1 Turn-taking

The study of turn-taking essentially began with Sacks, Schegloff and Jefferson (1974) who found that one of the most striking features in a corpus of naturally occurring conversations they had access to was that even when there was no artificial structure – such as might be seen in a union meeting or lecture – imposed on the ordering of inputs to dialogue, the organisation of turn-taking was undeniably evident. They noted 14 features which any model of the turn-taking behaviour would have to account for (see Figure 1/2).

1. Speaker-change recurs, or at least occurs.
2. Overwhelmingly, one party talks at a time.
3. Occurrences of more than one speaker at a time are common, but brief.
4. Transitions (from one turn to a next) with no gap and no overlap are common. Together with transitions characterised by slight gap or slight overlap, they make up the vast majority of transitions.
5. Turn order is not fixed, but varies.
6. Turn size is not fixed, but varies.
7. Length of conversation is not fixed in advance.
9. Relative distribution of turns is not specified in advance.
10. Number of parties can vary.
11. Talk can be continuous or discontinuous.
12. Turn-allocation techniques are obviously used. A current speaker may select a next speaker (as when he addresses a question to a next party); or parties may self-select in starting to talk.
13. Various "turn-constructional units" are employed; e.g., turns can be projectedly "one word long", or they can be sentential in length.
14. Repair mechanisms exist for dealing with turn-taking errors and violations; e.g., if two parties find themselves talking at the same time, one of them will stop prematurely, thus repairing the trouble.

Figure 1/2 Fourteen features of turn-taking in human-human dialogue.
Reproduced from Sacks, Schegloff and Jefferson (1974).

Sacks et al. (1974) hypothesized that two components would be used to meet these criteria; a turn construction component and a turn allocation component. The turn construction component determines which type of utterance (sentential, clausal, phrasal or lexical) a speaker will use to fill their turn, whereas the turn allocation component determines how turns are passed between speakers when a "transition relevance place" is reached.

Essentially, the turn allocation component runs on three rules (see Figure 1/3).

1. If the current speaker tries to select a speaker to follow them, that speaker is obliged to take that turn.
2. If the current speaker does not select a following speaker, then self-selection by other speakers is optional – the first speaker to start gets rights to the turn.
3. If the current speaker does not select a following speaker then they may continue to speak unless interrupted.

Figure 1/3 Sacks, Schegloff and Jefferson's (1974) turn allocation rules.

These rules are recursive, such that they apply equally well at the next and following transition relevance places.

What the model describes is a locally and interactionally managed, party administered means for transferring turns. It is locally managed because the only criteria for controlling the transition of turns develop within the two turns which encompass that transition, it is party administered because only the participants in the conversation determine the size and order of turns, and it is interactionally managed because turn size and order can be adjusted by any of the participants at any time.

What this means is that any occurrence of turn transition is a co-operative, not a one-way action. Sacks et al. (1974) call their model a "recipient design" because:

> *"the talk by a party in a conversation is constructed or designed in ways which display an orientation and sensitivity to the particular other(s) who are the co-participants" (Sacks et al., 1974, p. 727)*

Whether it occurs when the speaker explicitly hands the turn to another, for which the taker has to be aware it is being given, or whether the turn is being wrested from the current speaker, in which case that speaker must agree to cede the floor, the transition of turns requires co-ordination between the participants.

The work of Sacks et al. (1974) was seminal to the development of Conversation Analysis, which was referred to in Sections 1.1.2 and 1.2.3.3.

1.3.1.1.2 Reference

One of the critical points in conversation is when a new object is introduced to a conversation, for it must be determined that the item being referred to is understood to be the same by both participants. For instance if the speaker wishes to refer to a particular pair of shoes, they must introduce those boots to the conversation in such a way that the hearer can understand not only that a pair of shoes is being introduced, but also which pair. If the object is not suitably introduced, then any clauses which depend on it will fail to be understood correctly. The introduction and establishment of a co-reference to an object has been termed the use of "preliminaries" by Schegloff (1980).

The techniques which have been discovered whereby conversants build this co-reference can be divided into two categories; a "telling" form and questions (Schegloff, 1980). In "telling" forms, the speaker specifies the object in such a way as to remove ambiguity, whereas in questions the speaker presents one means to determine reference and allows their partner to further that specification. Either of these might be followed by a side-sequence which is used to further determine the correct reference, but the original intention at the onset of the presentation is where the difference lays.

"Telling" form presentations of reference include referential instalments, indicative gestures and alternative descriptions (Clark and Brennan, 1991).

Referential instalments occur when a speaker refers to an object, and establishes that the hearer has understood the same object before continuing with a description of the relevance of that object. For instance, in English a speaker may use left-dislocation, (see Example 1/2), to establish the reference to an object before saying what needs to be done with it:

S: Okay now, the small blue cap we talked about before?

J: Yeah

S: Put that over the hole on the side of that tube –

J: Yeah

S: – that is nearest to the top, or nearest to the red handle.

Example 1/2 Example of left dislocation.
(Cohen, 1984; quoted in Clark and Brennan, 1991)

Indicative gestures occur when the reference is achieved by deixis within the real world, by one speaker pointing, looking or touching the item they are referring to.

Alternative descriptions seem to occur in side-sequence where the partner is not fully sure they understand a reference and use a different set of features to identify them. For instance, in Example 1/3 the second speaker uses the person's name to confirm that they have understood the description given by the speaker:

A: Well, that young gentleman from - ((the park))

B: Joe Joe Wright you mean? - - *(laughs)*

A: *yes, (laughs) yes*

B: ((God)), I thought it was old Joe Wright who((‘d))
walked in at first

Example 1/3 Example of alternative description.
Reproduced from Clark and Brennan (1991, p.136)

Although B's alternative description is presented as a question, used to confirm that the description is correct, it is the presentation of an alternative description which gives evidence that the reference has been shared.

Question style presentations of reference include introductory questions and trial references. They occur when the speaker doubts the referring expression they are forced to use, either because they are not sure their use is correct or because they are not certain the hearer will understand it.

Introductory questions are when the speaker explicitly introduces objects with questions such as "Do you remember ... ?" They allow the hearer to indicate that they have understood the reference before the conversation continues.

Trial references are when the speaker finds themselves introducing an object in mid-sentence that has not previously been referred to. When this occurs, the speaker introduces the referring term in its normal place in the utterance, but gives it a "try marker", querying rising intonation, and leaves a short pause for the hearer to indicate whether they understand the reference.

In Example 1/4 the phrase "A man called Annegra?" is a trial reference, which the second speaker corrects with "Yeah, Allegra" to provide the correct reference.

A:	So I wrote off to Bill, uh who ((had)) presumably disappeared by this time, certainly, a man called Annegra?
B:	Yeah, Allegra.
A:	Allegra, uh replied,

Example 1/4 Example of trial sentence.
Reproduced from Clark and Brennan (1991, p. 138)

The important feature of all these forms of introduction of reference is that they depend on interaction between the participants. Even "telling" forms require an acknowledgement from the hearer that the reference has been understood, otherwise a side-sequence such as the alternative description would be initiated. The dialogues do not simply develop as one speaker presents information which they believe will be understood, then later discovers that some failure has occurred and must try again. At each introduction of information, checking occurs to determine that knowledge has been shared.

1.3.1.2 Models of Dialogue

Having described several features of dialogue which demonstrate how coherence is maintained, typically by interaction between involved partners, the next subsection will describe models which have been proposed to explain how conversation takes place, and why the features of coherence we have seen work in the way they do.

1.3.1.2.1 One-way models

Although there are notable earlier attempts to study the interaction within dialogue (e.g. Bales, 1950), its modern study of interaction in dialogues often traces its ancestry back to Austin's presentation (1962) and Searle's (1969, 1979) specification of speech act theory in the 1960s and 1970s. This is because their work has been taken to show that when a speaker says something they are not merely transferring information, but trying to affect the world, particularly their correspondents or collocutors, in some way. It is unlikely, however, that Austin would have understood the way his programmatic suggestion has developed.

His intention was to show that there were occasions on which the mere utterance of a sentence was equivalent to making an action in the world. Indeed one of the defining features of Austin's speech acts was that they could be achieved equally well by performing an action in the world as by uttering the speech act. For instance, war can be declared by a direct military strike as well as by a formal declaration.

Searle took this idea further and built a five class taxonomy into which speech acts could be classified (assertives, directives, commissives, expressives or declaratives). Although he agreed that almost every utterance could be considered one of these five classes, he claimed to be unhappy with its unrestricted use, for instance in coding dialogue. Indeed, although many people who code dialogue pay lip service to "speech acts", the taxonomies they use are frequently very different from Austin's and Searle's.

There are two probable reasons for this, and Austin's intentions and Searle's misgivings seldom figure, which are an uncertainty as to whether a scheme based on such unusual situations as declaring war and naming ships could be applied to everyday dialogue (Schiffrin, 1994), and that because the scheme deals with utterances in isolation it is incapable of dealing with the to-ings and fro-ings intrinsic in dialogue (Clark and Brennan, 1991).

This failing is shared with all early models of communication, such as the code or conduit model (Reddy, 1979) or the related inferential model (Bach and Harnish, 1979).

In the conduit model, the speaker is viewed as a transmitter who encodes a message into a form which can be transferred through the medium available, and the hearer is a receiver who decodes the signal they receive to recompose the message the speaker intended. The conduit model was successful because it explained how messages could be conferred, how if the hearer failed to decode the signal the act of communication had failed, and reflected the metaphors of communication entrenched in language itself (Reddy, 1979) such as "Try to *get* your thoughts *across* better" or "Let me know if you *find* any good ideas in this essay."

The inferential model added the ability for the hearer to see beyond the encoded signal to achieve some understanding of the intended message, essentially using Grice's maxims of Relevance, Sincerity, Truthfulness, Quantity and Quality (Grice, 1975) in the decoding stage. This explains how hearers can understand such indirect messages as irony, the illocutionary force of sentences such as "Can you pass the sugar?" when used as a request not a query, or the institutionalised "speech acts" Austin (1962) described.

One reason for the success of these models was that the study of spoken language is relatively recent, being preceded by the study of made-up sentences and of the written word (e.g. Saussure, 1990, p.24), and without that there could be little interest in conversational behaviours (Schiffrin, 1994).

1.3.1.2.2 The Contribution Model

The interactive to-ings and fro-ings of dialogue are, however, explicitly modelled by Herbert Clark's contribution model of dialogue (e.g. Clark and Schaefer, 1987). He and his collaborators have put forth and provided evidence for a model of dialogue and communication which takes as the basic unit of communication not the individual utterance but a "contribution" (Clark and Schaefer, 1987), which is a single interaction between the speaker and their audience. The aim in any of these contributions is to understand or develop the "common ground" shared by the participants (Clark and Brennan, 1991).

Common ground is the system of shared knowledge, beliefs and assumptions which is necessary for the participants in a conversation to coordinate the content and process of their utterances. Although a great deal of this common ground is brought into conversations as cultural conventions, it must also be updated on a moment-by-moment basis by a process known as grounding.

Grounding implies that any single contribution to conversation will have two parts which referred to as the presentation phase and the acceptance phase (see Figure 1/4).

Presentation phase: A presents utterance *u* for B to consider. He does so on the assumption that, if B gives evidence *e* or stronger, he can believe that she understands what he means by *u*.

Acceptance phase: B accepts utterance *u* by giving evidence *e* that she believes what A means by *u*. She does so on the assumption that, once A registers that evidence, he will also believe that she understands.

Figure 1/4 The presentation and acceptance phases in the contribution model of human dialogue.

In the minimal case, if a speaker A makes a statement the hearer B will give some positive evidence that they have understood it, for instance by acknowledging that they have heard, giving continued attention to show that they are up to date and willing to hear more, or initiating a relevant next turn.

If one of these forms of response is not given, then negative evidence has been given that the hearer does not believe they have common ground with the initiator, and an insertion sequence (Schegloff, 1972) in which they try to ground themselves is initiated[30]. Clark and Brennan (1991) use the example in Example 1/5 to demonstrate this:

Alan:	Now, - um, do you and your husband have a j- car
Barbara:	- have a car?
Alan:	Yeah
Barbara:	No -

Example 1/5 Example of negative evidence for grounding. Reproduced from Clark and Brennan (1991, p. 130)

After Alan's utterance, Barbara has not understood the question and so initiates an insertion sequence which she intends to rebuild local common ground. With Alan response, and embedded acceptance phase, of "Yeah", this ground is rebuilt and the acceptance phase to Alan's questioning presentation phase can be completed with her "No -".

This interaction is explained, according to Clark and Schaefer (1987) by the grounding criterion (see Figure 1/5)

Grounding Criterion	The speaker and addressees mutually believe that the addressees have understood what the speaker meant to a criterion sufficient for current purposes

Figure 1/5 The grounding criterion.

According to Clark and Schaefer (1987), it is the satisfaction of this criterion which is one of the most important goals of conversation, and it explains both

30 Although it is not the structure for dialogues presented in Clark and Schaefer (1987), a potentially important implication of the Presentation phase/Acceptance phase structure, is that dialogue could be interpreted as following the binary branching X-bar structure which is almost universally accepted in studies of syntax. This is worth noting because it conflicts with other models which presume a ternary structure, such as Power's (1979) or Labov and Fanshel's (1977). Such a model might not easily support simultaneous backchannels however.

how people recognise when they are in trouble and how they get out of trouble through the process of mutual questioning and acceptance.

The grounding criterion on its own does not avoid the theoretical problem of infinite recursion inherent in any hypotheses involving mutual knowledge. The basic tenet of any mutual knowledge hypothesis was best described by Harman (1977) who stated that for any speaker to generate an utterance they must assume, where p is the set of information necessary for comprehension of that utterance:

(i) The speaker and the audience mutually know p.

and

(ii) (q) The speaker and the audience mutually know p and that q.

The problem comes when it is realised that there must always be a further level of assumption in which the speaker and audience believe they mutually know the previous level, right down to the level at which mutuality itself can be assumed (Sperber and Wilson, 1986).

The grounding criterion does not directly avoid this because the level "sufficient for current purposes" is far from specific. Clark, however, adds the principle of least collaborative effort (Clark and Wilkes-Gibbs, 1986) in an effort to resolve this problem (see Figure 1/6).

The principle of least collaborative effort	In conversation, the participants try to minimise their collaborative effort – the work that both do from the initiation of each contribution to its mutual acceptance.

Figure 1/6 The principle of least collaborative effort.

This principle, based on Grice's (1975) maxims of quantity and manner, but taking into account the pressure of time, errors and ignorance of collocutor, explains how grounding in the concrete matter of discussions can be reached even if lower levels of mutuality have to be assumed.

The combination of the grounding criterion and principle of least collaborative effort explain how a speaker will deal with believed gaps in mutual knowledge, by suggesting that a speaker will use tactics such as try references[31] to enlist their partner's aid in building their own understanding of common ground. Through extensive study of spoken corpora they have shown that this model can be applied to explain behaviour in real world dialogues.

1.3.2 Technologically Mediated Dialogue

In the last section, the most basic features of human-human dialogue were described, and it was contended that the maintenance of coherence in dialogues was central. In this section the effects of mediation on human-human dialogue will be covered. Again, anecdotal evidence and comedy can capture these effects. Anecdotally, in the former Soviet Union, where telephone systems were basic, until recently the roles of caller and called were defined, and the called typically responded with only "Yes" or "No" to the caller's questions. In the Ealing comedies, when using a telephone Irene Handl frequently used exaggerated formal accents and structures, termed hypercorrection by Labov (1966), intended by her characters to display increased social status but ironically emphasising their lower middle class background to the audience.

In academic research, far more differences have been found. Although the coherence of dialogues is affected, and several evaluative tools depend on this to show the difference between face-to-face communication and various media, mediation can also have an effect on the outcome of tasks which are performed over them and be perceived differently by the users.

These effects are most clearly seen when media are grossly divergent, for instance the distinctions between spoken communication and written communication has been well-documented (e.g. Chapanis, 1977; Halliday, 1989), but even within these categories differences have emerged. Because the current interest is in human-human dialogues between customer service agents and their customers, and these dialogues are overwhelmingly spoken, the concentration in this review will be on media which permit synchronous spoken dialogue, and

---

31 A try reference is a reference to an object which is presented with a try marker, such as a questioning inflection, to indicate that the speaker is unsure that they are using the correct referring expression or that they do not know if the hearer will recognise the referring expression.

differences with the multitude of other variations of media will be ignored. For rigour, more than just face-to-face and audio-only communication will be covered, and studies which have included video-mediation or in which subjects are copresent but obscured from each other will also be included.

The section has several roles within the thesis. Firstly, since the majority of tripartite communication takes place in customer service and the majority of this is mediated, typically be telephone, it provides a background to this aspect of tripartite communication. Secondly, it represents another example of how human-human dialogue and technology interact which may be informative in the current case. Thirdly, it includes specific reference to occasions in which human-human dialogue has been used to evaluate the effects of technology on its users; potentially providing methods which could be applied in the formal study of tripartite communication. This last role is the most important to the thesis because, as was seen in Section 1.1.4 and 1.2.4, the current studies of tripartite communication lack formal study of the dialogues between the human participants.

As with the section on natural dialogue, the structure will be a description of the features of dialogue which have been shown to be affected by mediation, and then a description of models which have been proposed to capture the relationships between these features.

1.3.2.1 Features of Dialogue Affected by Mediation

The features of dialogues which can be affected by mediation are succinctly captured by the three P's of perception, performance and process, which were also seen as measures of HCI. Perception is the category which deals with the participants' attitude towards the results of the dialogues they have over varying media and towards the media itself. Performance is the more objective measure of how well the tasks taking place during the dialogues are performed, for instance how long a goal-oriented task takes to complete or the quality of a decision made.

Process is the measure of the internal construction of dialogues; how ideas are shared and how the participants interact socially during the dialogues[32].

There are clearly other distinctions which could be made. For instance, Rutter (1987) makes a single major division between outcome, which includes both perception and performance, and process, which he subdivides into content and style, and Monk et al. (1994) include items from questionnaires which could very easily be construed as perceptual among their measures of process.

1.3.2.1.0 Task and mediation

Before entering the description of the features of dialogues which change under mediation, it is important to step aside momentarily and comment on the well recognised interaction between task and media.

The first effect of task was first noticed by Short (1974), who found that subjects arguing for something they believed in against subjects who did not were more successful face-to-face, whereas this situation was reversed in audio-only discussions. This finding, that there is little or no effect of media when the tasks are co-operative, but more noticeable effects when then there is negotiation and conflict and the participants have vested interests in the tasks, has been amply supported in reviews (Williams, 1977; Rutter, 1987; Clark and Brennan, 1991).

For instance, Chapanis (1977) may have found that subjects performing co-operative tasks took significantly longer, more turns and various differences in the use of language, between synchronous written and spoken media of communication, but finds no more than hints of difference within the two written and two spoken media he investigated. This finding is upheld to this day by such studies as Gale's (1990), whose subjects had to co-operate in the sharing of information to design slides or build schedules, and whose judges found it impossible to mark any difference in the outcome of the tasks in two of those tasks.

---

32 In all cases, though particularly in the interpretation of turn-taking (Section 1.3.2.1.3.2), care must be taken in the interpretation of the effects of mediation. Although early researchers intimated that the effects of face-to-face communication were indicative of better communication, such value judgements may be unwarranted. It cannot be said that mediated communication is worse just because it differes from face-to-face. Halliday (1989) would argue that such differences are more or less irrelevant, and in some cases it might be argued that the effects of mediation are beneficial.

On the other hand, when the tasks involve conflict, increases in the degree of mediation lead to several effects. Participants prefer face-to-face meetings to telephone for some management tasks and telephone for others (Furnham, 1982), there are changes in performance with increased mediation leading to decisions or outcomes which are more based on the objective value of sides cases than on the sides' interests (Short, 1974; Stephenson, Ayling and Rutter, 1976) and there are several effects on process such as less overlapping speech, longer utterances and more speech errors when conversations are mediated (Stephenson, Ayling and Rutter; 1976).

It is at the same time worth noting the difficulty there can be in defining whether a task is co-operative or negotiative. For instance Gale's study mentioned above involves three tasks; *information dissemmination,* in which a team share the same information and must produce a slide to represent that information, a *creative cooperative work* task, which is similar to the first except that each member has different information from their partners, and a *meeting scheduling* task, in which the participants must fit a meeting into the hours a particular room is free and into their own schedules in which only certain meetings may be moved. Gale refers to all of these tasks as requiring negotiation, but there is little conflict involved because the participants have little vested interest in their problems. Because there is only a shared goal, and no vested interest in individuals' requirements, Gales' task is essentially co-operative.

Similarly, Veinott, Olson, Olson and Fu (1999) find effects of video-mediation on the "negotiation of common ground" during the otherwise collaborative "Map Task", however this involves no obvious conflict and other authors would consider them to be effects on the process of communication. Their finding that non-native speaker pairs' performance benefited from video more than native speaker pairs might suggest that there is more conflict as their ideolects collide, but would actually seem to be because when the had video they used gestures in the air to draw map features they could not describe verbally. Such activity turns the video mediation into a means of sharing physical objects rather than for supporting personal awareness, and this has been shown to be of benefit by several authors such as Olson, Olson and Meader (1997) who found benefits of whiteboards in collaborative design tasks and Gaver, Sellen, Heath and Luff (1993) who showed that video cameras which permitted views of workers in relation to the furniture in their office and focused on objects of discussion could be beneficial in both collaborative and competitive visually based tasks.

More recently, a second distinction in task has been drawn between the equivocality and the uncertainty of tasks and the appropriateness of different media for dealing with them (Daft and Lengel, 1986). Uncertain tasks are ones where clearly defined problems exist, but the solutions have to be found; the measure of uncertainty is the number of possible solutions. Equivocal problems are ones where it is not even certain what the issues are and people must work with ambiguity and an absence of information. Their research covered the spoken media of face-to-face and telephone conversations as well as written media which they divide into personal letters or memos, impersonal written documents and numeric documents, and has found that face-to-face meetings were preferred when the subject of discussion was equivocal.

1.3.2.1.1 Perception

Measures of perception may cover several features. The most obvious is participants' judgement of the media and their suitability for particular tasks, but the perception they build of their partners and the tasks may be equally as important. One problem when reviewing this is that the latter measures have often been considered, quite rightly, as measures of outcome. For instance, in conducting interviews over the phone, it is assessing the impression the interviewee gives of themselves which is the goal, and therefore also a measure of outcome, yet it is also a perception. This section will include measures which capture the effect of media on personal perceptions of any part of the dialogue or its participants.

Studies which have directly looked at preferences for media to resolve particular tasks have included work from the Communications Studies Group (e.g. Williams, 1975, reported in Rutter, 1987), Daft and Lengel's (1986) Furnham's (1982) and Gale's (1990).

The Communications Group (Williams, 1975, reported in Rutter, 1987) found that during discussions among civil servants, although not all comparisons reached statistical significance, the order of preference for the conversations which took place was face-to-face, video-mediated and then telephone. As has already been mentioned Daft and Lengel (1986), in studies of activities taking place in organisations, found that face-to-face discussions were preferred where there is a high degree of "equivocality" in the subject under discussion, that is where there is ambiguity about the nature of the problem to be discussed, not just "uncertainty" as to what the answer is. In other studies of work practice, Furnham (1982) found that face-to-face conversations were preferred for reprimanding, whereas telephone conversations were preferred for refusing unreasonable requests. Gale (1990), in a study of the use of whiteboards used on their own or supported by audio-only or video communication, with working subjects performing realistic business tasks over long periods repeated the common finding that audio or audio and video was perceived to improve discussion about what was being done on the whiteboard. Video was also perceived to have slightly improved the co-ordination of the discussions, in assessments of co-workers' participation.

People are also able to perceive effects which are otherwise manifested as differences in process. Sellen (1992) found that her subjects, who were involved in multi-party discussions conducted either face-to-face or over one of two video-mediated systems and in which they had no vested interest, considered face-to-face discussions as "more interactive", felt that they were more able to take control of the conversation, felt better able to attend to other speakers individually and felt more aware of when people were paying attention to them. Olson, Olson and Meader (1997) found similarly in their study of the interaction of whiteboards with face-to-face, audio-only and video-mediated co-operative tasks. Their subjects found it harder to tell what other group members thought about what was said when using the audio-only link. They also rated the work with audio-only less satisfactory than with either face-to-face or video-mediation.

Watson and Sasse (1996) directly assessed the impact of several technical failings in modern low-cost video-conferencing systems on communication. Packet loss inherent in internet video-conferencing was found to have an effect on listeners' comprehension, but certain reconstruction algorithms could ameliorate this. Synchronisation between audio and video was only found to be important if the video was updating at above 5 frames per second. Good audio quality was found to be particularly important to students in video-supported foreign language learning. These findings provide perceptual foils to the effects of delay and packet loss on audio which has been discovered in the performance and process of mediated communication (see Sections 1.3.2.1.2, 1.3.2.1.3.1 and 1.3.2.1.3.2).

Finally, those studies which have looked at participants' perceptions of their co-participants will be covered. In the same study from the Communications Study Group mentioned above (Williams, 1975, reported in Rutter, 1987) subjects' preferences for their partners in discussions were also measured and the same order of preference (Face-to-face, CCTV, telephone) was found, though again it has to be noted that not all the comparisons they took were statistically significant. In other studies from the Communications Study Group (Reid, 1970; Young, 1974; both reported in Rutter, 1987) which involved the perception of people centrally in interviews for a travel scholarship, no differences were found between face-to-face, CCTV and audio-only ratings of the interviewees.

1.3.2.1.2 Performance

One of the most simple measures of outcome is the time taken to perform tasks, and it is widely accepted that telephone calls are shorter than comparable face-to-face conversation (Short, Williams and Christie, 1976) although this may be due to several features other than the direct loss of visual information such as thoughts to the expense of telephone calls and the discomfort of using telephones. The finding has been somewhat supported in experimental situations, such as Chapanis' collection of work (e.g. 1975) who found that audio only communication took about 10% longer to complete co-operative information sharing tasks than face-to-face though the differences were not statistically significant, and work at the Communications Studies Group (Davies,1971a, 1971b; reported in Williams, 1977 and Rutter, 1987) where it was also found that when solving problems with objective solutions face-to-face dialogues were slightly longer than audio-only ones. More recently, Gale (1990) did not find differences in time taken to perform co-operative tasks, but his study involved the addition of audio and video to remote whiteboard use, and no face-to-face condition.

Olson, Olson and Meader (1997) report no time difference in performing their design task, but it is unclear what limits were imposed, and there is a suggestion that subjects may have been using them to their maximum extent.

Another intuitive measure of performance is the degree of success in performing a task, but often the tasks employed have made such a measure difficult or impossible to define precisely before the experiment, for instance Cook and Lalljee's (1972) study of students told to "make each other's acquaintance". Many studies have involved discussions of social dilemmas, for instance much of the work at the Communications Studies Group, and no objective criteria for success were predicted or indeed available as for instance in Rutter, Stephenson and Dewey's (1981) study of people trying to persuade others their beliefs about subjects on which they disagreed. In a slightly different vein, Chapanis (1975), used tasks with only one solution and so the time taken was the only measure of success.

Nevertheless, there have been many studies which have used fairly closed, deductive tasks. In almost all cases little effect of the medium of communication has been found, but this is largely because the majority of work has involved co-operative tasks, and what little conflict there may have been has not shown through and as was explained in Section 1.3.0.1 what differences there are are most evident where conflict is involved.

One of the earliest studies to show differences when conflict was involved was Morley and Stephenson (1969, 1970). They asked pairs of subjects to resolve union conflicts in circumstances where the case given to one house was manipulated to be stronger than that given the other. When their discussions took place face-to-face there tended to be compromise (0/10 wins by stronger side), but when the discussion was audio-only, the side with the stronger case tended to win (7/10 wins by stronger side). Their conclusion was that audio lacked social cues, which they termed being more "formal", and that this increased formality relegated interpersonal concerns and left only the interparty ones. In the audio only discussions, issues such as presentation became less important and the importance of the merits of each case were raised. In later experiments, reported in Rutter (1984) closed-circuit television was added and found to be significantly different in outcome from the audio-only condition; bringing the outcomes closer to those reached face-to-face.

In co-operative tasks, the effect of mediation on performance is less pronounced. Davies (1971a, 1971b, reported in Williams, 1977 and Rutter, 1987) found little difference between the solutions to his deductive tasks. Gale's (1990) judges were unable to distinguish between his subjects' performance in producing slides which represented the group's recommendation after a planning session, but as has already been commented this study only involved the addition of audio and video to distant collaboration over whiteboards.

Olson, Olson and Meader (1997) comparing video mediated co-operative design tasks with audio-only and face-to-face; and also varying access to a whiteboard, found a significant difference in the quality of the designs between trials performed face-to-face or with audio-only, but no differences between either and video-mediation, even if the whiteboard was also removed from the face-to-face condition but not from the video-mediated.

Anderson et al. (1997) report on a number of studies in which no difference was found in performance between face-to-face, video-mediated or screened conditions when subjects were performing collaborative tasks such as giving directions to people with maps or organising journeys around the United States unless other features, such as delays to the audio, were added[33].

Where there have not been differences of perception or performance there have however often been differences of process, and it is these which will be covered next.

1.3.2.1.3 Process

Even where measures of performance and perception have been insensitive to the effects of media on dialogues between participants, measures of process have often captured them. Indeed, because because participants' goals may be very different during a short experiment from during longer-term use, Monk, McCarthy, Watts and Daly-Jones (1994) suggest that measures of outcome may not be as useful as measures of process in any case. They contend that in short experiments the subjects concentrate on the task at hand, whereas in longer term use, tasks such as getting to know other participants demeanour and more general abilities may be just as important as the ability to perform that primary task. Isaacs and Tang (1994) also argue that media's primary effects are going to be on the mechanics of conversations which they mediate, and this may have a longer term effect on collaboration and satisfaction.

Process captures how participants progress towards the outcomes they reach, and has generally concentrated on the use of language during this progress although the patterns of gaze shared between participants has also received attention. In this review, gaze will not be covered because the primary interest is in differences between face-to-face and audio-only communication and gaze is obviously not a consideration in the latter.

---

33 Both Gale's (1990) and Olson et al.'s (1997) research include the use of mediation technologies to augment dialogues in ways other than supporting gaze and direct interpersonal awareness. Their addition of a whiteboard appears to be more beneficial than the inclusion of a video link showing the participants, as long as a good quality (synchronous) audio link is present. This is further supported by studies which have used video for other purposes such as Xerox EuroPARC's work on Media Spaces in which video was used to set up virtual shared offices (Gaver, 1992) or work on worn video cameras used to afford distant collaborators a view of the object being worked on (Kraut, Miller and Siegel; 1997).
At EuroPARC it could be used to display physical desk tops and objects of discussion (Gaver, Sellen, Heath and Luff, 1993; Harrison, Bly, Anderson and Minneman, 1997) and was often used for monitoring the availability of co-workers for co-present conversation rather than for direct communication (Heath and Luff, 1992b).

Other than a small number which have captured surface features of dialogue, such as the use of first person pronouns or the co-referring expression "it", "that" and "they", which have received only limited support in the literature (Brennan, 1991; Monk et al., 1994), measures of process can be divided into the categories of density, turn-taking and function.

Density, loosely following Halliday (1989), is being used to refer to measures such as utterance length and word density. Turn-taking refers to the information which can be divined about participants' interpersonal awareness of their co-participants and the effects of technology on the co-ordination of dialogue. Function refers to schemes which have been developed to measure the ways in which individual utterances are used to bring dialogues closer to resolution, essentially capturing the ways in which common ground (Clark and Brennan, 1991) are achieved. A subsection will be given to each of these measures.

The review will not include studies of gaze, or mutual eye-contact, because it is not universally available in these studies; for instance in any involving audio-only communication.

1.3.2.1.3.1 Density

Density, which is being defined as measures of the amount of communication which takes place, has been used since the beginning of research on mediation.

To evaluate the process of his co-operative tasks, Chapanis (1977) used counts of messages, sentences, words, different words and the number of words per minute, but found no difference between face-to-face and audio-only, though there were on average more of each measure used in face-to-face than audio-only. In an experimental study of industrial relations debates (Rutter and Stephenson, 1977) which was manipulated to ensure the debaters were pro the brief they were given, it was found that utterances tended to be longer in an audio-only condition than face-to-face, though there was no significant difference in the total number of words (slightly more on average in audio-only than face-to-face) or utterances (slightly more on average in face-to-face than audio-only).

More recently, similar analyses have been applied to video-mediated communication. Sellen (1992), studying group discussions of debates in which there were no vested interests conducted either face-to-face or over one of two video-mediating systems, found no difference between the three conditions in the number of turns, how they were distributed among speakers, or their duration but did not report word counts.

O'Connaill, Whittaker and Wilbur (1993) also compared face-to-face with two forms of video-mediation, one of which had a notable delay, in their case for the performance of real work-related meetings. They found significant differences between both face-to-face and delayed video, and between the synchronous video and delayed video in the total number of turns taken during the meetings, the number of turns taken by each participant and word per turn, though not in the total number of words or words used by each participant; but no significant difference between the synchronous video and face-to-face. There were fewer turns produced in the delayed video than in either other condition, but more words per turn.

Boyle, Anderson and Newlands (1994) found that in sharing routes on schematic maps, subjects needed to say significantly less in face-to-face than when screened from view. Similarly Anderson et al. (1994) found that fewer words were used face-to-face than audio-only when subjects were attempting the co-operative Travel Game task in which they had to organise a route around the United States. In related studies involving video-mediation, no differences were found in turn and word counts between video-mediated and audio-only conditions except where they implemented a video-tunnel system which permitted direct eye-contact; finding this system to led to dialogues with more words than audio-only or video-mediation without eye-contact.

1.3.2.1.3.2 Turn-Taking

Another frequently used measure is that of how turn transitions take place (see Section 1.3.1.1.1). This uses analyses of how well the transition from one speaker to another is co-ordinated to determine how the subjects' intersubjectivity is affected by mediation.

One of the first studies to use this was Cook and Lalljee's (1972) study of students told to "make each others" acquaintance" either face-to-face or using only audio. They found that there were more interruptions face-to-face than audio-only, and inferred that people could not interpret how their partners were reacting to their interruptions and waited for more explicit turn-transition places.

Rutter and Stephenson (1977), who, to avoid confusion with the term "interruption" referring to attempts to take the floor, used the term "simultaneous speech", analysed simulated union negotiations in which the subjects were known to be partisan, and found more and longer occurrences of simultaneous speech in the face-to-face than in their audio-only condition. They concluded that the increase in simultaneous speech was because the additional non-verbal signals available face-to-face supported the continuity of the conversation while permitting the participants to speak spontaneously. The increased overlaps were indicative of increased comfort in the conversation; the face-to-face condition permitted listeners more opportunity to take the floor or give backchannels[34] without disturbing the conversation as a whole.

More recently, Sellen (1992), using an automated tool to measure periods of time when two or more of her speakers were talking simultaneously, found more occurrences of simultaneous speech in face-to-face than in either of two video-mediated conditions and no difference between the video-mediated. She also found that the amount of time spent in simultaneous speech was greater in face-to-face. Simultaneous speech during speaker transitions also varied, with there being more during face-to-face debates than during either video-mediated condition and again no difference between the video-mediated. Indeed, overall there was an average overlap during transitions in face-to-face debates, but an average pause between them in both video-mediated conditions.

In their studies of real work-related meetings, O'Connaill, Whittaker and Wilbur (1993) found no overall significant difference between face-to-face and two video-mediated conditions, but did find a difference in particular uses of the overlapping speech. The video-mediated condition which had delays contained less occasions of speakers overlapping while completing utterances for their partners, and fewer occurring while speakers are trying to take or hold the floor, but more occasions of simultaneous speech when people began speaking simultaneously. The synchronous video system only differed from the face-to-face in number of occasions of simultaneous speech while speakers were sparring for the floor; producing less occurrences.

Anderson et al. (1997) reported several studies which compared face-to-face with screened and video-mediated systems, some of which used "video tunnels" to afford direct eye-contact, in the performance of the co-operative task of giving directions on maps. They found less overlapping speech in the face-to-face than

---

[34] Short acknowledgements of having heard, understood and typically agreed, usually non-verbal (eg "Mhm") (Yngve, 1970).

in the screened condition, and less in the audio-only condition than in either video-mediated condition. Although no statistical analysis was reported, there appeared to be more (over twice as many) interruptions in the copresent (face-to-face and screened) conditions than in the remote mediated ones.

Both Sellen's and O'Connaill, Whittaker and Wilbur's studies broadly support Rutter and Stephenson's conclusion, as do Anderson's though the relationship between audio-only and video-mediated conditions is slightly different from that predicted by Rutter's cuelessness model (see Section 1.3.2.2.1).

1.3.2.1.3.3 Function

The process measure of function can include a number of measures (see for instance Monk, McCarthy, Watts and Daly-Jones; 1994), but in this section the concentration will be on coding schemes which have been used to measure the way in which information is transferred during conversations.

Davies (1971a, 1971b, reported in Williams, 1977 and Rutter, 1987) at the Communications Studies Group performed one of the first studies of the effects of mediated communication on function when he looked at the number of potential solutions mentioned. He found that in the completion of his co-operative tasks, more potential solutions were mentioned when the discussion took place face-to-face than over an audio-only link, but did not look any closer at the way in which topics were referred, shared or closed.

Though Cook and Lalljee (1972) made some ad hoc study of the functional process of their dialogues, noting that more questions were asked audio-only than face-to-face, Stephenson, Ayling and Rutter (1976) can claim the first study which looked in depth at the way in which content was discussed, using a coding scheme devised by Morley and Stephenson (1977) to code one of Morley and Stephenson's (1970) studies of union disputes where the union side was considered to have a stronger case than the management.

Morley and Stephenson's (1977) Conference Process Analysis scheme recorded three features of statements within each utterance; it's mode, resource and referent. The mode encoded what sort of exchange was taking place, whether for instance it was an offer of information or a request for information. The resource encoded what was being exchanged, and had nine classifications including such activities as structuring the discussion, settling an open topic or making comments on an open topic. The referent noted whether anyone was referred to during the statement, and if they were a party directly related to the discussion.

Overall, the face-to-face dialogues tended to be less task-oriented and less personalized than the audio-only, with more concrete information being given and more reference being made to the parties being discussed in audio-only conversations. It was also noted that while the union subjects, who had the stronger case, were more task-oriented over an audio-only link, the management subjects tended to be more task-oriented face-to-face. Furthermore, in the audio-only conversations there tended to be a stronger structure to the dialogues, with one person taking the lead and the other responding rather than an equal division of initiation and response.

More recently, comparisons of functional codings of mediated conversations have been carried out by people such as the Olsons (Olson, Olson and Meader; 1997) and Anderson (Anderson, O'Malley, Doherty-Sneddon, Langton, Newlands, Mullin, Fleming and Van der Velden, 1997).

Olson, Olson and Meader (1997) use a coding scheme with 22 categories, which capture details about each utterance, such as whether they were primarily related to the organisation of the meetings themselves, mention of the mediating technology, digressions and several types of on task categories. They apply this to problem solving discussions between groups using a whiteboard either face-to-face, audio-only or with a video link. They found that both audio-only and video-mediated groups spent more time managing their meetings and clarifying what they meant to each other than face-to-face groups. A rare finding was a difference between audio and video, in that video-mediated discussions were found to spend more time stating and clarifying task related issues.

Anderson et al. (1997) report on several studies which have used Conversational Games Analysis (Kowtko et al., 1991; Carletta et al., 1996) to make comparisons between co-operative tasks performed under several mediated conditions, including subjects screened from view, the innovative video-tunnel, and various forms of video-conferencing.

Conversational Games Analysis makes no explicit distinction between the tasks which may take place during discussions, but concentrates on the way information is passed; coding utterances as being commands, statements, questions about new information or confirmations of information which has already been raised.

In a comparison of face-to-face and screened subjects trying to convey directions on maps (Boyle, Anderson and Newlands, 1994), it was found that there was less confirmation that common ground had been established, with fewer occasions of speakers confirming they had been understood and checking they had understood their partners. In a later study, which compared video-mediation without direct eye-contact with video-tunnels and an audio-only condition, it was found that the speakers questioned whether their partner had understood them more in the audio-only condition than in either video-mediated one. Again the tendency for an improvement in co-ordination from audio-only to video-mediated to face-to-face is seen.

Coding schemes will receive more attention in Section 4.1, where a coding scheme is developed for the evaluation of tripartite communication.

1.3.2.1.4 Summary of the effects of mediation

In summary of the mediation on human-human dialogues has been shown to produce different effects on each of the three classes of measure; perception, performance and process. People perceive the different media to each have value for particular tasks, with face-to-face and video being preferred over audio-only for co-operative tasks, but there being more interaction between precise task and preferred medium when confrontation is involved. In defining the type of task for this purpose, equivocality or uncertainty of a task may be as important as whether it involves conflict or co-operation. There does, however, appear to be a general belief that the co-ordination of dialogues is made easier in face-to-face than in video-mediated and in video-mediated than in audio-only.

As for the effect of media on performance, this seems more clear-cut. Although there may be a general tendency for comparable audio-only dialogues to be shorter than face-to-face, there is a clear interaction between the type of tasks being performed and the effect of media on their outcome. When the tasks being performed are co-operative, there is little effect of media, but when the tasks involve conflict, with sides being personally involved in the results of a negotiation, audio-only tends to raise the importance of objective features of the negotiation whereas face-to-face personal interests can over-ride them.

Measures of process have been capable of detecting differences between media, even when tasks are co-operative and no effects on performance have been detected. They may also be more useful than measures of performance because they capture inherent properties of the media, not effects on short-term tasks, which may be supported by the benefits of video-mediation typically having been shown in longer term studies (Gale, 1990; Tang and Isaacs, 1993).

There is a generally supported finding that more words are used face-to-face than audio-only, but the situation with other media is less clear. Screen studies would be expected to lead to fewer words than face-to-face, but Boyle et al. (1994) conflict with this. High-quality video does seem to lead to more words than audio-only, and fewer than face-to-face. Turn counts show a similar pattern, with more being used in face-to-face than in audio-only, but in this case the placement of video-mediation is a little clearer and in most cases it leads to more turns than audio-only but fewer than face-to-face. Where there are differences in turn counts, there tend to be more words per turn when there are fewer turns.

In studies of the structuring of turns, there tend to be more and longer periods of simultaneous speech face-to-face than in mediated conditions, though there are counter-examples to specifying the relationship between different sorts of mediation.

When the functional progression of dialogues is investigated, again there appear to be fairly clear findings even when the tasks are co-operative. When the tasks involve conflict, audio-only tends to be more task-oriented than face-to-face. When they are co-operative, there are hints that information is given more freely face-to-face, with more topics being raised and fewer questions being necessary, but some of this increase in questions under mediation may be due to poorer co-ordination because some studies have shown differences specifically in questions concerning the understanding of information which has already been presented. These studies tend to support the existence of a continuum in the improvement of co-ordination from audio-only through video-mediation to face-to-face.

Finally, it is worth pointing out that some caution has to be taken in the consideration of mediation when more advanced technology has been involved, because other factors have often come into play. The introduction of whiteboards in addition to any other means of communication including face-to-face tends to lead to an improvement in performance, though not necessarily in satisfaction (Olson, Olson and Meader, 1997). Technologies which have necessitated the inclusion of delays during transmission, or voice-switching so that only one person may speak at once have also had clear-cut effects on the process of dialogues (Sellen, 1992; O'Connaill, Whittaker and Wilbur, 1993)

1.3.2.2 Models of the Effects of Mediation

In this subsection of the section on the effects various technologies have on the dialogues they mediate, the ways in which the findings in the last section have been fitted into models. Although there were earlier such models, the first one mentioned will be Rutter's *cuelessness* model, which has fairly comprehensively captured the effects which had been discovered when he was writing, and most since. Following that, Clark's comments on mediated grounding, which are essentially further specification of Rutter's model, will be relayed. Finally, some mention will be made of Daft and Lengel's Media Richness model, which builds a similar continuum to Rutter's, though arriving at it through a different route.

1.3.2.2.1 The Cuelessness Model

It is very tempting to say that all the work on mediated communication in the last 12 years has been footnotes to Rutter (1987). With his cuelessness model, which grew out of Argyle and Dean's Intimacy model (Argyle and Dean, 1965), Morley and Stephenson's Formality model (Morley and Stephenson, 1969, 1970) and Short's Social Presence model (Short, Williams and Christie, 1976), he specified a continuum in which degrees of separation of speakers interacted with the interpersonal and interparty relationships of the speakers.

Since that time, work can be seen as either developing evaluative tools similar to the ones he had made use of, investigating the use of media in different situation (such as field studies) or testing the minor effects of varying task, combination of speakers with different abilities or predispositions, or further specification of varieties of media of communication on the position of the media within Rutter's model. As Anderson (1994) notes, the wide range of studies has lead to difficulty in interpreting the findings and that it is hard to distinguish the factors which contribute to the differences which have been found.

Rutter's (Rutter, 1984; Rutter, Stephenson and Dewey, 1981) cuelessness model was developed through the meta-analysis of several series of work on mediated communication which took place in the late 1960s and early 1970s and which are reviewed in Rutter (1987) and Williams (1977). What it presented was a unified continuum along which various media would fit, and according to which the effects on the dialogues between participants could be predicted. The media could affect the ways in which the participants interacted in many ways, but essentially what they did was change the amount of social cues which could be shared between the participants. The smaller the aggregate – of cues from whatever source, the more task-oriented and depersonalised the content, the less spontaneous the style, and in negotiations, the more likely the side with the stronger case to win a favourable outcome. The lack of cues, *cuelessness*, did not act directly on these features however, but affected an intermediate state they referred to as "psychological distance".

> *"Cuelessness leads to psychological distance, psychological distance, leads to task-oriented and depersonalised content, and task-oriented, depersonalised content leads in turn to a deliberate, unspontaneous style and particular types of outcome."*
>
> *Rutter, 1987, p. 74.*

The nature of cues could be somewhat difficult to define however because, from their own experiments, it was discovered that mere copresence, even when no visual co-reference was possible because the participants were separated by curtains – seemed to act as an important cue. Gaze or mutual eye-contact, which was originally expected to be one of the major benefits of the addition of video-mediation has certainly not been found to be as beneficial as expected (Rutter, 1987), even when fully supported by video-tunnels (Anderson et al., 1997).

1.3.2.2.2 Mediated Grounding

Clark and Brennan (1991) build a model based on the costs of grounding when certain facilities of communication are present or removed, which goes some way further to defining the cues and how they are affected. Like many researchers, his model extends into written and asynchronous media, which affords such facilities as reviewability and revisability, but the majority of the facilities mentioned in Clark and Brennan (1991) are related to spoken conversation. In that paper, he does not specify an order of media based on their cost of grounding, but such could be inferred from the benefits gained by the inclusion of particular facilities for grounding.

The facilities mentioned, which are not claimed to be inclusive, are copresence, visibility, audibility, cotemporality, simultaneity, sequentiality, reviewability and revisability (see Figure 1/7).

Copresence	Speakers share the same physical environment. They can see and hear what each other are looking at.

Visibility	Speakers are visible to each other, but cannot necessarily see what each other are doing.
Audibility	Speakers can communicate by speaking. Timing and intonation can be interpreted by hearers.
Cotemporality	The listeners receive at roughly the same time the speaker produces. There is no delay such as in long distance telephony, or longer delay such as in letter-writing.
Simultaneity	Speakers can send and receive at the same time.
Sequentiality	Speakers' turns cannot get out of sequence. In speech, turn-taking protocols control sequentiality. In written communication or when exchanging answering machine messages turns may become disordered.
Reviewability	Messages may be reviewed after having ben received. An answering machine message can be listened to several times, a letter may be reread. Speech fades quickly.
Revisability	The sender of a message can edit it before sending.

Figure 1/7 Facilities of communication.
Reproduced from Clark and Brennan (1991)

They achieve this by affecting the cost of grounding in several ways. Each medium will have only certain of the facilities, and this will affect the cost of formulating, producing, receiving and understanding messages, as well as affecting costs associated with delaying , asynchrony, changing speaker, gestural display, faults in production and repairs.

At the same time, although lack of certain of the facilities may make using media without them costly, these costs can be outweighed by the presence gained by the presence of others, which may make certain certain media which would be considered poor on the cuelessness scale preferable for certain purposes. Face-to-face conversation which has copresence, visibility, audibility, cotemporality, simultaneity, and sequentiality may be better for sharing more personal information, for instance reprimanding, but a media which includes reviewability may be better for building schedules.

1.3.2.2.3 Media Richness

Coming from the field of management studies, Daft and Lengel (1986) have built a model of media based on media richness. The principle of this model is that certain information is more rich than others, and the ability of certain media to transfer this richness varies. This richness is defined as "the ability of information to change understanding within a time interval" (Daft and Lengel, 1986, p.560). Information which can overcome different frames of reference or clarify issues quickly are rich. Information which require a long time to achieve the same ends are lower in richness.

They are able to produce an ordering of media (see Figure 1/8) based on their ability to transfer richness, which subdivides written media more than Rutter or Clark and spoken media less, but which nevertheless produces a similar order.

High Richness	Face-to-face
	Telephone
|	Personal Documents (letters)
	Impersonal Written Documents
Low Richness	Numeric Documents

Figure 1/8 The Media Richness of Common Media of Communication.

The reasons for the ordering are similar to Rutter's as well. Richness differences in media depend on such criteria as the media's capability for allowing immediate feedback, the number of channels utilized, personalisation and language variety. Face-to-face permits immediate feedback, body language and natural language to be used. Less rich media will restrict the ability for immediate feedback, and the reading of body language and may have rules which interfere with natural language production.

Essentially both cuelessness and media richness define the amount of social cues which can be transferred through media. They differ in that, whereas Rutter sees the effect of media being caused by an interaction between the values speakers put in their points-of-view and the media, Daft and Lengel see it as an interaction between the equivocality[35] of tasks and the media, thus keeping the reason for differences external to the participants. The differences which are seen between media may be due to the inability to express richer interpersonal information through certain media, but the requirement to express that information is due to the equivocality of the tasks.

1.3.3 Dialogue in Formal Evaluation

As has been seen in the preceding sections, evaluation of the dialogues between participants in controlled experiments has provided a mainstay of the investigation of mediated technologies' effects on communication.

It has been used to discover the effects of the mediating technologies on the process of dialogues, and the measures taken can be classified as density, turn-taking or functional.

Density is a relatively simple measure, reminiscent of measures of performance such as the time taken to perform tasks, because in its simplest form it measures the amount of words taken to perform tasks, but permits more inference about the way in which technologies influence participants means of maintaining coherence. More words, a higher density, tends to imply that participants are having to be more explicit with each other and may indicate difficulties in maintaining co-reference.

The investigation of turn-taking is based on the sociological work in conversation analysis, but has been formalised by delimiting categories of turn transition and comparing technologies on the basis of the number of instances of particular categories co-occurring with each technology. This indicates technologies' effects on the co-ordination of dialogue, and hence participants' intersubjectivity. In tripartite communication, differences in turn-taking as an effect of the system would suggest the system is interfering with the attention the humans pay to each other in order to structure their dialogues.

The investigation of functional progress in dialogues has predominantly involved the design and application of coding schemes intended to classify the

---

[35] Equivocality was defined in Section 1.3.2.1.0; it is essentially a measure of ambiguity. High equivocality implies the existence of multiple and confusing interpretations of a situation.

purpose of utterances within dialogues. The aim is to thereby model how individual technologies affect the transfer of information.

All three methods show potential for use in the study of tripartite communication.

If the style of an interface affects the density of dialogues between the two human participants, this would suggest the participants are having trouble co-ordinating their goals and references. In tripartite communication, this might be the case because the system's plan of a task differs from the humans'.

Effects on turn-taking would suggest that interfaces are affecting the participants' ability to share their time between interacting with the system and their human collaborators.

Effects on the functional progress could have various implications dependent both on the categories of utterance the coding scheme identifies, and the patterns of categories observed. They often serve to specify effects on density, but may also prove useful where no effects on density have been found. In tripartite communication, the design of an interface might influence the degree to which dialogues between the two human participants are lead by the computer user or by the third party. This might, for instance, be captured by an increase in the number of questions by the agent if they are being influenced to lead, or an increase in the amount of spontaneous description by the third party if they are permitted more liberty.

1.4 Summary and Conclusions

The review of what little work there has been on tripartite communication involving a computer (Section 1.1) shows that it has been approached from broadly two perspectives.

The majority of work has been based on sociological methodologies, such as participant observation or analysis and conversation analysis. This has concentrated on analysis of the dialogues between the two human participants and has suggested that there is an interaction amongst the three participants in tripartite communication, so it should be considered as more than a set of dyadic communications. The approach has recently lead to these tripartite communications being labelled triadic, to indicate that each participant and every dyadic relationship can interact with every other.

An alternative stream of work on tripartite communication has used formal methods drawn from HCI, but this has paid little attention to the triadic nature of these forms of communication and has found little more than that variation in the interface the agent uses can change the task the participants are involved in.

The first stream benefited from consideration of the human-human dialogues, but lacked formality; the second stream brought this formality but suffered for lack of the human-human dialogue. The other sections of the review sought means to rectify these gaps.

Section 1.2 provided a review of current methods in HCI, showing that although the use of formal methods and the standard experimental tools in HCI had little to offer the controlled study of tripartite communication, techniques for studying the use of computers in context could provide a means to ensure that the studies within the thesis had practical as well as model building applications.

Section 1.3 reviewed what is known about human-human dialogue, and showed how this has been applied in researching of the effects of mediating technology on dialogue. The ways in which dialogue has been used as an evaluative tool in these studies were considered to show more promise for the experimental investigation of tripartite communication.

Measures of language density in the human-human dialogues could point to various effects, most associated with poor co-reference between the conversants such as might occur when the system and third party human expected different patterns of interaction. Turn-taking could indicate when systems interfered with the participants ability to monitor each other's attention and structure their utterances accordingly. Functional analysis could indicate differences in the ways the human participants shared information, with one possible distinction being between voluntary donation by the third party and explicit request from the system's primary user.

In the following chapters the viability of these techniques in the controlled study of tripartite communication will be tested.

Chapter 2

—

Experiment 1

—

A Comparison of Open and Scripted Interface Styles in Simulated Customer Service Interactions with Audio-Only Communication

2.0 Introduction

The aim in this first experiment is to build the fundamental framework for the rest of the thesis since the interrelationships between the three parties in tripartite communication between two humans and a computer has not received experimental attention before. This will involve the evaluation of a task to simulate tripartite communication and methods to investigate effects of variation in the design of the computer's system on the consultation as well as testing of a broad variation in the interface to provide basic evidence that there are inter-relationships.

Although there is technically no reason to have the task simulate a naturally occurring tripartite communication, doing so will make the results all the more compelling. The "Travel Game" (Anderson et al., 1994) was selected because it provided such a task which had the benefit of already having been proven as an experimental task which was sensitive to the effects of technology on communication (see Section 1.3.2.1.3). This task simulates a customer service interaction in which subject "customers" collaborate with a "travel agent" to arrange an itinerary for a trip around the United States. The simple goal is to fit as many airports into the itinerary as possible, but this is complicated because only the agent has access to information about the availability of flights between airports (see Sections 2.0.1 and 2.1.1 for details of the task).

Because previous controlled investigations of tripartite interactions which concentrated on single aspects of the communication have failed to recognise the interrelationships between the three parties (e.g. Gray et al., 1990, 1992, 1993, see Section 1.1.3), it was recognised that multiple measures would be needed. Following the pattern established in the study of mediated human-human communication (see Section 1.3.2), it was decided to evaluate measures of PERCEPTION, PERFORMANCE and PROCESS.

Because only broad differences in the design of the agent-computer interface were being investigated, the questions used were simple, intended only to find main effects on the participants' opinions of the consultations. Using the travel game meant that its concrete measures of PERFORMANCE, the number of airports and the time taken to complete the task, could be used. The measures of PROCESS were those which have become common in the study of mediated communication (see Section 1.3.2.1) involving measures of density, in the form of counts of turns and words, and measures of turn-taking, in the form of counts of overlapping speech. Measures of function were not employed in this study because they are time consuming to apply and were considered unnecessary in the search for the effects of broad differences in the agent-computer interface.

In the search for basic evidence that the design of agent-computer interface could impact on the agent-customer consultation, it was decided to change the interaction style used over the whole consultations. This sort of comparison has become rare in human-computer interaction[1], because it has been recognised that the selection of a suitable style depends on a number of issues; the nature of the task, the experience of the user and the amount the software will be used (see Section 2.0.2.1.6).

The remainder of the introduction will be concerned with the explanation for the selection of the Travel Game as the particiants' task (Section 2.0.1), and a review of the guidelines which have evolved for the use of various interaction styles (Section 2.0.2).

2.0.1 Selection of the experimental task

The aim of this section is to explain why the Travel Game was proposed as one which would be suitable for studying tripartite communication with computers in customer service.

If the aim of the thesis was merely to study tripartite communication, there would be little need for ecological validity, though it would make the findings

---

1 A search of the BIDS and HCIBIB on-line bibliographic services showed that there had been only one such study during the 1990s. Wiedenbeck and Davis (1997) compared the effects of direct manipulation (see Section 2.0.1.4), menu-driven (see Section 2.0.1.3) and command language (see Section 2.0.1.1) on user's perception of a word processing system, the only important findings being that the interaction style did not influence judgements of the usefulness of the system, but that subjects who had previously used direct manipulation interfaces did not like less direct styles.

more compelling[2]. Because, however, the aim is to study tripartite communication in customer service, this need becomes more pressing.

Using an ecologically valid task, reflecting the occasions in which such tripartite communication can be observed to occur, will not only point towards the value of studying the tripartite communication per se, but will also ease the transition between experimental findings and the application of those findings in future design[3].

Furthermore, the particular interactions which take place within customer service may serve to limit the potential range of interactions between the agent, customer and computer which are likely to occur. For instance, it is difficult to imagine a customer service situation in which the customer has access to the computer database but the agent does not, or in which the customer is trying to persuade the agent to sell them something; though it may not be so far-fetched for the agent not to be willing to disclose everything they have access to.

The search for the requirements of a suitable task will be reported in three sections. In the first (Section 2.0.1.1), the range of tripartite interactions which have been observed in customer service situations will be reviewed, including both those reported in previous studies and gained through the author's personal visits to British Telecom call centres. In the second (Section 2.0.1.2), a selection of tasks which have been used in previous HCI and CSCW studies will be reviewed and evaluated for their suitability for this thesis' purposes. The final selection is made in Section 2.0.1.3.

---

2 Argument over the value of laboratory experiments in HCI has raged for some time, and there are vociferous exponents of both parties. A summary of the arguments appeared during a panel named "The role of laboratory experiments in HCI; Help, hindrance, or ho-hum?" at CHI '89, and little progress has occured since then.
One reason for the disupte appears to be an equal division between HCI practitioners whose aim is to apply HCI tools in the design of software and HCI researchers whose aim is to uncover the basic processes of HCI.
Both sides would do well to recognise that in other fields – for instance ethology – though field studies are used to suggest avenues of research, it is only after laboratory verification that results are considered sound.

3 See Section 1.2.1.3 for more comment on the importance of ecological validity in HCI-related studies, including reference to Baecker et al's (1995, p.571) comment that such validity is frequently impossible.

2.0.1.1 A model of customer service tasks

Although, for ease of reference, the studies of tripartite communication which were reviewed in Section 1.1 were divided into the three categories of Computers in Emergency Dispatch Facilities (Section 1.1.1), Computers in Medical Consultations (Section 1.1.2) and Computers in Customer Service (Section 1.1.3), they all in another sense represent customer service situations. This is because in all cases, there is one paying human party who seeks a service from the company which the second human party represents.

This relationship is self-evident in the calls in the Customer Service section, but also present in the other two. In the Emergency Dispatch, the service the call-taker provides is helping the caller provide suitable details to ensure they get the emergency service they require. In Medical Consultations, doctors provide the service of determing from patients' symptoms and their medical expertise which treatment would suit their patients best. Muller et al.'s (1995) reconception of Directory Assistance operators as knowledge workers (see Section 1.1.3) brings the similarity into even clearer focus.

The key is that in all cases the "agent" mediates between a huge knowledge or service space, often with its own arcane language, and a "customer" who has little or no personal experience of the languages or spaces but requires something from them.

This means that all three sections should be considered in making a decision of which features define tripartite communication in customer service.

2.0.1.1.1 Defining features of tripartite customer service

In the studies which were reviewed in Section 1.1, three defining features of the tripartite customer service interactions were indicated.

2.0.1.1.1.1 Asymmetry

The first generalisation which can be made about these interactions is that the access to the third-party computer, and implicitly the information it holds, is asymmetric.

> *Tripartite customer service interactions involve asymmetric access to the computers*

In the studies of Toll and Assistance and Directory Assistance (Gray et al., 1990, 1992, 1993; Lawrence, Atwood and Dews, 1994; Muller et al., 1995), as well as in those of Emergency Dispatch Facilities (Whalen, 1995a), in which the agent and customer converse remotely by telephone, the computer is only accessible to the agent or call taker. In the studies of doctor's consultations (Fitter and Cruikshank, 1982; Greatbatch, Luff, Heath and Campion, 1993), although the doctors and patients converse face-to-face, the doctor has almost exclusive access to the computer – the patients are reported to be able to observe the doctor-computer interaction, but there is no indication they are aware of the computer's output. Even in Bright's (1991) study, there is no indication that patients were given access to the computers' displays.

This contrasts with other situations in which machine's are present in human-human communication, such as Suchman's (1987) studies of pairs using photocopiers. In this case, the tasks modelled an office environment in which both subjects were permitted equal access to interact with the machine.

This restricts the task which should be chosen to one in which asymmetric access to the computer is inherent.

2.0.1.1.1.2 Strangers

The second generalisation is that these interaction involve people who have not previously met each other.

> *Tripartite customer service interactions involve communication between strangers*

In the studies of Toll and Assistance and Directory Assistance (Gray et al., 1990, 1992, 1993; Lawrence, Atwood and Dews, 1994; Muller et al., 1995), as well as in those of Emergency Dispatch Facilities (Whalen, 1995a), there is no indication that the agents or call takers knew their callers. This is less likely to be true of the doctor-patient consultations observed by Fitter and Cruikshank (1982) and Greatbatch, Luff, Heath and Campion (1993); but still the contact between the participants is likely to be infrequent and contained within the doctor's surgery, not personal.

This feature is important because the majority of the work on computer-supported collaborative work has modelled typical business tasks (e.g. Olson, Olson and Meader, 1997; Gales, 1990), in which participants are likely to have closer relationships.

One of the criticisms of studies of such tasks is that the effects of technology on the building of these relationships is missed because they are one-off snapshots and a longer-term view is needed (Isaacs and Tang, 1994).

In the case of customer service, however, the tasks chosen should be ones in which the participants should not be known to each other, and Isaacs and Tang's (1994) criticism of other such tasks is not relevant.

2.0.1.1.1.3 Cooperation

The third, and final, generalisation is that these interactions involve cooperation not negotiation.

> *Tripartite customer service interactions involve cooperation not negotiation.*

In all the studies of tripartite communication reported in Section 1.1 the tasks the participants were involved in were essentially cooperative, and both were aiming to reach the same goal. In the Toll and Assistance calls (Gray et al., 1990, 1992, 1993), the customer is trying to gather information or make calls, and the agent's role is to facilitate this as quickly and pleasantly as possible. In Directory Assistance calls (Lawrence, Atwood and Dews, 1994), the customer is trying to find a telephone number, with often only partial information about the desired number's owner, and the agent's role is to gather enough information to complete the search. In Emergency Dispatch (Whalen, 1995a), the caller needs access to an emergency service, and the call taker's role is to transfer that need to the relevant service. In doctor's surgeries (Fitter and Cruikshank, 1982; Greatbatch, Luff, Heath and Campion, 1993), the patient is in need of medical assistance, and the doctor's role is to select the best remedy.

It is important to recognise this because previous studies of the effects of technology on collaboration and communication have shown that the effects depend on whether the task the participants are engaged in is cooperative or negotiative (see Section 1.3.2.1.0).

Although there are elements of sales in some such customer service situations – doctors may be paid for their expertise, customers may pay for the directory assistance and toll and assistance calls – and therefore the "agents" must satisy the customers' requirements to retain their business, there is none of the discussion of heart-felt issues studied, for instance, by Stephenson, Ayling and Rutter (1976).

This generalisation implies that the task selected for use in this study should be cooperative, as well as involving asymmetric access to the third-party computer and strangers as its participants.

2.0.1.1.2 Personal Observations of Customer Service

During discussion with British Telecom human factors experts and management, and personal observation of customer service interactions, the generalisations made in the previous section were supported.

2.0.1.1.2.1 Asymmetry

In all cases observed, the agent and customer had asymmetric access to the third-party computer, and implicitly to information. The agent and customer conversed remotely, and the agent had unilateral access both to input to and view the output from their systems.

Anecdotal evidence suggests that even when agent-customer interactions are face-to-face, companies in many cases prefer to restrict customer access to their databases. Consider, for instance, high street travel agencies, where, although the agents and customers are co-present, the customers are not inherently given access to the agent's computer screens, and interfaces use abbreviations and jargon which make them unsuitable to be read by untrained personnel.

Future situations in which this asymmetry would be reduced had been proposed, for instance giving the customer limited access to the computer's output through interfaces which could be displayed simultaneously with the consultations on home computers or over cable televisions, but there was little indication these would be implemented in the near future.

2.0.1.1.2.2 Strangers

In the majority of cases observed, the interactions were between strangers. Bigger corporate customers had staff who were dedicated to them – therefore the customers' representatives would deal with the same agent on different occasions – but this comprised a small part of the total number of customer service interactions which take place. Most customers would not expect to encounter the same agent more than once; and if they were to, would be unlikely to recognise them.

Management indicated that there some moves to improve the relationships between the company and customer, particularly trying to emulate the personable relationships which grow between customers of local stores. The aim here was to keep track of customers' usual requirements so that the service could be tailored specifically to them, in much the same way a general store's owner keeps track of events in his customers' lives so that he can suggest suitable new products to them.

The impression given, however, was that this was all to be automated – most likely because of relatively high staff turnover in telephone customer service – and there would be no attempt to give individual agents' attention to a wider range of customers.

2.0.1.1.2.3 Cooperation

Although the archetypal telephone interaction between companies and customers is cold tele-sales – usually offering fitted kitchens, conservatories or double glazing – this sort of interaction was not observed at British Telecom. Such calls as these would be negotiation tasks, because the agent is typically trying to persuade the customer to buy something they had no previous intention to buy, and which is often of doubtful benefit to them.

The nearest to such calls which were observed at BT's call centres were outgoing calls offering improved services to customers, such as charging schemes which offered discounts to customers whose bills were consistently above a certain amount. Agents were also not intended to perform a "hard sell" but merely to suggest options to customers – British Telecom was trying not to lose its image as "Auntie BT" (Rob. Davis, pers. comm.) in the face of stiff commercial competition in the recently deregulated telecoms industry.

2.0.1.2 Selection of a task

A new task could have been designed to fulfil these specification, but because the task is also required to produce consultations which will be sensitive to the effects of variation in the agent's interface within the tripartite communication, it was decided to base the task on one which had already been developed and shown to be sensitive to other experimental variations.

Many of the tasks which have been developed in psychology are manifestly not suitable to fulfil the above requirements. Although it might be academically interesting to discover the effects of technology on the collaborative performance of tasks which involve problem solving – particularly standardised tests or tasks such as the Prisoner's dilemma – the direct benefit of this to software design would be limited (see Section 2.0.1).

To select a task which would be more suitable, the tasks used in the experiments on computer-supported collaborative work and mediated dialogue which were reported in Section 1.3.2 were evaluated.

2.0.1.2.1 Ecology

The majority of the tasks fulfilled the restriction that they should have some ecological validity. Although not all represented typical customer service tasks, any that modelled real world tasks were not rejected immediately, on the basis that they could all have captured some aspect of customer service interactions.

Rejections were Davies' (1971a, 1971b) closed, deductive tasks, Cook and Lalljee's (1972) tasks in which subjects were asked to make each others' acquaintance, and Reid's (1970) and Young's (1974) interviews. Davies (1971a, 1971b) and Cook and Lalljee's (1972) are at the extremes of human social contact, the former unrealistically abstract and the latter emphasizing the social aspects of conversation too strongly. As regards Reid's (1970) and Young's (1974) interviews, although the gathering of information about customers is an integral part of many customer service calls, selection interviews are qualitatively different from these interactions.

2.0.1.2.2 Asymmetry

It is also hard to see how Cook and Lalljee's (1972) task could be made asymmetric. A computer dating agency situation in which the "agent" is invited to introduce the "customer" to a third person, represented by database file, might be conceived; but seems unlikely.

Although many other tasks in Section 1.3.2 could be levered into situations where there is asymmetric access to computers and information, these were also rejected. They include Short's (1974) civil servants' discussions of budget cuts – essentially balloon debates – Morley and Stephensons' (1969, 1970) and their collaborators' union conflicts, and Sellen's (1992) informal discussions.

2.0.1.2.3 Strangers

Of the remaining tasks, although few demand that the participants be strangers, it was considered that several were less likely to be performed by people who knew each other as little as those in the situations recounted in Section 1.1 or observed at British Telecom.

These were the slide creation and scheduling tasks used by Gales (1990), simulations of the work-related meetings evaluated by O'Connaill, Whittaker and Wilbur (1993), and the collaborative design task used by Olson, Olson and Meader (1997).

2.0.1.2.4 Cooperation

All of the tasks which would be rejected under the restriction that they be cooperative have already been rejected under the previous restrictions. Because this restriction confirms those decisions, they are Short's (1974) planning of budget cuts, Morley and Stephenson's (1969, 1970) and their collaborators' union dialogues, and Gales' (1990) scheduling tasks.

2.0.1.3 The selected experimental task

The remaining tasks are the various tasks used by Chapanis (e.g. 1977) almost all of which were cooperative, and involved experts advising novices how to perform physical tasks, Brennan's (1991) task involving the retrieval of information from remote databases, and Anderson and her collaborators' (e.g. Anderson et al., 1997) map task and travel game.

Chapanis' (1977) tasks may reflect those which occur in some customer service situations – computer support personnel frequently have to guide their customers through the use of their software – there is clearly asymmetry in that the expert knows how to perform the task but the subjects do not, and expert-subject relationships are more naturally ascribed to strangers than friends. However, customer support of this sort is rare at British Telecom's call centres[4].

---

4 BT's ISDN and on-line gaming service support personnel may be involved in interactions such as this, this was not observed.

Many customer service interactions involve database searches similar to those simulated in Brennan's (1991) study. At British Telecom, calls to directory assistance or in which information about the company's products are discussed, such interactions are taking place. Such database searches involve asymmetry because the agent is the only one to have access to the database, and as with Chapanis' (1977) tasks, are more naturally ascribed to strangers than friends. On the other hand, although British Telecom's agents frequently have to provide information for their customer, there is also usually an aspect of their interactions which involves business transations; this is not captured within simple database searching.

The latter point also contributes to the rejection of the Anderson et al.'s (1997) map task, but this also has less ecological validity than the others. Although the giving of directions remotely is common – and it involves asymmetric access to information and is frequently performed by strangers – it is not a task which is core to customer service.

The travel game, on the other hand, was specifically intended to model the sorts of interactions which take place in customer service. It is a cooperative planning task, such as might be undertaken when British Telecom's customers arrange new services, which includes aspects of business transactions in that the "customers" have to balance their preferences against a budget.

Needless to say it is inherently asymmetric in that the customer does not have access to the agent's database of flight routes, and it is assumed that the subjects will be strangers.

2.0.2 Selection of the interaction styles

2.0.2.1 Interaction styles in HCI

Shneiderman (1991, 1998) identifies five broad styles of interaction which can be used in the design of human-computer interfaces; command language, form fill-in, menu selection, direct manipulation and natural language. As well as reviewing some evidence which will predict better design within each style, he also provides some guidelines for the selection of each style. To decide the basic style of interface which would be used for the agent-computer interface in this study, reference was made to these papers. In the next few subsections, very brief descriptions will be made of each style, and benefits which are ascribed to each will be given.

2.0.2.1.1 Command Language

Command language interfaces require the user to formulate their commands before telling the computer what actions to perform, and input their command through a language which has its own vocabulary and syntax, though these are often related to natural language or Boolean logic. These often permit the creation of complex structures in single lines of commands, which would have to be built up through strings of instructions in other interaction styles. They often require some time to learn, and are typically employed where users will be making frequent use of a system and the benefits of concision outweigh the benefits of other styles. Users may also gain satisfaction from mastering the system (Shneiderman, 1991, 1998). The archetypal system will have a single line in which the users issue their commands, and the rest of the screen is used for description of states or giving feedback after commands.

2.0.2.1.2 Form Fill-in

Form fill-in systems appear as screens with fields which can take entries, accompanied by labels which give some guidance as to the data which should be entered in each field. Traditionally, fields were moved between using <TAB>, other cursor control keys and increasingly by mouse. This form of interaction requires some more keyboard skills and training than menu selection of direct manipulation, but simplifies the entry of data, by reminding the users of data which is required and permitting the user to see the context of their entries (Shneiderman, 1991, 1998).

2.0.2.1.3 Menu selection

Menu selection interactions involve the presentation of a number of options to the user, from which they must make a selection. They therefore have to remember the bare minimum about how to interact with the computer. It also decomposes a potentially complex interaction into a sequence of simpler steps, thus structuring the making of decisions. The combination of features may reduce error rates and performance times. It may also restrict the interaction more than form-fillin or command languages, and may slow knowledgeable frequent users, though this is empirically unsubstantiated (Shneiderman; 1991, 1998).

2.0.2.1.4 Direct Manipulation

Direct Manipulation interfaces are screen-oriented, visual and graphic interfaces which require the use of pointing devices to manipulate objects on screen to desired positions. For instance, in the selection of a time, where a command language or form fill-in would require the typing of that time, direct manipulation might require the movement of hands on an analogue clock face. As long as the objects being manipulated represent a good metaphor to real world practice, this makes these interfaces very easy to learn and to retain knowledge of how to use them and reduces error. Users may also find these interfaces more enjoyable to use, than others. Although development tools are improving, these require more programming effort than other interaction styles, and similarly, although the standard of equipment available is increasing, these require higher quality displays than other styles (Shneiderman, 1991, 1998).

2.0.2.1.5 Natural Language

Natural language systems, in the ultimate case, involve the user conversing with the computer through their own natural language. These would, it is argued, reduce the amount that the user would need to know about how to interact with the computer to its extreme. The quality of such systems does not currently permit their widespread use, and users are still required to know how the system will react to their commands, so they may not be the panacea some see them as. Speech input technology also needs to be improved, since typed natural language commands are likely to be more error prone than other interaction styles unless the typist is expert (Shneiderman, 1991, 1998).

2.0.2.1.6 Comparisons

The following table (Figure 2/1) reproduced from Shneiderman (1991) summarises the advantages and disadvantages of each interaction style.

Menu Selection

- Advantages: shortens training, reduces keystrokes, structures decision-making, permits use of dialog management tools, easy to support error handling
- Disadvantages: danger of many menus, may slow frequent users, requires screen space, requires rapid display rate

Form fill-in

- Advantages: simplifies data entry, requires modest training, assistance is convenient, shows context for activity, permits use of form management tools
- Disadvantages: consumes screen space, requires typing skills

Command Language

- Advantages: flexibility, supports user initiative, appeals to "power" users, potentially rapid for complex tasks, supports macro capability
- Disadvantages: requires substantial training and memorisation, difficult to retain, poor error handling

Natural Language

- Advantages: relieves burden of learning syntax
- Disadvantages: requires clarification dialog, may require more keystrokes, may not show context, unpredictable

Direct Manipulation

- Advantages: visually presents task, easy to learn, easy to retain, errors can be avoided, encourages exploration, high subjective satisfaction
- Disadvantages: may require graphics display/pointing devices, more programming effort until tools improve, may be hard to record history or write macros

Figure 2/1 Overview of interaction styles with their advantages and disadvantages.
Reproduced from Shneiderman (1991, p. 326)

Shneiderman's suggestions for the selection of interaction style, appear to be largely based on his own intuitions and experience, since what studies have compared styles may typically be criticised for producing results which are too specific to the tested users and tasks to be generally applicable.

Ogden and Boyle (1982, referred to in Shneiderman, 1998), for instance, found a preference for form fill-in over a command language, but their study involved database update and results might have been reversed for extensive repetitive file manipulation which would permit the building of macros in UNIX or for users with more experience.

The majority of studies have concentrated on specifying issues within each interaction style. For instance a considerable effort has gone into defining best practice in menu design, comparing depth versus breadth (e.g. Kiger,1984), and considering the ordering of items (e.g. Card, 1982, referred to in Shneiderman, 1998). Kiger (1984) showed the performance advantage of breadth over depth in graphical menus. Card (1982) showed performance advantages of alphabetic ordering over random ordering, and that functional ordering was closer to alphabetic than random in performance.

The degree of task and computer knowledge, is in fact the most pertinent influence on the choice of interaction style which Shneiderman recognises. In the absence of other constraints, he suggests that when users are novice, menu selection or some form of direct manipulation should be employed. This applies whether they are novices with computers, in which little keyboard skill is required, or whether they are novice in the skill, since the presence of choices or restrictions on choices reminds them what they are capable of. Form fill-in might be preferable when the users have better computer skills, greater task knowledge or both. Expert users may prefer to use command languages, since they will prefer the typically shorter response times. The degree of skill may also affect the explicitness of choices or prompting when menu selection or form fill-in is being used.

At the same time, combinations of all styles may also be suitable, and have been implemented. A legacy system in use at British Telecom uses form fill-in for data entry, but movement between screens is accomplished by placement of the cursor in a particular field and the entry of a three letter command word. Similarly, a form fill-in interface may permit free entry in some fields, such as names, but require command language like entries in others, for instance defining airports using their international three-letter abbreviation.

2.0.2.2 Interaction Styles in Customer Service

Observation in British Telecommunications' call centres showed that in customer service, the three earliest interaction styles described by Shneiderman (1991, 1998), command language, form fill-in and menu selection are all in common use where agents are acting as intermediaries between the computer and customer. The other two, direct manipulation and natural language are less common, though direct manipulation is being considered because it may increase agent motivation (Hole et al., 1998), and natural language, or at least limited speech based interfaces, are being used for direct interaction between the agent and customer (Westall, 1995).

At the same time, it was observed that customer service has bred one style of interaction, scripting, which is not present in more general HCI, and which may have direct implications for the relationship between the agent, customer and computer. Its defining feature is that the output from the computer is only indirectly intended to inform the agent. Instead, a phrase which is intended to be read, for the most part, verbatim by the agent is presented, and the agent is intended to enter the customer's response through an appropriate style of interaction from those described in Section 2.0.2.1.

2.0.2.2.1 To Script or not to script

Although scripting is accepted to have several disadvantages, such as limiting the flexibility of response to customer requests, making it easy to miss opportunities outside the scope of the script and possibly sounding dull and forced, in some cases these may be outweighed by the benefits. Essentially, the benefit of scripts is providing consistency in the way agents interact with their customers, but this implies a number of singular benefits, such as delivering a specific and accurate message, targeting specific results, providing standards of performance, and ensuring adherence to legal requirements. Another major benefit, in situations where agent turnover is high, is the short training time and the lack of technical skills required by the agents. A consequence of this is that, when agents might be called about multiple subjects, they do not have to retain as much knowledge about each since the script holds it for them, thus reducing their cognitive load (Liederman, 1990).

Scripts are less suitable when agents are liable to remain in their jobs for a long period and a higher expenditure on their training is acceptable, and when agents must deal with a number of services which are complex and in which they require sizeable familiarity (Liederman, 1990). Such situations are often referred to as knowledge work (see review of Muller et al., 1995; Section 1.1.3).

When designing scripts, several considerations must be made, beyond the logic of the progression through the script and iterative testing. These include ensuring the script meets the objective of the call, to the level that every word should have a value and every sentence should be directed towards the desired close, being conversational and uncomplicated using short words and sentences and not using jargon, being positive, encouraging dialogue, asking questions which elicit a "yes" response and getting commitment from customers (Millard, 1995).

2.0.2.3 The selected interaction style

The decision was made that, rather than comparing two different interaction styles, which would have lead to complications because of the different training times which would probably have been necessary, as well as inviting criticism for the applicability of one or other style for the task, a comparison would be made of a scripted and open style interface. In both cases, menu selection was used because this would depend less on individual agents' keyboard skills than other methods and reduce training times on the open interface so that training time on both could be similar.

A side-effect of using menu selection as the agent's interaction method in both the open and scripted interface was that the order in which information had to be entered by the agent was the same in both cases. In the open interface, agents had to move the cursor over the airport whose flight times they wanted to check, before the flight times in the second level of the tree would pop up, so it was beneficial for them to know the customer's choice of airport first[5]. In the scripted interface, the first dialog box in the script asked "Where would you like to fly to from <current airport inserted automatically>?" and only once this had been selected would a second dialog box prompt the agent with "When would you like to fly to <selected airport inserted automatically>? There are flights at ..."

This meant that in both cases the agent had to perform some adaptation of the customer's desires to the system structure, so that the effect of the scripting alone was being isolated from other interface effects. In both cases the airport which the customer wished to travel to had to be selected before the times of flights to that airport would be presented to the agent.

2.0.3 Hypotheses

The overarching hypothesis is that the broad difference between scripted and open interfaces will permeate to the dialogues between the agent and customer. Following the taxonomy of measures which evolved in the study of mediated dialogues (Section 1.3.2), the variation of the dialogues on measures in the three categories of PERCEPTION, PERFORMANCE and PROCESS will be

---

[5] The airport level was ordered alphabetically. The time level had earlier flights at its top.

investigated. PERCEPTION will indicate the way the agents and customers perceived the consultations they had. PERFORMANCE will indicate the speed and success of the consultations, using clearly defined criteria. PROCESS will measure how the agent and customer interact to achieve their goals during the consultation.

2.0.3.1 Predictions

The most basic prediction about the way the interaction between agent and customer will be affected by the agent-customer interface is that in the open interface, decisions will be much more lead by the customer's suggestions, whereas in the scripted interface, they are more likely to be lead by the scripts' questions. So in the open interface, once it is clear that a decision has been made and the information has been entered and accepted by the computer, the customer will pre-empt any queries by the agent and initiate their own turns detailing their requests. In the scripted interface, the computer will instead prompt the agent to ask the customer for their next choice before the customer can make the turn transition.

This difference will have several knock on effects on the PERCEPTION, PERFORMANCE and PROCESS which will form the bases of the hypotheses.

2.0.3.2 Hypotheses

The most directly related difference is that because the scripts force the agent to ask questions which would be pre-empted by the customer, there will be more turns and words used in the scripted interface than in the open. The difference will be most extreme in the number of words spoken by the agent, since he will be compelled to make utterances when scripted, which will be unnecessary in the open interface. This confusion around turn transitions may also lead to more overlapping speech when the scripted interface is used, because the scripts may impose themselves when the customer is trying to initiate their own turns. This is debatable however, because the customer may wait for evidence from the agent that their dialogue with the computer is finished before initiating their turn, and so the transitions will be cleaner because of this explicitness.

It is hypothesized that the impact of the scripts will also show in the PERFORMANCE measures. Two of these will be used, time taken to perform the task and the number of airports arranged. The number of airports is intended as an overall measure of the success of the consultations, but is most directly related to the customers' goals because they were aware that the customer who visited the most airports would win a prize. Time taken is more of a shared measure of success. Customers paying for calls would want to resolve them as soon as possible, but similarly the time taken in calls can imply a huge cost to the agents' employers and these will want to reduce call times. It is most likely that the difference in turns and words hypothesized will in turn effect the time taken in each interface, so it would also be hypothesized that the scripted dialogues will take longer than the open. It is unlikely that a difference would appear in the number of airports visited because in both cases the same information is available to the agent and customer through a similar interaction between the agent and customer.

Finally, it is expected that there will be an effect from the influence of the scripted and open interfaces on the PROCESS of the agent-customer dialogues to the agents' and customers' perception of the consultations. The constraints of the scripting will lead to consultations which both agent and customer find unsatisfactory because they are unable to progress in their preferred ways.

In the order these hypotheses will be investigated in the results section, these hypotheses are:

H1 The agents and customers will perceive the scripted consultations less favourably than the open.

H2 The scripted consultations will require more time to complete than the open.

H3 More turns will be used in the scripted consultations.

H4 More words will be used in the scripted consultations.

H5 The agent will speak proportionally more words when using the scripted interface than when using the open.

H6 There will be a proportional increase in the number of initiations of overlapping speech by the agent when they use the scripted interface.

2.1 Task and Apparatus

2.1.1 Task

The task which was selected to investigate this question was a variation on "The Travel Game" (Anderson et al., 1996) in which customers are asked to organise the itinerary for a trip around the United States for themselves and a friend.

Customers may only travel between towns via a single airline with whom they have bought a 3 week travel pass, they must remain three days in each state and they are surcharged for returning to an airport they have already visited. The customer knows which airports are available, but only the agent has direct access to a database containing information about flights between them.

The customer's aim is to get to as many airports as possible within the restrictions of the pass, and a prize is offered to the customer who reaches more airports than any other.

This provides a task which is similar to those which might be carried out in a customer service situation, weakening arguments about the artificiality, but has the easily measured goal of how many airports the customer visits. Simulations of customer service tasks in general have the advantage that they involve communication between strangers both in the simulation and in the real world; avoiding criticisms similar to Isaacs and Tang's (1994) that, whereas studies of video-conferencing use strangers, collaborators typically know each other in the real world.

The agents were not advised to keep the consultations as short as possible, as is common in customer service, but were aware of the time limit, and were advised to warn the customer if they were approaching that limit.

2.1.2 Apparatus

A dedicated piece of software was designed and built by the author to support the agent and research the question of what effect greatly differing styles of interface have in such a customer service situation.

CLARIS' Hypercard 2.3 was used to build a single interface which would display basic information about the customer (see Figure 2/2). This showed the agent the customer's name, the time which had elapsed and the itinerary built so far. It repeated information about the customer's current location to make these easier to find.

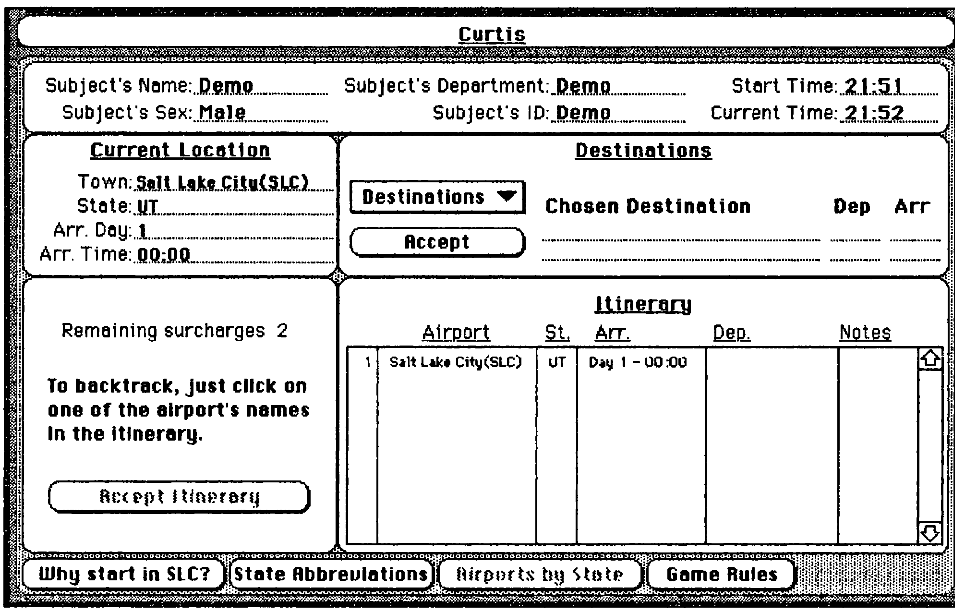


Figure 2/2 The open interface. The scripted interface did not display the buttons and fields in the "Destinations" section.

On top of this interface the variations for each style of interface, open and scripted, were added.

For the open interface a section of the interface which was left blank in the other interface contained a pop-up button, built using an XFCN called Hierarchic Popup Menu 3.3.1 (Pugh, 1995), and fields for the description of the selected destinations. Using the button, the agent could find out which airports were available from the current airport, and what flight times were available. The fields indicating which destination had been selected were included so that the agent had to confirm these choices before they were entered in the itinerary. This was to increase the similarity with the scripted interfaces in which the agent was prompted to confirm that the customer was happy with their choices.

For the scripted interface an XCMD called ShowDialog version 1.6 (Hodgdon, 1990) was used. This displayed non-modal dialogue boxes which contained phrases which could be spoken verbatim to the customer (see Figure 2/3). Where the scripts required a choice from a list of options rather than a binary decision, the list was presented in a scrolling field within the script box (see Figure 2/4).

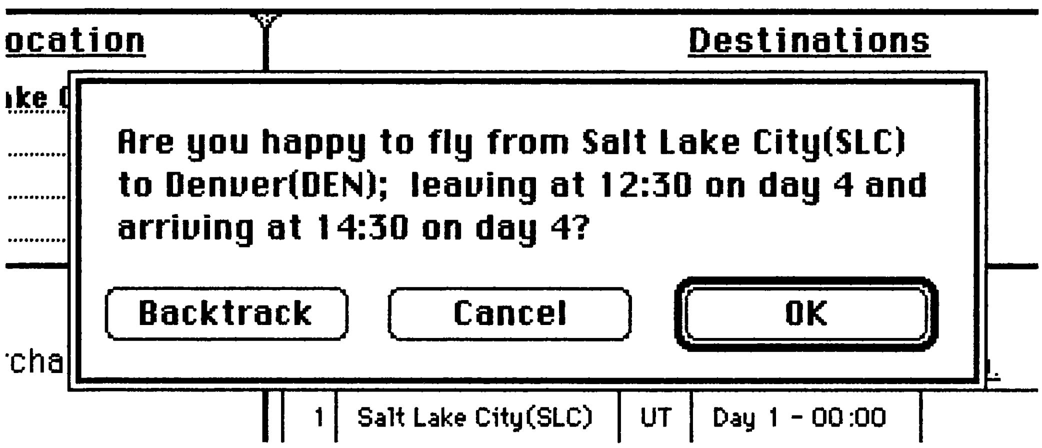


Figure 2/3 The scripted interface in use; confirmation of decisions.

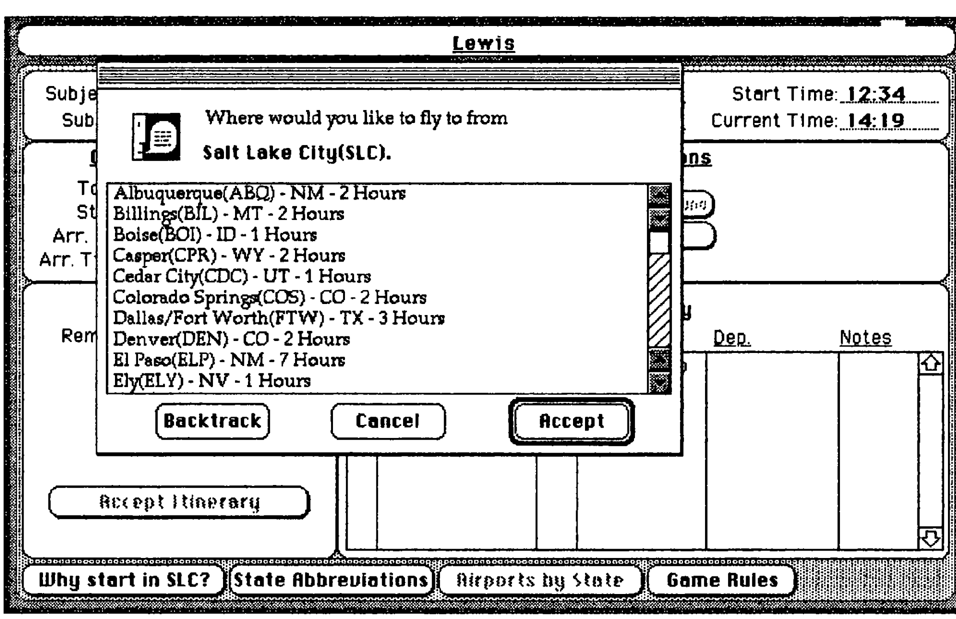


Figure 2/4 The scripted interface in use; selection of destination.

As far as possible the scripts were designed according to the rules mentioned in Section 2.0.2.2.1, for instance using short words and sentences rather than script monologues.

In both cases the agent had several off-line resources available to them. A sheet of paper contained a summary of the rules of the travel game, a list of the airline's abbreviations for State names, since the airport listings only use the abbreviations, and a list of suggested phrases to use if the customer remained silent or was having trouble selecting a prospective destination.

To aid in dialogue identification and to simulate real customer service practice more closely, dialogue boxes prompted the agent to gather subject details before the task began and to confirm their itinerary once the task was complete. OFTEL demands that all calls to British Telecom are logged with customer details, and best practice instructions require confirmation that customers' desires have been accurately recorded before they are set in motion.

Because such broad differences between the interfaces could have altered the task in such a way that more was demanded of the agent, as occurred in Gray et al.'s (1990, 1992, 1993, see Section 1.1.3) studies, it was informally ensured that there were a similar number of actions in the search for and confirmation of airports in the agent's interface.

In the open interface, the agent had to click on a popup button to open it, drag through a single hierarchic level and release on the desired flight time then double-click on a separate button to confirm the decision. In the scripted interface, the agent did not have to click to access flight information, but had to select a destination airport in one script box with a double-click, then select a flight time with a double-click in a second script box, and finally confirm the decision with a single click on a button in a third script box.

2.1.3 Subjects

Thirty six customers (19 male, 17 female) participated. All subjects were students at Glasgow University who were paid £3 for taking part, and were given the incentive that a £30 reward was available to the persons who performed best on a simple criterion.

The three confederate "travel agents" (1 male, 2 female) were all postgraduate students at Glasgow University. They were trained in both interfaces before trials with either began, and so were aware what conditions were being varied, but were kept naive to the experimental hypotheses. The training for each interface consisted of a half hour's instruction and a single trial run with one of their confederates.

Twelve subjects planned their route with each of the travel agents, and each agent used both interfaces six times.

2.1.4 Procedure

The customer read a set of instructions which introduced the game rules (see Figure 2/5) and were advised to plan a route they would like to take around the United States on the map they were given. They were also advised that the airline did not have links from every airport to every other, so they might have to adjust their plans during the consultation with the agent.

Instructions to Travel Game Participants

Instructions

Please read these instructions carefully before you say that you are ready to start the game.

You have a map of the western states of the USA in front of you, marked with the names of states and the cities which have airports used by Atlantic Airports. You should use this map to help you plan a 22 day trip around these states using an "Atlantic Air Travel Pass".

The pass allows you unlimited travel along any air routes which Atlantic Airways has with the one restriction that each airport may only be visited once. If you wish to visit an airport more than once you will incur a surcharge of $50. Your budget permits 2 such surcharges.

Furthermore, you must spend at least three days in every state you visit, for sight-seeing, but you may travel within the state during that time.

The agent will be keeping a record of your itinerary, but you may want to write a note of it yourself.

Your objective in this game is to visit as many cities as you can within the 22 day duration of the pass.

There will be a ten minute time limit on your consultation.

The person who visits the most airports wins a £30 prize.

Game Rules

- You must plan to visit as many airports on Atlantic Air routes as you can in 22 days.
- You must spend 3 days in every state you visit, but may travel between airports within the state during that time.
- You can afford 2 surcharges for visiting the same airport more than once.
- You may restart or change your itinerary at any point in the game without cost; it is your final itinerary which counts.
- There is a real time limit of 10 minutes on your consultation with the travel agent.

Remember you are in a competition!

Figure 2/5 The customers' instructions.

When the customer had made a preliminary plan, the experimenter replied to any questions they had about the rules, as long as this did not provide hints to tactics for getting to more airports and clarified the rules about surcharges and backtracking. The consultation began as soon as it had been checked that they could hear each other comfortably over the headphones they would be using during the consultation.

Before commencing the Travel Game, the agent asked for the customer's name, and basic details such as their department and matriculation number. Once the itinerary had been booked, the agent confirmed that what they had recorded in their software was what the customer had intended to plan. These two tasks modelled the structure of a customer service call, providing a small amount more ecological veracity than the isolated performance of the Travel Game, as well as identifying the customer and recording their results on tape.

Once the consultation was complete, both the agent and customer were given direct questionnaires on which to rate their consultation; containing 4 and 12 questions respectively.

2.1.5 Transcription

After the trials, the audio recordings were transcribed and marked with occurrences of overlapping speech. This process took approximately 1 1/2 hours for every 10 minute dialogue.

The transcription divided each speaker's utterances into turns, which were defined as sections of the dialogue in which they were the primary speaker, but during which the other participant could produce sections of overlapping speech.

The way in which these sections of overlapping speech was marked captured the exact nature of the overlap. The transcription indicated whether the utterances which overlapped with the original floor holder's speech ended before the floor holder's utterances or not, and only overlapping speech which extended beyond the original floor holder's utterances were considered to be new turns. Occasions in which the overlapping speech extended beyond the original speaker's utterances and new turns were added to the transcript were referred to as INTERRUPTIONs, and occasions where the overlapping speech ended before the original speaker's utterances were referred to as OVERLAPs. These distinctions reflect those suggested by Sellen (1992), who used them to detect effects of different styles of video-mediation on casual discussions.

2.1.6 Vignette

To give a feeling for how well the agent and customer could fulfil their roles, a vignette has been provided (see Example 2/1).

- Customer
  [um] can I go to Houston from there or not
- Agent
  Yes you can
- Customer
  Yeah can I fly down to Houston then
- Agent
  oh, Lots and lots of flights would you like to go in the morning or in the afternoon
- Customer
  [em] in the morning please
- Agent
  ok how early an a riser are you there's seven "a" "m", nine "a" "m", ten, ten
- Customer
  {laughs} take, I'll take the ten this time
- Agent
  the ten [uh] not that early
- Customer
  {chuckles}

Example 2/1 Vignette demonstrating participants' immersion in their roles during comparison of open and scripted interfaces. Open interface, Agent: Lorna, Subject: 6.

Previously, the customer has commented on being an early riser. The agent has remembered this to add a more informal tone to the selection of flight times.

2.2 Results

The results which were measured came from three categories; PERCEPTION, PERFORMANCE and PROCESS. The agents' and customers' PERCEPTION of the quality of the consultations were determined from their answers to the questionnaires they answered. PERFORMANCE results will cover the agents' and customers' performance of the task; the number of airports the customer included in their itinerary and the time taken for the consultation. The measures of PROCESS used here involve analyses of the dialogues which took place during the consultations to ascertain how the agent and customer achieved their goals.

2.2.1 PERCEPTION

The hypothesis regarding the measure of PERCEPTION was that the scripted interface would be perceived less favourably than the open, because the scripted interface will constrain the dialogues in ways which interfere with the natural progress of the consultation.

> **H1** The agents and customers will perceive the scripted consultations less favourably than the open.

The agents were given a short, four question, questionnaire after each trial. The 5 point Likert-type scales ran from very positive answers to very negative answers. Although each "agent" took part in 12 trials, each subject took part in only one trial so the style of interface was considered a between subjects measure; the unit of analysis is each agent-customer interaction.

Questions asked to the agents:

- Did you find the conversation natural?
- How satisfied do you think the customer was with the final outcome of the consultation?
- How friendly do you think the consultation was?
- How happy were you with the way you dealt with the customer?

Notable features of the agents' responses were that whereas 16/18 of the open interface consultations were rated "fairly friendly" or "very friendly", only 13/18 of the scripted interface consultations ones were, and that whereas the agents thought 6/18 of their customers would be "dissatisfied" or "very dissatisfied" when they were using the open interface they thought only 2/18 would be dissatisfied when they were using the scripted interface.

It was considered appropriate to use parametric analysis of these questionnaires because the ANOVA is a "*very robust statistical procedure, and the assumptions (normality and homogeneity of variance) can be violated with relatively minor effects*" (Howell, 1997, p. 321) and because Cohen (1997) says that there must be random assignation and sampling, but as long as there are equal sample sizes homogeneity of variance can safely be ignored, and that "*even with fairly small samples, the F-test for ANOVA is not very sensitive to departures from the normal distribution*" (Cohen, 1997, p. 468).

One-way ANOVAs with type of interface as a between subjects factor (2 levels: open and scripted) found that the two questions mentioned both showed a trend in the effect of the interface on the responses (Friendly: $F[1,34]$ 3,726, $\underline{P}$: 0.062; Satisfaction: $F[1,34]$ 3.778, $\underline{P}$: 0.060), but that neither of the other questions showed an effect with a probability of less than 0.1. The trends were towards the agent finding the open interface consultations more friendly, but that the customer would be more satisfied with the scripted consultations.

The customers received longer, 12 question, questionnaires, but many of the questions related to past experience, so only the six directly related to the experiment will be reported:

- How easy was it to hear and understand the travel agent?
- Did you find the consultation with the travel agent natural?
- How easy was it to make changes in your itinerary?
- How helpful did you find the travel agent?
- How satisfied were you with the final outcome of the consultation?
- How friendly did you find that the travel agent was?

In the case of the customers' responses there were clearer differences. In response to the question "How easy was it to hear and understand the agent?" when the agent was using the open interface, every customer responded that it was "Very Easy" but when the agent was using the scripted interface this number was only 11/18 with the other 7 all finding it "Fairly easy". Although there is a small difference between "Very easy" and "Fairly easy", the large number finding a difference makes it interesting. In fact the difference may be stronger than it appears because of ambiguity in the question; the agent and customer used the same means of communication in all cases, and so we would be highly surprised by any effect on the customers' hearing. The lack of difference could be masking any effect on the customers' understanding, which is more likely to be what the question captures.

The other noticeable difference was in the responses to the question "How helpful did you find the travel agent?". Of the customers whose agents used the open interface, 17/18 considered the agent "Fairly Helpful" or "Very Helpful" whereas only 13/18 considered the agents using the scripted interface as such. One-way ANOVAs with type of interface as a between subjects factor (2 levels: open and scripted) showed that there was a significant effect of the interface in both questions mentioned (Easy to Hear and Understand: $F[1,34]$ 10.818, P: 0.002; Helpful: $F[1,34]$ 5.443, P: 0.026), but that none of the other questions had a probability of less than 0.1.

The customers' responses suggest that they are able to perceive a practical effect of the interfaces on the agent but no effect on the overall performance of the task or interpersonal aspects of their conversation with the agents. Both of the questions mentioned above ("How easy was it to hear and understand the travel agent?", "How helpful did you find the travel agent?") relate directly to the agents' effectiveness, rather than the customers' performance ("How satisfied were you with the final outcome of the consultation?", "How easy was it to make changes in your itinerary?") or interpersonal aspects of their consultations ("Did you find the conversation with the travel agent natural?", "How friendly did you feel that the travel agent was?")

An important issue now is how this PERCEPTION relates to the PERFORMANCE and PROCESS of the task, since it was originally inferred that this difference in perception would be due to process differences caused by the variation of the interfaces.

2.2.2 PERFORMANCE

PERFORMANCE was assessed by two measures, the number of airports visited and the time taken to build the itinerary. The number of airports visited is most obviously a measure of customer success, because they were all made aware that the subject who visited the most airports would win a prize. It is of course potentially related to the style of the interface the agent uses, because that interface might make relevant information more or less available, but in this case no such difference would be expected and it was not expected that there would be a difference in the number of airports visited as a function of the interface the agent used. Any difference in the number of airports is likely to be subsidiary to the difference in time, for instance due to subjects feeling more time pressure if one interface lead to longer dialogues than the other.

Time taken is more a measure of shared success. A customer contacting a service agent would probably want to resolve their query as soon as possible. The company that the agent is employed by would be aware – and probably make the agents aware – that the shorter each individual consultation is, the more customers the agent can deal with. In this experiment the importance of time was conveyed to both agent and customer by putting a ten minute time limit on the consultations. This probably makes the time taken less important to the agent and more important to the customer than it is in a real customer service situation, but nevertheless reminds them both that it is an issue.

In this case it was expected that there would be an effect of interface, with the scripted interface leading to longer consultations.

H2 The scripted consultations will require more time to complete than the open.

No significant difference was found in the number of airports visited in each condition (One-way ANOVA, F[1,34] 9.027, P: 0.870), but there was a significant difference in the time taken (One-way ANOVA, F[1,34] 6.531, P: 0.015). The consultations in which the agent used the scripted interface were on average 25% longer (see Table 2/1).

Measure	Open	Scripted	ANOVA	P
Airports Planned	10.1	10.0	F[1,34] 0.027	0.870
Time Taken	400	498	F[1,34] 6.531	0.015

Table 2/1 Summary table of PERFORMANCE results by style of prompt.

This means that although the customers perceive the agents as being more helpful when they use the open interface, any difference does not impact on their performance of their goal. The impact seems more to have come from the quality of the dialogues. Considering only the data presented so far, it would appear that the quicker the agent can resolve the customer's requirements, the more helpful they may be perceived. It has been hypothesized that the length difference could only be a knock on effect from finer level differences in the PROCESS of dialogues however, and this contention will be investigated in the next section.

2.2.3 PROCESS

To determine what was accounting for the difference in time, the dialogues between agent and customer were analysed. In this way it could be seen who held the floor most of the time, and how turns were passed between the speakers. The major prediction was that the difference in agent-computer interface would show up in the PROCESS, and that the effects on the PERCEPTION and PERFORMANCE were only subsidiary. After some definition of terminology, the hypotheses related to the effect of the interfaces on PROCESS will be repeated.

The dialogues were transcribed by an audio-typist, checked by the experimenter, then marked for overlapping speech; a process which took in the region of 2 hours per 10 minute dialogue. Counts were then taken of the turns and words used by each speaker and of the use of overlapping speech.

Turns in this case were taken to be sections of speech where an individual held the floor of the dialogue. Once a turn began it did not end until the other speaker began an uninterrupted turn of their own.

Because there were just two speakers, the turns could only be passed one way, and therefore the potential difference in the number of turns used by each speaker would be limited to one. Even if a consistent difference were to be found, it would most likely be an artefact of the task, for instance caused by the agent telling the customer that the trials were about to begin; therefore differences in use of turns by each speaker were not investigated.

Typically, an increase in the number of turns would either be interpreted as an indication of fluent communication (Daly-Jones et al., 1998) though it might indicate difficulty in reaching agreement because more toing and froing is necessary to achieve common ground (Clark and Brennan, 1991). In this case, a more mechanistic explanation is likely, in that the scripted interface is expected to lead to more turns because it will prompt the agent to introduce the scripts' queries when these would be assumed by the customer if they were not presented. Further more turns may be added as the customer indicates that they were not expecting the turn from the agent.

H3 More turns will be used in the scripted consultations.

For the same reasons it would be expected that more words will be used in the scripted style than in the open.

H4 More words will be used in the scripted consultations.

More specifically, but again because it is the script prompting the agent to add queries to the dialogues which would be assumed otherwise, the majority of the extra words in the scripted style will be produced by the agent.

H5 The agent will speak proportionally more words in the scripted consultations than in the open.

The transcripts were marked for overlapping speech in the dialogues. If a section of overlapping speech extended beyond the turn it overlaps, this was considered an INTERRUPTION, which seized the floor from the original speaker, otherwise if the overlapping speech did not extend beyond the turn it OVERLAPs.

Because the scripted system prompts the agent to tell the customer information as soon as a previous decision has been entered, it is also hypothesized that there will be an increase in overlapping speech when the scripted interface is used. This is because the customer will infer that a turn transition place has been reached as the agent is being prompted to introduce the next scripted message, and consequent overlap will occur. This will result in a proportional increase in the amount of overlapping speech produced by the agent when they are using the scripted interface.

H6 There will be a proportional increase in the number of initiations of overlapping speech by the agent when they use the scripted interface.

This is not a certain finding because, there is some evidence that customers can balance their desires to take turns with the agent's interactions with their computer, and spot suitable transition places.

2.2.3.1 Turns and Words

A One-way ANOVA with style (2 levels: open and scripted) as a between subjects factor were performed on the total number of turns per dialogue. A Two-way ANOVA with style as a between subjects factor (2 levels: open and scripted) and role (2 levels: agent and customer) as a within subjects factor was carried out on the counts of words. The details of these analyses are shown in Table 2/2.

Measure	Open	Scripted	ANOVA	P
Turns	129	145	F[1,34] 1.734	0.197
Words	977	1264	F[1,34] 9.282	0.004

Table 2/2 Summary table of turn and word counts by style of prompt.

Although, as hypothesized, there are 16 (12%) more turns used in the scripted interface than in the open, there is no significant difference between the two conditions (One-way ANOVA, F[1,34] 1.734, P: 0.197). The difference does appear in the number of words used, where there are 287 (29%) more words used in the scripted than in the open interface (Two-way ANOVA, F[1,34] 9.282, P: 0.004).

As well as a significant effect of the role of the speaker (agent or customer) (F[1,34] 51.008, P<0.001) the Two-way ANOVA on the words used, also showed a significant interaction between the role of the speaker and the style of interface the agent was using (F[1,34] 11.774, P: 0.002) (see Table 2/3).

Measure	Open	Scripted
Agent Words	569	861
Customer Words	408	403

Table 2/3 Summary table of mean words by role of speaker and interaction style.

Post-hoc Bonferroni T-tests on the PERFORMANCE of the two roles in each level of style showed no significant difference between the number of words produced by the customer in each style (Two-tailed Bonferroni T-test, df: 34, t: 0.073, P: 0.942) but there was a significant difference in the number of words used by the agent in each condition (Two-tailed Bonferroni T-test, df: 34, t: -5.21, P<0.001). As can be seen in Table 2/3 the agent uses 292 (51%) more words when using the scripted interface than when using the open one, clearly supporting the hypothesis that the agent will use proportionally more words then using the scripted interface.

2.2.3.2 Words per Turn

Because the first hypothesis – that there would be a difference in turns between the dialogues in which each interface was used – was not supported, but the second hypothesis – that there would be a difference in the amount of words used – was, it was decided to determine the relationships between words and turns by performing an analysis of the words used per turn.

Measure	Open	Scripted
Agent Words per turn	9.12	12.35
Customer Words per turn	6.15	4.91

Table 2/4 Summary table of words per turn by role of speaker and interface style.

As can be seen in Table 2/4, the agent uses 35% more words per turn when using the scripted interface, whereas the customer uses 25% more in the open interface. Statistical analysis using Two-way ANOVAs with interface style as a between subjects factor (2 levels: open and scripted) and role as a within subjects factor (2 levels: agent and customer) showed that the difference between the interface styles was significant (Two-way ANOVA, $F[1,34]$ 5.145, $\underline{P}$: 0.03) and there was a significant effect of role (Two-way ANOVA, $F[1,34]$ 53.738, $P<0.001$) and a significant interaction between the style of interface and role (Two-way ANOVA, $F[1,34]$ 9.905, $\underline{P}$: 0.003).

Post-hoc Bonferroni T-tests performed on the agents' and customers' ratio of words per turn in each interface style showed that there was a significant difference between the ratio of words used by the agent (Two-tailed Bonferroni t-test, df: 34, t: -3.4, $\underline{P}$: 0.002) and a trend in the ratios of words used by the customer (Two-tailed Bonferroni t-test, df: 34, t: 1.8, $\underline{P}$: 0.088).

Although the trend is interesting because it suggests that the customer is using slightly fewer words per turn when the agent is using the scripted interface, possibly hinting that they are giving more short responses to questions asked by the agent or giving more verbose questions, it is not strong enough to be assigned much credence.

The significant difference in the number of words used per turn by the agent is more interesting because it emphasises the increase in the amount of words produced by the agent.

2.2.3.3 Overlapping Speech

A Three-way ANOVA with style as the between subjects factor (2 levels: open and scripted) and role (2 levels: agent and customer) and type of overlapping speech (2 levels: INTERRUPTION and OVERLAP) was performed on the raw counts of overlapping speech. The means of the raw counts are shown in Table 2/5.

Measure		Open	Scripted
Agent	Overlapping Speech	7.2	6.9
	INTERRUPTIONs	5.2	5.1
	OVERLAPs	2.0	1.8
Customer	Overlapping Speech	8.2	9.7
	INTERRUPTIONs	4.2	5.2
	OVERLAPs	4.0	4.4
Both Speakers	Overlapping Speech	15.3	16.6

Table 2/5 Mean raw counts of overlapping speech by agent and customer by interface style.

Although the style of interface used by the agent did not have a significant effect on the use of overlapping speech (Three-way ANOVA, F[1,34] 0.149, P: 0.702), role did have a significant effect (Three-way ANOVA, F[1,34] 4.776, P: 0.036) as did the use of the two types of overlapping speech (INTERRUPTION or OVERLAP) (Three-way ANOVA, F[1,34] 17.9, $P<0.001$). Since the only significant interaction was between the role of the speaker and the two types of overlapping speech (Three-way ANOVA, F[1,34] 14.5, P: 0.001) the investigation of the overlapping speech was not carried further.

This means there were no significant differences relevant to the investigation of the two styles of interface in customer service, because the differences in role and types of overlapping speech are most likely an artefact of the travel game task rather than the interfaces. The hypothesis that the scripted interface would lead to more overlapping speech being initiated by agents using the scripted interface is therefore not supported, indeed the descriptive statistics show a slightly lower mean number of occurrences of overlapping speech by the agent when they use the scripted interface than when they use the open one.

2.3 Discussion

The aims of this study were to evaluate the travel game as a suitable task for the experimental investigation of tripartite communication, the methods of analysis derived from research in mediated communication for capturing the inter-relationships between the three parties, and providing fundamental evidence that the design of the computer can impact on the human-human dialogue.

The subjects were observed to find the travel game immersive, and the effect of the interaction styles on the time taken to perform the task shows that it can be sensitive to the effects of changes in the relationships between the three parties in tripartite interactions.

The ability to formulate hypotheses about the effects of changes in the relationships in terms of the measures employed, and their observed sensitivity to the effects, validates their use.

Finally, experimental evidence was found for the broad differences in the agent-customer interface's impact on the agent-customer dialogue, and the analysis of the process gave some indication of what that impact might be.

It was proposed that the interface could exert influence on three aspects of the consultation – its PERCEPTION, PERFORMANCE and PROCESS – and based on predictions about the influence of the agent-computer interface on the PROCESS of the agent-customer dialogues, the following hypotheses were made:

H1 The agents and customers will perceive the scripted consultations less favourably than the open.

H2 The scripted consultations will require more time to complete than the open.

H3 More turns will be used in the scripted consultations because the interface prompts the agent to mention information which would be assumed otherwise.

H4 More words will be used in the scripted consultations because the interface prompts the agent to mention information which would be assumed otherwise.

H5 The agent will speak proportionally more words when using the scripted interface than when using the open.

H6 There will be a proportional increase in the number of initiations of overlapping speech by the agent when they use the scripted interface.

The most important hypotheses were supported, but no difference in turns (**H3**) and overlapping speech (**H6**) were found.

The first failure may be explained by a combination of the reasoning behind the hypothesis and the means of capturing the data. The hypothesis rested on the fact that there would be more turns because the scripting would force the agent to introduce queries which would otherwise be assumed by the customer. Turns were measured as the sequence of words during which one speaker held the floor and no one else spoke. If the agent did not give up the floor between acknowledging the end of the previous decision cycle, and the initiation of the query prompted by the script, there would be no extra turn generated. Instead the turn during which the agent held the floor could contain the performance of two functions. It was suggested that there would be further turns as the customer tried to align their pattern of how the dialogue would progress with the way the agent was progressing. This might not be enough to form a significant difference, but combined with a few occasions of the agent introducing scripted queries instead of ceding the floor to the customers' pre-emptive questions, might explain the non-significant 12% increase in turns which was observed.

The second failure – not finding a difference in overlapping speech – was considered when the hypothesis was presented, and may be because the customer is able to recognise the agent's dialogue with the computer, and waits for suitable turn transition places in that interaction to be signalled by the agent.

The remaining four correct hypotheses, lead by the increase in volume of agent speech during the scripted consultations, support the thesis that variation in the agent-computer interaction can permeate to the agent-customer dialogue in predictable ways.

The increase in agent words when they used the scripted interface lead to an overall increase in words, and a consequent increase in the time to perform the travel agency task.

This provides experimentally derived evidence that broad differences between the interfaces an agent in tripartite communication uses can influence the human-human dialogue. In the following chapters, the range of this influence will be investigated in closer detail.

Chapter 3

—

Experiment 2

—

A Comparison of Sales Prompting Styles in Simulated Customer Service Interactions with Audio-Only Communication

3.0 Introduction

In following up the comparison of open and scripted interfaces, it was decided to investigate the extent and manner in which potentially less invasive and pervasive variation in the design of the agent-computer interface could permeate to the agent-customer dialogue. The intention was to discover whether variations which did not alter the entire style of the interaction between the agent and computer, in other words which only occurred for short periods of the entire consultation, and which did not constrain the agent as much as scripting potentially does, could lead to detectable changes in the dialogues.

To discover a suitable aspect of the agent-computer interface to vary, agents at work in various British Telecom call centres were observed. More recently developed software at Inbound call centres, where agents receive calls from customers rather than initiating the calls themselves, provided a suitable topic.

The GUI software used by agents at this centre included floating windows which could display messages to the agent when algorithms detected the likelihood that the customer would benefit from particular products the company could provide. For instance, if the customer's bill for calls was above a certain amount the agent could be prompted to suggest a suitable discount scheme which would reduce the overall bill's cost. Although attention had been paid to correctly determining when the agent should be prompted, for instance Houghton, Gardner and Gould (1991) investigated the use of expert systems in this role, little attention had apparently been paid to the method in which these prompts were presented to the agent. Indeed during the observations, several agents were seen with the floating window hidden behind other, more important windows, apparently because they found the instructions a distraction from their primary task. Of course, in doing so, many potential improvements in the customers' service were being lost.

It was decided to investigate how variation in the manner such prompts were presented to the agent would transfer to the dialogues between them and their customer. The underlying aim was to determine if close analysis of the dialogues between agent and customer could inform decisions to be made in designing the interface between agent and computer.

To aid in the description of the fairly complex pattern of events which occur during the consultations, the following set of terms have been used to represent salient events. Some are borrowed from basic human factors and human computer interaction literature, while others are being introduced for the first time. The circumstances which lead to the prompt being presented is known as the *trigger*. When this trigger occurs, a suitable *instruction* for the agent will be selected or generated, and presented in the form of a *prompt*. If the agent considers the instruction worth transferring to the customer, they will *mention* this to the customer. The discussion of this *opportunity* is taken to have reached its *resolution* when the customer decides whether to buy the offered product or reject it.

3.0.1 Prompting

3.0.1.0 General Prompting Guidelines

Very little empirical attention has been paid to the way in which alerts should be presented to their users. Literature searches of the last five years of journals including HCI, The International Journal of Human Computer Studies, Behaviour and Information Technology and other journals listed in the BIDS and HCI Bibliography databases, as well as searches of abstracts in the proceedings of CHI and Interact conferences, unearthed not a single related reference.

Textbooks and other on-line resources pointed only to guidelines and standards documents, and these in turn generally depended on some presumption of underlying standards. For instance, these guidelines presume that the messages will be displayed through unavoidable means such as dialogue boxes, and do not consider that messages which could be presented in a less salient manner, such as in the floating fields observed in call centres. One of the few exceptions is Smith and Mosier (1986) who do mention the inclusion of constant supplementary instruction in consistent locations such as a message bar, but say no more than that such inclusions will be useful to novices without hindering the expert, and yet be available to experts encountering new situations, and provide no evidence to back up their decision.

A selection of guidelines on building user guidance from one of the most frequently quoted guides (Smith and Mosier, 1986) will be presented below, since these should be relevant to the design of any visually based prompt.

- Tailor the display for any transaction to the current information requirements of the user.
- Design display formats so that user guidance is readily distinguishable from displayed data.
- Alarms and warning messages may require output of auxiliary auditory signals as well as display highlighting, to help assure that they attract the user's attention.
- Adopt task-oriented wording for labels, prompts and user guidance messages, incorporating whatever special terms and technical jargon may be customarily employed in user tasks.
- Choose wording for user guidance that speaks directly to a user, rather than talking about users.
- Adopt active rather than passive voice in user guidance messages.
- Use concise wording for prompts; eliminate extraneous words.

As a final comment, it is also worth noting that one of the most important and recurrent features of user guidance that Smith and Mosier (1986) mention is that it should be consistent with the user expectations.

3.0.1.1 Speech Guidelines

The guidelines on giving alerts tends to presume that prompts will be presented visually, but recent work in human factors and human-computer interaction has recognised that in some circumstances sound, and more importantly for the current study, speech may be an effective mode of communication.

Human factors experts (Deatherage, 1972; Baber et al., 1992) have recommended the use of audio and speech to convey information for essentially two reasons – that it is omnidirectional and that it utilises a channel of communication which is otherwise unused – but also recognise that they are particularly suited to some types of message.

One of the first sets of human factors guidelines were devised by Deatherage (1972) for the US Joint Army-Navy-Air Force Steering Committee, and many of his recommendations reflect that audience.

The first recommendation he made were for choosing between auditory and visual methods of presentation, and these are presented in Figure 3/1.

Use auditory presentation if:	Use visual presentation if:
1. The message is simple.	1. The message is complex.
2. The message is short.	2. The message is long.
3. The message will not be referred to later.	3. The message will be referred to later.
4. The message deals with events in time.	4. The message deals with location in space.
5. The message calls for immediate action.	5. The message does not call for immediate action.
6. The visual system of the person is already overburdened.	6. The auditory system of the person is already overburdened.
7. The receiving location is too bright or dark-adaptation integrity is necessary.	7. The receiving location is too noisy.
8. The person's job requires him to move about continually.	8. The person's job allows him to remain in one position.

Figure 3/1 Deatherage's guidelines for the use of auditory and visual forms of presentation.
(Reproduced from Deatherage, 1973, p. 124)

He considered auditory signals to be particularly valuable in the following situations:

- when signals are originally acoustic -- audio signals should only be converted to visual ones when conditions preclude its use;
- warning of imminent danger -- because they omnidirectional and hearing cannot be involuntarily shut off;
- when there are too many visual displays -- for instance in piloting a plane;
- when information must be presented independently of head movement or body position;

- when darkness limits or precludes vision;
- when there are conditions of anoxia, for instance at high altitude or under high g force -- because auditory sensitivity is more resistant to anoxia than visual sensitivity;
- when signals must be distinguished from noise -- because the ear is an effective detector of periodic signals in noise;

He further suggested that, when auditory presentation was preferable, the guidelines in Figure 3/2 should be followed for the selection of tonal or voice signals.

Use tonal presentation:	Use spoken presentation:
1. For simplicity	1. For flexibility.
2. When listeners are trained to understand coded signals.	2. To identify a message source.
3. For designating a point in time which has no absolute value.	3. When listeners are without special training in coded signals.
4. When immediate action is desired.	4. There is a necessity for rapid two-way exchanges of information.
5. In conditions unfavorable for receiving speech messages. (Tonal signals can be heard at noise levels where speech is unintelligible)	5. The message deals with a future time requiring some preparation. (Example: the countdown preparatory to firing a missile - tonal signals could be miscounted)
6. When security of the message is desired; coded signals or noise signals may be used.	6. Situations of stress might cause the listener to "forget" the meaning of a code.
7. If speech communication channels are already overloaded	
8. If speech will mask other speech signals or annoy listeners for whom the message is not intended.	

Figure 3/2 Deatherage's guidelines for the use of tonal or spoken presentation.

In the sections which follow, a summary will be given of more recent human factors and human computer interaction work which has extended these guidelines, as well as giving them empirical support, and work in cognitive psychology and linguistics which has distinguished the particular advantages of spoken and written language for different types of message.

Much of the human factors research done on speech interfaces has been investigating the utility of speech in aircraft cockpits funded by the US Department of Defence. For this reason, findings have often appeared in military research reports and not academic journals; many of the reports have, however, been cited in the involved researchers' academic papers. In the following sections, when such citations occur, both the source and the referring document will be given.

3.0.1.1.1 When to use Audio or Visual signals

Speech's omnidirectionality means that users do not have to be attending a particular display point, so if it would be desirable for a number of people to receive the alert, or when users attention may be focused on other issues, speech could redirect them to more pressing developments. This would avoid alarms being missed when operators are not at their desks, which Baber et al. (1992) note as one of the major perceived reasons for missing alarms in control rooms.

At the same time, in some circumstances the omnidirectionality of sound can be a disadvantage. A recurrent argument for the failure of audio interfaces being more widespread in desktop computing at the 1998 International Conference on Auditory Display was the distraction caused to co-workers if auditory displays were used in offices. Stern (1984) and Hone et al. (1998) found that subjects preferred visual displays to spoken in automated teller machines; disliking the spoken because other customers could overhear their transactions.

As can be seen in Deatherage's guidelines, the use of an otherwise unused channel of interaction between software and its users also has several benefits.

One of the most often proposed reasons for using sound in man-machine interfaces to reduce other capacities' workloads. This is reflected in Deatherage's guide-line that audio should be used when there are too many visual displays, and is reminiscent of Paget's quote (1930, referred to in Schafer, 1995):

> *What drove man to the invention of speech was, as I imagine, not so much the need of expressing his thoughts (for that might have been done quite satisfactorily by bodily gesture) as the difficulty of "talking with his hands full."*

This has received support in the human factors literature. For instance, Simpson et al. (1987) note that, when an alarm must be given in situations when the target is involved in visually demanding tasks, such as reading flight charts in busy airspace, checking numerous visual readouts in a control room or editing text on a visual display spoken messages, auditory signals for relevant events might result in more efficient task performance. They provide no empirical evidence because they argue that the work which had been done was too application specific. Mountford and Gaver (1990) note that visual overload may even occur in desktop computing because the persistent nature of visual signals from computers means that they may become obsolete or irrelevant soon after their presentation and screens may become cluttered to the extent that important messages are missed.

Brewster (1997) tested whether audio signals could improve workload and performance measures in two tasks involving search through randomly generated text files. He compared performance with a simple text browsing system, comprising a view window and a scroll bar, with the same browsing system augmented by earcons (Blattner, Sumikawa and Greenberg, 1989; Brewster, Wright and Edwards, 1992) which indicated the direction of scrolling, the relative location in the document and when page boundaries were crossed. Earcons represent specific events and objects with abstract synthesized tones or musical sounds. Brewster considered visual overload to be present in even this simple task, because the user needed to be both looking at the browser window (the task) and the displayed position on the scrollbar (the interface).

He found that the sound enhanced interface reduced the users' mental workload on a recognised workload measurement scale. But the improvement in performance was not so clear. In a navigation task which required the subjects to find a particular page there was a significant improvement due to the earcons, but in a second task where subjects had to find particular strings of letters, there was no significant improvement, although it is possible the earcons' small benefits were lost in the large visual search times. One reason for the significant improvement in navigating was that, when earcons were present, subjects took less time to recover from errors.

It is unfortunate that a third condition did not include a comparable visual based navigation aid to avoid the possibility that any aiding system would lead to improvements over the simple browser, but it should also be considered that the scrollbar provided some feedback about position in the document. Nevertheless, the experiment provides evidence that even in such apparently simple tasks as document browsing, auditory feedback can reduce visual workload.

On the other hand, because vision and hearing are complementary senses with individual features, there may be overload in auditory signals as well. Baber et al. (1992) working in the human factors tradition, point out that there need not be a correspondingly lower audio workload just because there is a high visual load. For instance, when pilots are flying at low level, there is clearly a high visual load; but there is often a similarly high audio workload as pilots keep in verbal communication to co-ordinate with other aircraft. They also note that workload in one channel may not be completely separate from workload on other channels, because language requires a considerable proportion of a person's auditory communication resources and general processing capabilities.

Some authors have suggested that the unique properties of visual and audio signals should be exploited.

In the field of human computer interaction, Gaver (1989) notes that, while many visual signals can be presented simultaneously, a much smaller number of simultaneous audio signals are likely to be discriminable. In his discursive paper, he suggests that audio should be used for changing events, whereas visual signals should be used for persistent events, and that in desktop computing audio may be more appropriate for giving alerts about background process than superimposing invented visual metaphors. This reflects the commonly quoted example of computer users listening to their hard drives to determine when file copying is complete.

In human factors, Williams and Simpson (1976, refered to in Simpson et al., 1987) found that pilots preferred speech to be used for critical warning information, but preferred more advisory information to be given visually.

More incisively, Wickens et al. (1983) tried to formulate principles to dictate when auditory or speech should be used, and provided evidence that the mode of presentation and response and task type are all inter-related. He took the cognitive psychology finding that central processing in human working memory uses two different codes of representation (Baddeley and Hitch, 1974), which can be labelled spatial and verbal, and suggested that relating stimuli and response to the preferred underlying representation of the task would lead to improvements in performance and workload when more than one task were being performed simultaneously. He considered there to be a spatial-verbal continuum with verbal tasks being those which required language or abstract symbol systems and spatial tasks requiring orientation and tracking in a dimensional space.

In two tasks involving constant flight like primary tasks, the subjects had to perform a discrete, intermittent side task.

In the first they had to use a joystick to keep a cursor on a moving target while simultaneously deciding whether a series of letters they were presented with belonged to a set they had previously memorised. The memory task stimuli were presented either visually or verbally; and were responded to either verbally, by saying "yes" or "no", or manually, by pressing one of two buttons. The workload of the primary task was varied by having the subjects either control the velocity (easy) or acceleration (hard) of the cursor. In this experiment, a primary spatial task is performed continually while a discrete side verbal task is performed intermittently but simultaneously.

In the second test, experienced pilots used a general purpose flight simulator to fly through a serpentine tunnel while concurrently performing a task which required either spatial or visual processing. The spatial processing task involved moving an on-screen cursor to one of three target stimuli, with the stimulus command either being written messages on the pilots' head-up display (HUD) screen or spoken and the movement being made with either a finger-controlled joystick on the throttle or spoken commands indicating which way the cursor should move with clock-like directions. The verbal "communication, navigation and identification" task involved three sub-tasks such as responding to a command to send an automated identification number.

Again commands were either given through the HUD or spoken. Manual responses involving the use of a 23 button control panel and verbal responses involving naming the functions of the keys on the control panel. In both cases, the same number of steps were involved to respond to the commands, and feedback was given through the control panel's digital readout. The primary task had two levels of difficulty, based on the apparent speed of the aircraft. In this experiment, a primary spatial task is performed continually while two possible discrete side tasks, one verbal and one spatial, are performed intermittently but simultaneously.

Even though in both cases the visual stimuli were written instructions rather than iconic images, there was potential interference with the visually-based primary spatial task. The effect of stimuli being spoken or written on the perception and comprehension of messages will be returned to in Section 3.0.1.1.3. The potential presentation of alerts through symbolic visual images has not been covered.

The effects were measured using reaction time, a composite measure intended to remove the artifacts of different time required to use each response type, and errors.

Wickens showed that the primary and side tasks competed for the visual input channels, and for the manual response channels. More importantly there was an interaction between the task and the means of stimulus and response. The memory task was disrupted when the inputs were the same, i.e. when the subjects had to split attention between the target they were tracking and the visual side task stimulus. The tracking task was disrupted when the outputs were the same, i.e. the subjects had to split attention between the manual response to the side task and the tracking task. Both of these latter findings were emphasised when the harder, acceleration control, tracking task was used, but audio stimulus and verbal response were almost insensitive to the difficulty of the tracking task.

The findings from the second experiment were as clear as those from the first. Competition between visual stimuli or manual response and the primary spatial task disrupted performance of the tasks more than auditory stimuli or verbal response; with visual stimuli leading to more disruption of the side task and manual response leading to more disruption of the primary task. Furthermore, it was found that there was still less competition when the audio and verbal stimuli were used to perform the verbal "communication, navigation and identification" task than when they were used for the spatial "target acquisition" task.

Essentially Wickens showed that not only were audio stimuli and verbal responses preferable for verbal type tasks, but that if more than one task is to be performed simultaneously performance is improved if the stimulus and response modalities for the tasks are not the same.

3.0.1.1.2 When to use Speech or Other Audio signals

Most of the benefits of audio signals already discussed ignore or do not acknowledge differences between speech and non-speech audio signals. As Deatherage (1972) suggested, however, when audio signals are to be used, there are cases when speech should be preferred to non-speech signals.

One of the most important findings since Deatherage has been that speech signals are inherently alerting. Byblow and Corbett (1989) tested the response speed of subjects to spoken prompts which either began with a tone or a redundant word. They found that responses were quicker when the prompts began with the redundant word, so it seems speech both alerts and informs simultaneously.

Roberts and Sikora (1997) made another important finding when they compared simple spoken messages, earcons (Blattner et al., 1989; Brewster et al., 1992) and auditory icons (Gaver, 1986) for alerting users to non-urgent events. The difference between earcons and auditory icons is that while earcons use abstract sounds, auditory icons represent events and objects with metaphoric sounds. For instance, an earcon for the a page boundary being crossed would be a simple tone, whereas an auditory icon might be an archetypal recording of a page being turned. Roberts and Sikora (1997) found that the simple speech alerts were mapped onto events better than either earcons or auditory icons, but reported anecdotal evidence that such spoken signals may become annoying after extended use. The same anecdotal observation was made by Blattner et al. (1989) about earcons.

Arguments for several potential problems with spoken alerts have appeared. One is that spoken alerts from machines may be confused with background speech where that is present (Smith and Mosier, 1986), and Simpson et al. (1987) and Baber et al. (1992) report guidelines documents and accumulating evidence which recommend machine voices should be distinct from surrounding human speech. Baber et al. (1992) recognises another when he argues that if speech is over-used and users do indeed respond to it faster than other signals, there is some danger that if the message is not urgent the system will be perceived as "crying wolf" and subsequent spoken messages will be paid less heed.

3.0.1.1.3 When to use Spoken or Written signals

Assuming a flexible messaging system is required and language has to be used, but the above guidelines do not indicate an advantage to audio or visual messages, there may still be reasons to use spoken prompts rather than written. Practically, Simpson and Navarro (1984, refered to in Simpson et al., 1987) propose that speech signals should be used for short signals, in the region of four and eight syllables, and Rosson (1984) proposes that longer phrases should be presented visually but might be introduced by a short verbal prompt. This section will, however, concentrate on studies which have indicated that there are distinct processing systems for each modality of language, which may reflect differences in the way information is encoded and perceived. This may inform decisions about types of information which would be better represented by each modality.

In cognitive psychology, this work has involved event-related potential studies, evaluation of specific impairments in brain damaged patients and modality specific subjective ratings and recognition behaviour. In theoretical linguistics, it has been suggested that the information typically encoded in each modality affects the way each modality is used to represent information, and hence the way in which information is perceived.

Event-related potential (ERP) work by Besson et al. (1997) introduced a 600 ms pause into sentences spoken or presented through a rapid serial visual presentation (RSVP) method.

Firstly, they found that when the RSVP method was used, there was a slow negative shift during the pause, but when they were spoken there was an emitted potential during the pause. Although this could be taken to indicate that there are quite different processes occurring during reading.and listening to speech, Besson et al. (1997) recognise that the findings are more likely to be artifacts of their methods of presentation. Essentially, the RSVP method, in which each word in the sentence was presented for 200 ms with 500 ms between each word, which meant that the additional 600 ms pause could have represented only a slight lengthening of the already unnatural 500 ms delay.

They also found that when sentences were written, unfamiliar sentences lead to larger negative shifts than familiar, idiomatic, sentences, whereas when sentences were spoken there was a larger emitted potential during familiar sentences than during unfamiliar sentences. Besson et al. (1997) expected the latter, both because of empirical evidence and common sense, because the higher the expectation, the stronger the surprise effect when that is not fulfilled. They argued that the former on the other hand might be due to the extreme pauses in the RSVP presentation technique, leading to the final word in the idioms already having been predicted and processed, so there was less uncertainty and less processing required.

The qualifications of Besson et al.'s (1997) work mean that very little interpretation of the way in which the processing of read and heard language differ can be made.

On the other hand, evidence from brain-damaged patients also suggests that spoken and written language may have modality specific registers (Rapp and Caramazza, 1997). They use several evaluations of the language performance of a post-CVA patient, to argue against written language skills being dependent on spoken skills.

Their patient produced neologisms in spoken naming although he did not display any articulatory deficits, but did not produce them in written language where most errors were easily recognisable. He also showed differences in the comprehension of spoken and written sentences. When asked to match spoken or written words to pictures, he was less accurate with spoken words than with written in a number of different tests, with most errors being the selection of semantically related pictures. A low memory span meant that he had trouble comprehending sentence in either modality, but his sentence production was characterised as "oral Wernicke vs. written Broca" because his speech contained suitable syntactic elements and prosody but lacked recognisable content words, while his writing contained content words but lacked function words.

They could not determine whether the observed use of spoken language was an effect of the grammatical function or word frequency, but did rule out an effect of the size of words. On the other hand, in writing longer words were spelt less accurately than shorter, suggesting a graphemic buffer impairment, but this would not have explained the lack of function words which tend to be short. Instead, Rapp and Caramazza argue that the presence of function words in their patient's spoken language, but not in written language points to the written deficit not being at the vocabulary level but at a lower level of production. They conclude that orthographic production is considerably autonomous from spoken production.

Another source of evidence that reading and hearing of language may involve different processing mechanisms comes from word recognition studies. These have typically found that higher frequency words are recognised more frequently and more accurately than lower frequency words, but Gernsbacher (1984) showed that there have been some inconsistencies.

She suggested that one reason for these inconsistencies was the use of objective word counts of printed material to determine word frequency which ignores the relative frequency of words in spoken language, and recommended the use of subjective ratings of word frequency using subjective ratings of word frequency to resolve the inconsistencies. She also argued that, because there were high correlations between the ratings of the same stimuli words being spoken or written, the internal lexicon keeps one record of each word's frequency; not one which is modality-specific.

More recently, Gaygen and Luce (1998) have found some evidence that there may be separate records of written and spoken word frequency. They presented words either visually or auditorily, and asked subjects to rate the frequency with which they heard, read, spoke and wrote the words. They found that when stimulus words were spoken they were rated more frequently heard and spoken than read or written, but that there was no effect when the stimulus words were written.

They speculated that, whereas a momentary increase in activation of the auditory representation lead to the higher frequency ratings of hearing and speaking words when the stimuli were auditory, when the words were presented visually, they were being phonologically recoded and this masked a corresponding effect of visual representation on the four frequency ratings.

Beyond these basic findings there were several other patterns of results, however, such as auditory presentation leading to higher frequency ratings of reading and hearing words but visual presentation leading to higher ratings of writing and speaking words, so there appeared to be fairly specific use criteria affecting judgements.

To further investigate the relationships, Gaygen and Luce (1998) performed a second experiment using the same words in the first, but which controlled both mode of presentation and mode of response. The words were presented either visually or auditorily as before, but this time the subjects had to either indicate whether the stimulus was a word or a nonword by pressing one of two buttons or to verbally repeat the stimulus word as quickly as possible. This meant there were four tasks: Visual Lexical Decision, Auditory Lexical Decision, Visual Naming and Auditory Repetition.

Important findings were that the only significant correlations in the Visual Naming task were with subjective ratings of reading and writing frequency, that the only significant correlations in the Auditory Repetition task were with subjective ratings of hearing and speaking frequency, and that the majority of the significant correlations in the Auditory Lexical Decision task were with subjective ratings of hearing and speaking frequency. As in the first experiment there were a handful of inconsistent correlations, which Gaygen and Luce again account for with recoding when words are presented in an unusual modality, but overall the suggestion was that the representations of words may be separate for written and spoken words.

In theoretical linguistics, Halliday (1989) has made some arguments for fundamental differences between written and spoken language, even though he was clear on the fact that both were language and it was a mistake to become too obsessed with the medium. For instance, he points out that while gaps, false starts and repetitions are taken to be the distinctive features of spoken language, they are no less widespread during the production of written language. The apparent difference is essentially an artefact of transcription techniques which do not permit the destruction of early drafts.

On the other hand, he does recognise that spoken and written language are used in different contexts and serve different purposes; although both can express the same subject matter, each is specialised to representing certain types of subject. Spoken language may allow for prosodic and paralinguistic information, but written language permits the definition of sentence and paragraph boundaries as well as signalling direct quotation. Spoken language may be preferable for performing brief service encounters, but written language evolved for remote, replicable and asynchronous communication. Most importantly, spoken and written language affect the way the information they represent is interpreted.

Written language is static and dense; the grammar tends to be simple but the substance is compressed into the simple form. Spoken language is dynamic and intricate; there tends to be less substance but the way in which it is represented is more fluid. Whereas spoken language tends to employ verbs which represent the process of events, written language would use nouns which are more metaphorical:

Spoken	Written
applause followed the announcement	after the announcement, people applauded

Halliday argues that this means:

> *Writing creates a world of things; talking creates a world of happening. (Halliday, 1989, p.93)*

Such an explanation may underlie Williams and Simpson's (1976, reported in Simpson et al., 1987) findings that airline pilots expressed less tolerance for false spoken messages than for comparable false visual messages. If the spoken message is taken to represent a dynamic, immediate event, the pilot may find it more compelling and will be annoyed when that call for attention turns out to unnecessary. The written message, being more static and metaphorical, will not be as compelling, will not demand as much attention, and so false messages will be tolerated better. On the other hand, if action is required and the accuracy of messages can be assured, the spoken message may be preferable.

The four studies in this subsection, coming as they do from four very different research paradigms all point to spoken and written language having distinct bases of processing. Besson et al. (1997) and Gaygen and Luce (1998) have provided evidence that there may be different representations for spoken and written words and sentences, but it is the work of Rapp and Caramazza (1997) and Halliday (1989) which is most important for making decisions about when the different representations should be used in the presentation of messages from machines to their operators.

Halliday (1989) suggests that the varying features of spoken and written language reflect the fact that information represented in spoken language is perceived as active and dynamic, whereas information represented in writing is static and about objects. Although Rapp and Caramazza (1997) did not go this far in their analyses, the same interpretation could be made from their results. Their patient lost the content words, but retained the function words in spoken language, but lost function words and retained content words in written language. It could be interpreted that they are maintaining perception of the pattern and flow of events in spoken language, but of the static content of what is involved in events in their written language.

From this, it could be suggested that, if a message needs to represent something changing or dynamic, then speech would be preferable, whereas if something less variable is being represented, written language would be preferable. More tentatively, it could be suggested that, when a message requires a more dynamic response, speech would be preferred, but if no such response is necessary a written message would suffice.

3.0.1.1.4 Design of spoken messages

Considerable attention has been paid to the design of speech output in human factors, with the two aims of increasing the comprehensibility of the output and disambiguating it from other spoken messages.

The clearest differences have been seen in comparison of synthesized and recorded human speech (Pisoni, 1982; Waterworth and Holmes, 1986; Baber et al., 1992).

Pisoni (1982) found that, although subjects were able to remember whether particular words as well in synthesized speech as in human speech, they were less able to recognise whether certain facts had appeared; and similarly, although subjects were able to distinguish between words and non-words, they took longer to do so when speech was synthesized.

Waterworth and Holmes (1986) found that novices listening to synthesized speech found it harder to recall both meaningful and nonsense sentences presented through synthesized speech than through recorded human speech. They found that this difference was removed, however, when a filled delay was placed between the presentation and the recall, which suggests that the difference is because the synthesized speech was harder to encode.

Baber et al. (1992) presented groups of student subjects with simple alarm messages, by both synthesized speech and recorded human speech. They found that although there was no difference in the perceived urgency of the messages, there were differences in the accuracy with which the messages were recalled, with human speech leading to improved recall and more suitable proposed responses.

US Department of Defence standards (1981, refered to in Simpson et al. (1987) and Baber et al. (1992)), specify that machine speech should be clearly disambiguated from human speech. An early suggestion was machine should use female voices, but this presumed that male voices prevailed and can no longer be respected (Simpson et al., 1987) with synthesized speech being the alternative. Reports from pilots in flight simulation studies (summarised in Simpson et al., 1987) suggest they preferred their systems' speech not to sound too human. Cotton and McCawley (1983; refered to in Baber et al., 1992) suggested that giving human features to machine speech lead to pilots assigning too much intelligence to their support systems. Contrary to these findings, however, Simpson et al. (1987) also reported their own studies which found pilots to prefer digitized male speech to digitized female or synthesized speech.

3.0.1.2 Summary of Speech Use Guidelines

The research which has been done on machine-generated speech suggests that it can be useful in many situations. It is potentially useful both for physical reasons, such as its omnidirectionality, and because of differences in the way humans perceive audio and visual, and spoken and written messages.

In some cases, the omnidirectionality of audio communication is required because it cannot be guarantied that operators' visual attention will be focussed in the correct direction to detect a visual message. In other – somewhat related – cases, there may be a potential for visual overload which can be relieved to some extent by presenting some information auditorily. Wickens et al. (1983) provides evidence that relating the machine alerting system to the type of response the operator must make may also be beneficial.

Speech itself has several faculties which make it a particularly useful form of audio communication. Most pertinent is the evidence that spoken messages are inherently alerting, so they may be preferred when immediate action is required. Another advantage speech has is that, whereas mappings between messages represented by abstract and even metaphoric audio messages need to be learnt, machine speech communication depends on a prior learnt language, so no training in the messaging system itself should be required; the ability to inherit languages' pre-existing syntaxes also means complex novel messages can be generated when an alternative audio messaging system would rapidly become unwieldy.

Of course, this latter argument applies equally well to written language as to spoken language, but there is evidence that each has their own beneficial features beyond those inherited from being audio or visual. There is some evidence that each has its own register in the brain, which suggests that overload might be reduced if machine messages provided information in the less used mode. More argumentative is the evidence that speech, due to the information it typically encodes and its typical structure, is more closely associated with action, whereas written language for the same reasons is static. This could suggest that where action is required speech would be preferable, but where the message is more informational or advisory, writing would be more suitable; an interpretation which is given some support by work on pilots' preferences in human factors.

Where speech is to be used, the two issues of discriminability and intelligibility are paramount. The machine's voice should be discriminable from other voices which are going to be heard, but this must be balanced with the difficulty hearers may have understanding distinctively synthesized voices.

3.0.2 The Current Task Design

Because there is little empirical work on many aspects of the way in which alerts are presented to the computer users, it was decided to concentrate on the fundamental issue of mode of presentation.

The design of wording for prompts is an intricate question which is likely to vary considerably with a particular application; and yet has an accepted set of guidelines, which were described in Section 3.0.1.0.

The mode of presentation of a message, on the other hand, has received considerable attention from varying sources, which permits the building of sensible hypotheses about how they will be received, but little empirical evidence to support them. For instance, although there has been considerable and diverse work on the effects of using audio or visual alerts, empirical work has typically investigated the benefits of audio when that channel of communication is unused, as in human computer interaction's desktop computing, or less used, as in flight simulation studies. The use of alerts in situations where the audio channel is otherwise more engaged has not been studied.

Similarly, although blanket statements can been made about speech being more alerting than visual; current graphical interfaces afford various ways in which alerts can be presented visually: each producing potentially different characteristics of response, but having received very little attention. For instance, a simple dichotomy could be made between dialogue boxes and integrated message bars or windows. A modal dialogue box is a common means of presenting an alert which is presented in the centre of the users' attention on a desktop screen and prevents further use of the interface until it has been dealt with. An integrated message window is a graphical messaging device which does not change the gross layout of a computer user's screen, but displays messages when relevant; this is more similar to the presentation of alerts in pilot's HUDs since they do not interrupt any ongoing visually based task.

For these reasons, it was decided to investigate the ways in which the three mentioned prompting styles (dialogue box, integral window and spoken prompt) interacted with an ongoing customer service task which required considerable auditory attention, and in which the outcome of the auditory task was highly important.

The travel game described in Section 2.1.1 was used as the background task, not only for the reasons described there, but also because in that chapter it was demonstrated to be an effective tool for investigating tripartite communication between computers, their operators and their remote customers in the comparison of the style of interface the agent used.

Essentially no changes were made to the task which the agent and customer would be performing, so they still had to build an itinerary of flights around the Western United States and the customer would be given slightly different instructions, which were intended to make them aware of a richer background to the task, so that they had some reason to consider whether the sales opportunities they would be presented had some value. In essence, they were told that they were organising the itinerary for a friend, who wanted to see as much of the locations they would be visiting as possible while there. This would make the product which they would be offered, a local bus and commuter rail travel pass, of potential benefit to them.

In addition to supporting the building of the itinerary, the agent's task also involved selling the local travel pass to the customer. This added more of negotiatory element to the predominantly co-operative itinerary building because the time and money which the customer would need to spend on deciding whether to purchase local travel pass could have been more efficiently spent in planning further links in their itineraries. The benefits in purchasing the travel pass for the customer were only abstract, in that they did not know the passes would be offered before the trial began and could only infer whether the friend who had asked them to build the itinerary would be interested. The benefits to the agent were more concrete, since they knew they would win a prize if they converted more of the prompts they received into sales than any of the other agents. At the same time, the agent had some incentive to support the customer's itinerary building since they were informed that the size of the customer's itinerary would be used as a tie-breaker if more than one agent sold the same amount of local travel passes.

The agent was given instruction in the products which they would be asked to sell, and in how they would be prompted. It was here, in the selection of the prompting styles, that the most important decisions for the design of the study had to be made.

Three styles of prompt were decided upon; a dialogue box, an integral window and a spoken prompt.

The dialogue box was included because it is essentially a default prompting style in human computer interaction; it is the style through which standards and guidelines presume new and important messages are to be presented (see Section 3.0.1.0).

The integral window was an area of the interface which was dedicated to presenting non-compulsory messages to the operators, the aim of this was to simulate the floating windows which had been observed in call centres, but it also reflected the continual guidance frames which had been suggested by Smith and Mosier (1986) (see Section 3.0.1.0) and was closer to the pilots' HUDs used as a visual presentation means in many human factors studies.

Speech prompts were proposed as an alternative to these two means for several reasons. They might reduce workload on the agent's use of their interface (e.g. Deatherage, 1972), but at the same time were likely to increase auditory workload which has received little attention. Inferences from Halliday's (1989) linguistic work and Simpson et al.'s (1987) reports on human factors work suggested that the spoken prompts would be better at encouraging the agents to mention the prompts. Similar inference from Wickens (1984) suggested that, since a verbal action was necessary, a spoken message would be preferable to a visual one. It would also permit the verification of anecdotal observations and suggestions that machine speech output could interfere with other simultaneous tasks involving speech.

3.0.3 Hypotheses

The overall intention in this chapter was to find evidence to support the contention that even relatively small variation in the agent-computer interface could permeate to the agent-customer dialogue. These effects could show immediately after the presentation of the prompt or may be more pervasive and appear throughout the dialogues. The former will be refered to as local effects, and the latter as global effects.

Before describing the hypotheses, though, predictions about the effects of the prompts on the agent's behaviour will be presented because the hypotheses which could be made about observable features in the consultations, were all derived from these predictions. This means that the predictions do not have to be explained during the building of each hypothesis, so that only relevant inferences need to be mentioned.

3.0.3.1 Predictions

The precise hypotheses about the effects of the different styles of prompt depend on a limited number of their features. The representation of the effect of these features in terms of the measures taken will be left until Sections 3.0.3.2 and 3.0.3.3. This section will concentrate on a higher level description of the reactions expected to each style.

It is expected that the dialogue box style of prompt will have immediate effects which force the agent to introduce the sales opportunity to the customer immediately. This may prevent the agent considering which product is being sold and target it effectively at the customer, so that their sales attempts will be less successful. This is because the dialogue box stalls the agent's search for information related to the building of the customer's itinerary, potentially compelling the agent to introduce the opportunity to the customer as soon as possible so that the sale can be resolved before the box is removed from the screen, and the original task returned to with the minimal disturbance.

The integral window is expected to produce a very different behaviour from the dialogue box prompt, because in this case – although the agent is alerted to the presence of an instruction – their ongoing interaction with the software is not interrupted in any way, and so they may select a suitable time during their interaction with the software and dialogue with their customer to refer to the instruction, and by extension may select a suitable time and manner in which to introduce the opportunity to their customer. At the same time, the instruction presented in the integral window might be ignored as the agent continues the ongoing building of the itinerary with their customer.

The spoken prompt has its own individual features. Just as the dialogue box prompt interrupts the agent's interaction with their software, the spoken prompt interrupts their dialogue with the customer. This may force the agent to lose track of any utterances which were being made by the customer, particularly if the spoken prompt overlaps one of the customer's utterances, leading to globally longer dialogues. Again similarly to the dialogue box prompt, the spoken prompt may compel the agent to introduce the opportunity to the customer immediately the prompt is given, leading to local effects. This could be because the agent will be aware that there will be no reminder of the prompted instruction and so will want to transfer the opportunity to the customer before they forget it, or because the spoken prompt is inherently more compelling to action than a visually presented written one (Byblow and Corbett, 1989; Simpson et al., 1987; Halliday, 1989).

Beyond expectations about the local effects of each prompt, it was also recognised that they could have different effects over the whole consultations. The way opportunities were introduced locally could have caused disturbance to the ongoing travel game task, altering the customer's perception of their consultations and potentially causing more dialogue to be required to overcome the disturbance. For instance, spoken prompts, if they interrupt the dialogue between agent and customer may lead to globally longer consultations than the integral window. Alternatively, related to Cotton and McCawley's (1983; refered to in Baber et al., 1992) suggestion that giving overly human speech to a system can lead to that system being assigned too much intelligence, and also taking into account other suggestions that the speech conveys information in a very different way from written language (Halliday, 1989), it could be that the style of prompt could have a more widespread effect on the interaction between agent and computer, which could itself permeate to the whole dynamic between the agent, customer and computer. Specifically, spoken prompts may lead to the computer being seen as a commander, whereas the written prompts are more more likely to be seen as advisory.

All three styles of prompt are therefore expected to have individual effects on the process of the dialogues immediately after they are introduced, and potentially over the whole dialogues. The hypotheses made were first divided into two sections, refered to as the local and global hypotheses, based on this distinction, and each of these sections was further divided into the three Ps of PERCEPTION, PERFORMANCE and PROCESS derived from the literature, described in Chapter 2 and already deployed in the analysis of the open and scripted interfaces. The local hypotheses did not include any PERCEPTION section due to the inappropriateness of asking for judgements of PERCEPTION in the middle of the consultations.

3.0.3.2 Local Hypotheses

The local hypotheses were predominantly about the PROCESS of the dialogues immediately after the prompts had been introduced to the customer. The one exception is the comparison of the conversion of prompts into successful sales which is considered a measure of the agent's success in their PERFORMANCE of the task.

3.0.3.2.1 Local PROCESS: Mentions

The first local hypothesis was based on the expectation that agents would forget or ignore the integral window prompts because they were only alerted and did not impose their message in either the human-computer interaction or the human-customer dialogue. The first hypothesis was, then, that the integral window would lead to fewer conversions of the prompts into mentions of the opportunities to the customer than either the dialogue box or spoken prompts.

H1 The integral window will lead to fewer mentions of the message to the customer than the other two prompting styles.

3.0.3.2.2 Local PROCESS: Turns and Words between prompt and mention

The second hypothesis took into consideration the expected differences in how strongly the prompts compelled the agents to introduce the prompt. It was expected that the dialogue box and spoken prompts would compel the agent to introduce the prompts immediately, but the integral window would permit the agent to consider the best time to fit it into the dialogue with their customer. To measure this, the amount of turns and words between the presentation of the prompt to the agent and its mention were counted. It was hypothesized that there would be more turns and words between the integral window prompt and mentions than between the introduction of the other two styles of prompt and the mention of the opportunity to the customer.

H2 The integral window will lead to more turns and words between the presentation of the prompt and the agent's mention of the sales opportunity, than the dialogue box and spoken prompts.

The counts included all words between the beep which indicated a prompt was about to be given to the beginning of the utterance in which the mention was made.

3.0.3.2.3 Local PROCESS: Overlapping Speech During Mention

It is also hypothesized that there will be some effect of the different prompting styles on the manner in which the agent mentions the opportunities to the customer, the most obvious measure of this is the amount of overlapping speech which occurs during the mention of the opportunities. Because the dialogue box and spoken prompts are expected to compel the agent to mention the opportunities to the customer before they have considered whether it is a suitable time, it can be hypothesized that these styles will lead to an increase in the amount of overlapping speech during the mention because the agent will not consider whether the dialogue is at a suitable turn transition place.

H3 The integral window will lead to fewer occasions of overlapping speech during the mention than either of the other two styles.

3.0.3.2.4 Local PROCESS: Accuracy of Mention

Once the opportunity had been mentioned, a number of additional differences in the reactions to the style of prompt could be measured. One was the accuracy with which the prompts were understood by the agent. It was expected that the integral window would lead to more accurate uptake of the prompted information than the other two styles because continual reference could be made to the instruction, that the dialogue box would lead to less accurate mentions because – although the message could be reviewed while the box was on screen – the agents might remove the box once they had read the instruction and so not be able to review it while conveying the opportunity to the customer. The spoken style was predicted to be the least accurate because the instruction was only conveyed once and permitted no review.

To permit this to be measured, two potential levels of error were added to the task. The agents were told that they might be prompted to sell one of two products to the customer, so they had to recognise which one had been prompted to them, and secondly they had to recognise which airport the product applied to: the airport was refered to by an index number in the prompt but the agent was advised to refer to the airport by its name.

The capability of each prompting style was measured by the accuracy with which the opportunity in the prompt was transferred to the customer. If the agent mentioned the correct product, and offered it in the correct airport, this was considered an accurate mention; if not, it was inaccurate.

H4 The integral window will lead to more accurate mentions than the other styles, the dialogue box to less accurate mentions, and the spoken style to the least accurate mentions of all.

3.0.3.2.5 Local PERFORMANCE: Sales

The final hypothesis about the effects of the style of prompt immediately after their presentation is that each style will lead to differences in the number of prompts successfully converted to sales. This is expected to be due to differences in the way in which the agent tries to sell the products as a consequence of the urgency with which the style of prompt compelled the agent to mention them. For instance, if the spoken and dialogue box prompts lead to the agent reacting to the prompt without considering how best to sell it to the customer, they may also not have time to plan how to sell the product to the customer and this may lead to fewer sales. If on the other hand, the integral window leads to a better timing of the mention, and more consideration of how best to present the opportunity, this may lead to more sales. The fifth hypothesis is, then, that the integral window will lead to more sales than the dialogue box and spoken prompts.

H5 The integral window will lead to more sales than the dialogue box and spoken prompts.

3.0.3.3 Global Hypotheses

Effects on global aspects of the dialogue as a function of the style of prompt could have occurred for two reasons. Either the way in which the agent reacted locally to each presentation of a prompt could have directly or indirectly lead to global differences, or the style of prompt could have lead to the agent having a different expectation of their interface which would change the dynamic of the tripartite communication between agent, customer and computer. In the former case, varying disturbance of the ongoing travel game task could have lead to differences in the amount of discourse needed to overcome that disturbance. In the latter case, for instance, the integral window may have lead to the agent considering their computer an advisor, whereas the spoken prompt and dialogue boxes may have lead to them considering it a commander. The visible effects of these two potential explanations for global differences need not be mutually exclusive, so some caution needs to be taken in interpreting any global differences.

A more complete description of the expectations of the effects of each style of prompt was provided in Section 3.0.3.1, and where it is relevant reference will be made to those features during the description of the global hypotheses. Following the distinctions drawn in the literature review in Chapter 1 and used during the analysis of the Open and Scripted interfaces in Chapter 2, the global hypotheses about this comparison will be divided into measures of PERCEPTION, PERFORMANCE and PROCESS.

3.0.3.3.1 Global PERCEPTION: Preference

It was predicted that the different styles of prompt might alter the customers' PERCEPTION of their interaction and service, particularly if the agents' distraction from the itinerary building task was evident. This was tested through attitude questionnaires.

For instance, the integral window, which was expected to lead to introduction of the opportunities to the customer in more sensitive manner than the other styles of prompt (See Section 3.0.3.1) was also expected to lead to the customers having a higher regard for those consultations than for those after the other two styles. Precise reasons for hypothesizing this perceptual difference could be the agent feeling compelled to mention the opportunity when they would predict the sale would be unlikely, increased disturbances to the agent-customer dialogue as the agent is distracted from the ongoing itinerary building by the more aggressive prompts and damage to the customer's faith in the agent because they introduce the opportunities when the customer expects the agent to be ready to perform other actions for them.

> **H6** The customers will have a more positive PERCEPTION of the consultations in which the agent uses the integral window than the consultations in which the other two styles of prompt are employed.

It was recognised, however, that the nature of the task could mean that questionnaires administered after the consultation might not capture the PERCEPTION effects well. Pilot tests had shown that the trigger for the prompt could only guarantee the presentation of 4 prompts within the 15 minute time limit.

In many cases, however, subjects triggered all four prompts early in the dialogues, leaving a large section at the end of consultations where no prompt was received. It is possible that any direct effect of the introduction of the opportunity on customers' PERCEPTION could be dissipated during such a period. In brief, if the 4 prompts were triggered early in a consultation the customer could have forgotten their effect by its end.

3.0.3.3.2 Global PERFORMANCE: Airports and Time

As during the experimental comparison of open and scripted human-computer interface styles, two global measures of PERFORMANCE were taken; the number of airports included in the itinerary and the time taken to complete the task. It is not hypothesized that there will be any difference in the number of airports visited, because any variation in airports was likely to be either more due to personal differences than the style of prompt, or subsidiary to more pronounced differences such as in the time taken. It was, on the other hand, hypothesized that there would be a difference in the time taken to complete the task.

As was noted in Section 3.0.3.2.2, when the hypotheses about turns and words between the prompt and its mention were being made, it is possible that the dialogue box and spoken prompting styles, which are expected to compel the agent to mention the opportunity before considering whether it is a good time to do so, may lead to globally longer dialogues. This is because the agent will leave the resolution of the customer's demands hanging while they deal with the sales opportunity, and will have to spend time rebuilding their common ground once the opportunity has been resolved. This difference may be accentuated if the further hypothesis that the integral window will lead to fewer mentions than the other two styles (**H1**) is supported, since no mention implies no time being spent on the sale.

H7 The dialogue box and spoken prompting styles will lead to longer consultations than the integral window.

The same hypothesis could also be made if the style of prompt was having a more general effect on the dynamic of the communication between agent, customer and computer. From a number of sources (e.g. Gaver, 1989; Halliday, 1989 and Williams and Simpson, 1976) it could be inferred that the spoken prompt could lead to the computer's role in the relationship between the agent and computer being commanding, whereas the integral window could lead to it being advisory. Gaver (1989) suggested that changing and changeable information should be represented through sound, whereas constant information would be better represented through visual means. Halliday (1989) suggested that speech was the medium of action, whereas written language was the medium of facts. Williams and Simpson (1976, referred to in Simpson et al., 1987) found that pilots preferred speech for critical information which required action, whereas they preferred more advisory information to be given visually, and were less tolerant of erroneous spoken messages than erroneous written messages possibly because the spoken messages compelled a response.

The dialogue box being written would be expected to produce a dynamic more like the integral window, but by demanding attention would be expected to produce a dynamic more like the spoken prompt. Extending this, given that Wickens' results (1983, see Section 3.0.1.1.1) suggested that spoken alerts would facilitate spoken responses, it could be hypothesized that the spoken prompts will lead to more speech within the agent-customer dialogue and therefore longer consultations than the integral window. Because there are two recognised influences on the way in which the dialogue box could affect the overall relationship between agent and customer, that style of prompt's global effects would however be less certain.

3.0.3.3.3 Global PROCESS: Turns and Words

The same three measures of PROCESS (turns, words and overlapping speech) used in the study of open and scripted interfaces (Chapter 2) will be used to assess the global impact of the different styles of prompt on the dialogues between the agent and customer.

As with the measure of time described above, and for the same reasons, it is expected that there will be an effect of the style of prompt on the amount of turns and words used. In particular, it is hypothesized that the dialogue box and spoken prompting styles will lead to more turns and words being used than when the integral window is used.

H8 More turns will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

H9 More words will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

As with the hypothesis about the time required for the consultations (**H7**), if the difference is being lead by a change in the overall relationship between agent and computer, rather than local effects, it is less clear whether the dialogue box would be expected to produce behaviour more like the spoken prompt or the integral window.

3.0.3.3.4 Global PROCESS: Overlapping Speech

Overlapping speech which occurred during the whole dialogues was marked and measured because it has received considerable attention since Sacks et al. (1974) recognised its usefulness as a measure of the structure of conversation.

Its primary usefulness has been in distinguishing between grossly different genres and loci of conversation. For instance Sacks et al. (1974) distinguished between lectures and spontaneous conversations, Stephenson, Ayling and Rutter (1976) were able to distinguish between negotiative face-to-face and telephone conversations and Sellen (1992) was able to distinguish between face-to-face and video-mediated discussions (see also Section 1.3.2.1.3.2). Boyle, Anderson and Newlands (1994) found that subjects produced more occurrences of overlapping speech when they were screened from each others' sight, and that familiar subjects produced less occurrences than unfamiliar subjects.

Greatbatch et al. (1993) also suggested that in tripartite communication between a doctor, customer and computer, the doctor's interaction with their computer was taken into account by the customer during the planning of their turns.

In the current study, it is unlikely that the style of prompt will produce a noticeable effect on the amount of overlapping speech over the whole dialogue. It is possible that there will be slightly more overlapping speech during the mention immediately after the more compelling styles of prompt than after the less compelling (see Section 3.0.3.2.3; **H3**), but since there are only four prompts per dialogue and overlapping speech seems to occur in 10-20% of utterances in spontaneous spoken dialogue, it is unlikely that this will produce a noticeable effect over the whole dialogues. Similarly, there is no reason to believe that variation in the overall relationship between agent and computer would lead to differences in the amount of overlapping speech between the agent and customer unless one style of prompt leads to major disfunction. Therefore no hypotheses were made about the effect the style of prompt would have on the amount of overlapping speech.

3.0.3.4 Summary of Hypotheses

The hypotheses are repeated here for ease of reference:

H1 The integral window will lead to fewer mentions of the message to the customer than the other two prompting styles.

H2 The integral window will lead to more turns and words between the presentation of the prompt and the agent's mention of the sales opportunity, than the dialogue box and spoken prompts.

H3 The integral window will lead to fewer occasions of overlapping speech than either of the other two styles.

H4 The integral window will lead to more accurate mentions than the other styles, the dialogue box to less accurate mentions, and the spoken style to the least accurate mentions of all.

H5 The integral window will lead to more sales than the dialogue box and spoken prompts.

H6 The customers will have a more positive PERCEPTION of the consultations in which the agent uses the integral window than the consultations in which the other two styles of prompt are employed.

H7 The dialogue box and spoken prompting styles will lead to longer consultations than the integral window.

H8 More turns will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

H9 More words will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

3.1 Procedure

The description of the procedure will be divided into 5 sections. The first, apparatus, will describe the software and off-line resources available to the agent. The second will describe the subjects which took part and their motivation. The third describes the training and instruction given to the subjects before the trials began, while the fourth describes the events which took place after each trial. The final briefly explains how the dialogues were transcribed.

3.1.1 Apparatus

The author used Hypercard 2.3 to build the interface which the agent would be using during the consultation.

Much of the code for discovering which airports were available and displaying that information was inherited from the interface used in the comparison of open and scripted interface (Chapter 2), but the interface was updated for the specific experiment (see Figure 3/3).

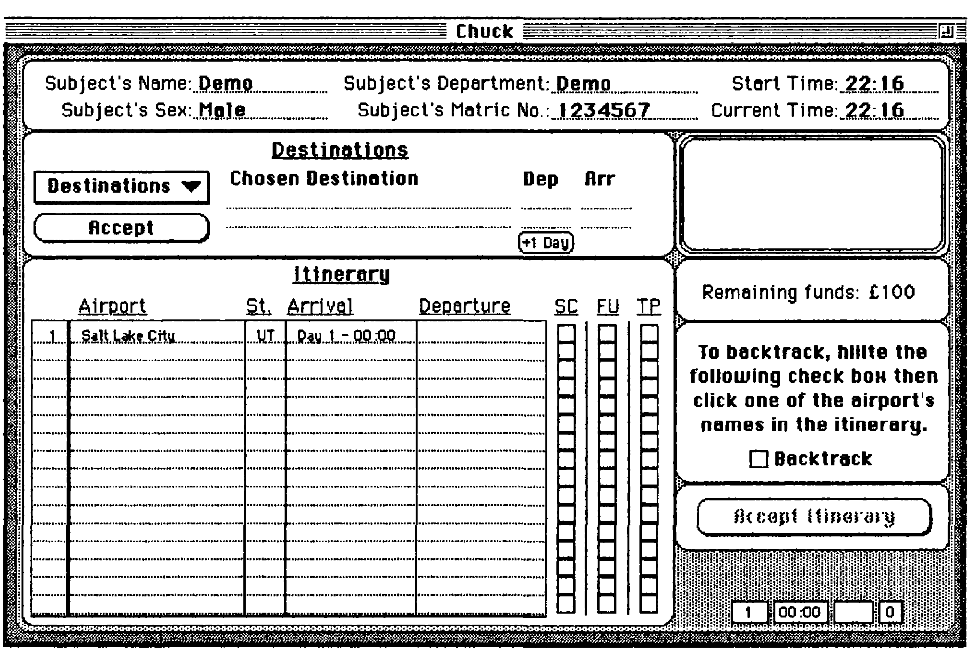


Figure 3/3 The Basic Interface.

Because the intention was to simulate an outbound sale being prompted during an inbound call[1], the open interface from experiment 1 was used as the basis for this experiment's interface. This interface is currently more common in inbound calls because it allows the agent more free movement through a company's databases when the customer's intentions are not known before they initiate a consultation. This interface was then altered to simplify it slightly to make the training of novice, non-confederate, agents easier, and to gain screen "real estate" for the integral window condition.

In addition to this, code was added to detect situations in which a sale was possible, then trigger the presentation of prompts to the agent through one of the three styles. These three prompting methods comprised the three conditions under investigation for their effect on the dialogues between agent and customer and were a dialogue box, an integral window or spoken prompt. In all cases the prompt was preceded by a beep which signalled that a prompt was forthcoming and served to mark the occurrence of a prompt in the audio recordings of the dialogues between the agent and customer.

In the dialogue box condition (Figure 3/4), when a sales opportunity was detected by the software, a message telling the agent to sell the travel pass and which airport the offer applied to was displayed in a dialogue box. When this box appeared, the rest of the screen was locked, and the agent could only continue with the planning of the itinerary by clicking a button within the dialogue box to acknowledge that they had read it.

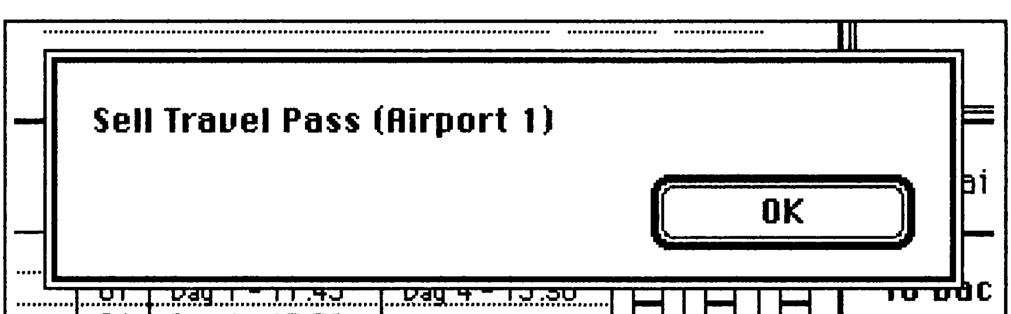


Figure 3/4 The dialogue box in use.

---

[1] I.e. the customer has initiated a call to the agent, but the agent is reacting to developments during this call which make the sale of a relevant product possible.

In the integral window condition, the prompts were presented in a field in the top right hand corner of the interface which was dedicated to this purpose (see Figure 3/5). These prompts were displayed in the same font and at the same size as in the dialogue box condition. At the beginning of the consultation, the field they would appear in was clear, and when the agent clicked a button in the interface to indicate they had sold the travel pass or entered a new destination into the itinerary its contents were cleared again. This was intended to restrict the presence of the prompt on screen to the duration of its relevance.

Figure 3/5 The integral window in use.

In the spoken condition, a digitised recording of the prompt, spoken by a university-educated female Scot, was played to the agent. This is against the US Department of Defence standards (1981, refered to in Simpson et al. (1987) and Baber et al. (1992)) for speech interfaces which specify that speech used as alarms should be clearly synthesized so that they can be distinguished from human speech and against suggestions by Cotton et al. (1983, refered to in Baber et al., 1992) that computer speech which sounds too human may lead to users attributing too much intelligence to their machines, and arguably contrary to the standard HCI contention that computer systems should behave like tools, and not have human facilities attributed to them (Shneiderman, 1998). The choice was made because experimental work had shown that – at least for novices – messages presented through synthesized speech was harder to recall and respond to appropriately than human speech (Pisoni, 1982; Waterworth and Holmes, 1986; Baber et al., 1992). Indication that the speech was emanating from the computer was provided by the simple beep which preceded all three prompting styles.

In the dialogue box and integral window conditions the wording of the prompt was the same; the airport to which the prompt applied being included in brackets below the mention of the product to be sold. In the spoken condition, the preposition "for" was introduced between the mention of the product and the airport it applied to.

The agents were told that they might receive prompts for two products, and those two products were described to them. This was so that they had to read the content of the prompts' instructions to determine which product should be mentioned, and so permit measurement of the accuracy with which each style of prompt lead to its message being transferred to the customer. In fact, all the agents received instructions to sell the same product each of the four times they were prompted.

The products which were prompted were chosen because they could be repeated several times throughout the journey, related to different airports (Figure 3/6). This would not have been possible if a product whose range was the whole itinerary, such as travel insurance, were being prompted.

Prompts

Since the ticket the customer has purchased is for two people the surcharges indicate the amount to be paid for both travellers, not individually.

Prompt:	**"sell travel pass"**
Details:	Local travel pass
	Surcharge: £10 for 2 people
	This pass permits unlimited travel on local bus and commuter rail for the duration of your stay in the town. It is offered for travel around the last location the customer visited.
	This provides a perfect way to get around the town and its surrounding area.
	The pass has been arranged privately between Pacific Airlines and the local travel companies and must be bought at the same time that the itinerary is booked.
	If the customer accepts, mark "TP" in the relevant airport's notes section.
	This will only be offered if the customer spends a complete daylight period in a town, but may not be available on all cities or at all times.
	Hilite the "TP" button by the penultimate airport.
Prompt:	**"sell flight upgrade"**
Details:	Flight upgrade
	Surcharge: £15 for 2 people
	This upgrades the quality of the in-flight arrangements on the last flight the customer booked.
	The improved arrangements include more comfortable seats and a personal entertainment system which permits the customer to choose the video they watch.
	The upgrade must be bought at the same time that the itinerary is booked.
	This will only be offered is the customer is taking a flight of 3 hours or longer, but may not be available on all flights.
	Hilite the "FU" button by the current airport.

Figure 3/6 Product information given to the agent.

As well as the off line description of the prompts, the agent also had access to the same resources available to the agent in the comparison of open and scripted interfaces (Chapter 2). This comprised a sheet of paper contained a summary of the rules of the travel game, a list of the airline's abbreviations for State names, since the airport listings only used the abbreviations, and a list of stock phrases to use if the customer remained silent or was having trouble selecting a prospective destination.

3.1.2 Subjects

Ninety subjects, all students in the Strathclyde region who did not know each other before the experiment, participated. 45 of these were trained and served as agents and 45 served as customers. All subjects were paid £2 for taking part. The agents were aware that a £15 pound prize was available for the agent who performed best on a simple criterion, and also that a similar prize was available to the customers. The customers were only made aware of the prize available to themselves.

3.1.3 Procedure

3.1.3.1 Customer Instructions

While the customer read a set of instructions similar to those used in the first experiment, but providing a background story to explain why they were planning the trip and intended to provide some motivation to purchase the products they would be offered (See Figure 3/7), the agent was trained in the use of the interface.

Customer Information

I will answer any questions you have about the rules after I have trained the agent.

Instructions

One of your friends has joined a scheme which allows students to work abroad in the US over the summer holiday.

Since they payed a deposit early they were entered in a prize draw and have won a 3 week trip for 2 around the western United States. You have been invited to come along, on the condition that you organise the route you will take. All the flights have to be booked before you leave the UK, and this is the purpose of the "call" you are now making.

The Pacific Airlines Travel Pass you have been given permits unlimited travel between airports the airline uses for 22 days, with only two restrictions.

Firstly, you must spend at least 3 days in each state, but may travel to other airports within the state during that time.

Secondly, every time you visit the same airport after the first there will be a £50 surcharge. You and your friend have agreed to spend at most £100 on such travel arrangements.

Your journey will begin at Salt Lake City; a Pacific Airlines hub which has links to most other airports.

Your friend wants to visit as many cities as possible so that they can decide where they would most like to spend their summer. Therefore, your objective is to visit as many airports as possible.

If at any time, you reach a location that you find it difficult to advance from you may "backtrack" to an airport you have previously been to. This will erase your itinerary from that point onwards, including any surcharges.

The call you make may take no longer than 15 minutes – the agent will advise you when this time is up.

To demonstrate your friend's pleasure if you organise the route well, the customer who arranges the itinerary visiting the most airports will win a £15 prize.

Rules Summary

- You must **plan to visit as many airports** on Pacific Air routes as you can in **22 days**.
- You must spend **3 days in every state** you visit, but may travel between airports within the state during that time.
- You can afford **£100** in excess surcharges.
- You may "**backtrack**" at any point in the game without cost; it is your final itinerary which counts.
- There is a real **time limit of 15 minutes** on your consultation with the travel agent.

Don't forget you are in a competition!
Good luck!

Figure 3/7 The customer's instructions.

3.1.3.2 Agent Training

After being given a description of the rules of the Travel Game, the agents were trained in the software they would be using. Firstly, it was explained that before they began the Travel Game, they would be asked to enter a few details about the customer, then they were taught how to find out information about flights, enter a choice into the itinerary and how to erase choices if the customer wanted to change their plans. Once the agent was familiar with this, they were told about the sales prompts they would be receiving, the style of prompt they would be receiving was demonstrated, and they were taught how to indicate that they had made a successful sale. They were not made aware of the other styles of prompt.

The agents were told that they might receive prompts for either of the two products described in Section 3.1.1, whereas in fact all the prompts for every agent were for the same product. The inclusion of two ostensibly possible products was so that the accuracy with which the agents mentioned the correct opportunity to the customer could be measured. If there was only one product, the agents would have had no need to consider the content of the prompts, and the task would have become one of stimulus and response.

The agents were not told how many prompts they would receive, but all received four which was the most pilot tests had shown could be guarantied within a suitable time limit.

The agent's active training took between 15 and 20 minutes, but they were also allowed around 5 minutes to review the prompt information and practice with the interface while customers' questions were being answered.

3.1.3.3 Training Completion and Trial

At this point the agent was left to practice with the interface and to read the information on the products they might be asked to sell, while any questions the customer had about the rules were answered. Once their questions had been answered, the difference between surcharges and backtracking was clarified because pilot tests had shown them to find this confusing.

The experimenter then returned to the agent to resolve any difficulties they had had with the interface and answer questions about the task or prompts. When they were satisfied that the agent could perform their role, the agent and customer tested the headphone link they would be using and when they could communicate comfortably the trial began.

3.1.3.4 Questionnaires and debriefing

After the consultation, when the ticket had been fully booked or the 15 minute time limit had been reached, both agent and customer were given the same 26 item questionnaire. Once they had completed the questionnaires and been paid for their participation and given a debriefing about the purpose of the study which included a description of the styles of prompt they had not received.

3.1.5 Transcription

After the trials, the audio recordings were transcribed and marked with occurrences of overlapping speech. This process took approximately 2 hours for every 15 minute dialogue.

The transcription divided each speaker's utterances into turns, which were defined as sections of the dialogue in which they were the primary speaker, but during which the other participant could produce sections of overlapping speech.

The way in which these sections of overlapping speech was marked captured the exact nature of the overlapping sections. The transcription indicated whether the utterances which overlapped with the original floor holder's speech ended before the floor holder's utterances or not, and only overlapping speech which extended beyond the original floor holder's utterances were considered to be new turns. Occasions in which the overlapping speech extended beyond the original speaker's utterances and new turns were added to the transcript were refered to as INTERRUPTIONs, and occasions where the overlapping speech ended before the original speaker's utterances were refered to as OVERLAPs. These distinctions reflect those suggested by Sellen (1992), who used them to detect effects of different styles of video-mediation on casual discussions.

3.1.6 Vignette

To demonstrate how well the subjects took to their roles in simulating a real customer service situation, the vignette in Example 3/1 is provided.

• Agent
now what you're going to [um], need there I would say is probably the travel pass, and this gives you unlimited travel while you're there [em] on the local bus and commuter rail [em] for the duration of your stay in the place, so obviously that's [em] fairly indispensable for getting around [eh] a large city like Seattle

• Customer
yeah

•

• Agent
so you you would like the travel pass then

• Customer
yeah I would, yeah

Example 3/1 Vignette demonstrating participants' immersion in their roles during audio-only prompting task.
(Spoken Prompt, Sixth Trial, First Prompt).

The main indication that the agent has taken to their role is their inclusion of information outside that given in their instructions about the travel pass. In this case, the agent reminds the customer that Seattle is a large city where the travel pass will be particularly valuable.

3.2 Results

Following the structure introduced in the description of the hypotheses, the results section will be divided into 2 subsections; one detailing the analyses of the consultations local to the prompts, and the other detailing the analyses of global effects on the consultations. Each of these subsections will be further divided into analyses of PERCEPTION, PERFORMANCE and PROCESS, though the local section included no analysis of PERCEPTION.

3.2.1 Local Results

The order in which the results of the analyses of the local effects of the prompts will be reported follows the order in which effects would be likely to appear immediately after the prompt's presentations. Therefore, contrary to the order of presentation of results in the comparison of open and scripted interfaces in Chapter 3, the analyses of the PROCESS of dialogues immediately after the prompt will precede the analysis of the PERFORMANCE. In this way, it should be possible to build a step-by-step picture of how the dialogues progress immediately after the prompts.

3.2.1.1 Review of local hypotheses

In the introduction (Section 3.0.3.2), the following hypotheses were made about the PROCESS and PERFORMANCE in the consultations local to the prompt, the description of the results will follow the order in which they are presented.

H1 The integral window will lead to fewer mentions of the message to the customer than the other two prompting styles.

H2 The integral window will lead to more turns and words between the presentation of the prompt and the agent's mention of the sales opportunity, than the dialogue box and spoken prompts.

H3 The integral window will lead to fewer occasions of overlapping speech than either of the other two styles.

H4 The integral window will lead to more accurate mentions than the other styles, the dialogue box to less accurate mentions, and the spoken style to the least accurate mentions of all.

H5 The integral window will lead to more sales than the dialogue box and spoken prompts.

3.2.1.2 Testing of local hypotheses

Counts were made of the number of prompts which the agent mentioned to the customer to test Hypothesis 1. Separate counts were made of the number of turns and words between the prompt and its mention to test the two related parts of Hypothesis 2. The number of occurrences of overlapping speech during the mention of the opportunity were counted to test Hypotheses 3. The number of times the agent mentioned the correct product and location were counted to test Hypothesis 4; because only one of the potential measures of accuracy showed any variation this measure actually represents occasions on which the agent mentioned the correct airport. Finally, a count was made of the number of prompts converted into sales to test Hypothesis 5.

All these measures were analysed by One-way ANOVAs with style of prompt as a between subjects factor (3 levels: dialogue box, integral window and spoken prompt). The mean number of occurrences of each measure within each dialogue, and the results of the inferential statistics are summarised in Table 3/1 and discussed in the following paragraphs.

Measure	Dialogue Box	Integral Window	Spoken Prompt	ANOVA	P
Mentions	3.7	3.3	3.7	F[2,42] 1.063	0.354
Turns between prompt and mention	4.5	3.9	4.1	F[2,42] 0.037	0.964
Words between prompt and mention	26.6	28.1	22.5	F[2,42] 0.075	0.928
Correct Airport	1.1	1.0	2.0	F[2,42] 1.826	0.174
Sales Made	1.5	1.8	2.3	F[2,42] 1.071	0.352

Table 3/1 Statistical analysis of local effect of style on the consultations between the agent and customer.

None of the these measures supported the local hypotheses. Although as the hypothesis suggested, there appear to be fewer mentions made after the integral style than after the dialogue box or spoken style, this is not statistically significant (One-way ANOVA, F[2,42]: 1.063, P: 0.354) . Similarly, even though there appeared to be minor differences between the conditions in the number of words between the prompt and mention of the opportunity to the customer in directions which would have supported the hypothesis, neither the variation in turns nor in words was significant (Turns: One-way ANOVA, F[2,42]: 0.037, P: 0.964) (Words: One-way ANOVA, F[2,42]: 0.075, P: 0.928).

The number of occurrences of overlapping speech was not analysed statistically because there were exactly 4 occurrences after each style of prompt out of a potential 60. Because there were so few occurrences and because there was no significant difference between the styles in the number of mentions, it was concluded that any analysis of the number of occurrences of overlapping speech as a proportion of the number of mentions would equally not be warranted. Although the percentage of overlapping speech is somewhat low after all styles of prompt, it is within the range of normal spontaneous conversation.

As was described in the introduction, two levels of accuracy in the mention of the opportunity to the customer were accounted for. The broad level accounted for the agent describing the incorrect product to the customer, the finer level was whether the agent mentioned the correct airport since the prompt refered to the airport's index number in the itinerary but the agent was advised to mention it by name. The measure counted occasions when the agent refered to "that airport" or "there" as incorrect because the closest references for these referents would be airports mentioned more recently in the dialogues than the ones the prompt's instructions refered to.

The broad level of accuracy was not analysed because in all cases the agents either mentioned the correct product immediately or rapidly corrected themselves. The apparent difference between the styles in the number of correct airports with the spoken prompt leading to roughly twice as many correct mentions as both the dialogue box and integral window prompts, which also appears to be exactly contrary to the hypothesis made, was not statistically significant either (One-way ANOVA, $F[2,42]$: 1.826, $\underline{P}$: 0.174).

Finally, the overall measure of the agent's success in presenting the opportunity to the customer, as measured by the number of prompts successfully converted into sales, was analysed. Although there did appear to be slight variation between the styles of prompt, the directions of variation did not directly match those hypothesized and no significant difference was found (One-way ANOVA, $F[2,42]$: 1.071, $\underline{P}$: 0.352).

3.2.1.3 A Composite Measure of Local Success

Because there was so much apparent variation between the styles, but no significant differences, a composite measure of the success of the prompts in persuading the agent to transfer their instruction to the customer was built by summing the number of mentions, correct mentions and sales. The other measures of turns, words and overlapping speech were not included because they assessed more interpersonal factors than those chosen. The analysis of this composite measure is displayed in Table 3/2.

Measure	Dialogue Box	Integral Window	Spoken Prompt	ANOVA	P
Composite measure of prompt execution	6.3	6.1	7.9	F[2,42] 2.421	0.101

Table 3/2 Composite measure of prompt execution.

Again, although there again appeared to be some variation in the styles, with the spoken prompt being apparently more successful than the other two styles, no significant difference between the styles was found (One-way ANOVA, F[2,42]: 2.421, P: 0.101). Were the spoken prompt to have lead to more more successful presentations of the opportunities according to this composite measure than the other styles this would have been surprising because the integral window was hypothesized to be more effective on two of the components.

3.2.2 Global Results

The lack of support for the hypotheses made about the PROCESS of the dialogues and PERFORMANCE directly related to the presentation of the prompts, is somewhat surprising, but it was also hypothesized that local variation could lead to effects which could be measured globally. For instance, the variation in demands which each style of prompt placed on the agent, could have lead to more common ground being lost and subsequently more discussion being required to regain that common ground.

Following the structure which had already been established, the measures which were used to test the global impact of the prompts were perceptual, PERFORMANCE related and PROCESS related. The four hypotheses related to these three measures will be presented in each section.

3.2.2.1 PERCEPTION

To assess the impact of the prompts on the customers' PERCEPTION of the dialogues they were given attitude questionnaires and these were statistically analysed.

3.2.2.1.1 Description of Questionnaires

Each customer received a 25 question questionnaire comprising three sections. One section used 7 point scales based on Osgood's semantic differential scores, one section used 5 point scales based on scales of customer satisfaction derived during British Telecom's Customer Interface Quality Measurement ("CIQM") project (Marle and Coyle, 1995), and one used questions of varying styles directly addressing features of the consultation and communication with their partner. The 5 point "CIQM" scales were repeated refering to the "consultation" in one case, but the "partner" in the other. This was primarily intended to test whether the customer could distinguish between features of the service which were due to the agent and features which were due to the interaction between the agent, software and themselves.

The "Osgood" section was introduced with the statement "Please rate your consultation on the following scales. *Note that you are rating the consultation as a whole, not just your partner.*", and included the following scales:

- Unpleasant -- pleasant
- Still -- lively
- Tender -- Tough
- Bad -- Good
- Passive -- Active
- Weak -- Strong

The "CIQM" questions were introduced in two different fashions, the questions refering to the consultation were introduced with exactly the same line as that used for the Osgood questions, whereas the questions refering to the respondent's partner were each introduced as separate questions with a line such as "How polite did you feel that your partner was?" This distinction served to remind the respondent to distinguish between the agent and consultation. In all cases the questions' introductory statements named the most positive degree of the scale on which the partner was to be judged, but the points of the scales the respondent had to mark were all named.

- Impolite -- polite
- Inefficient -- efficient
- Unfriendly -- friendly
- Unassuring -- assuring
- Incompetent -- competent
- Unsympathetic -- sympathetic
- Unhelpful -- helpful

The other questions varied in style. Four of the questions were five point scales, which used the pertinent adjective from the statements below. The other question, “Did you feel that you had lost touch with you partner at any time?”, requested a simple yes-no response.

- How easy was it to hear your partner?
- How easy was it to make changes in your itinerary?
- How satisfied were you with the final outcome of the consultation?
- Did you find the conversation natural?
- Did you feel that you had lost touch with your partner at any time?

3.2.2.1.2 Analysis of Questionnaires

The hypothesis had been made that the customers would prefer the consultations in which the agent’s prompts were presented via the integral window to those in which the other styles of prompt were used. This was essentially because the agent would introduce the opportunities to the customer in ways which showed more concern for the customer’s desires after the integral window.

H6 The customers will have a more positive PERCEPTION of the consultations in which the agent uses the integral window than the consultations in which the other two styles of prompt are employed.

It was however recognised that, due to the design of the task, there may have been some time at the end of the consultations during which no prompt or opportunity was presented, and this may have reduced any effect of the prompts.

The customers' responses to the scalar questions were analysed using Analysis of Variance which was considered appropriate for the same reasons described in Section 2.2.1.

Although several differences appeared to be occurring, none were found to be statistically significant. For instance, on the Osgood rating of pleasantness, whose 7 point scale ranked from "Very Unpleasant" to "Very Pleasant", although the customers rated 12/15 of the consultations using both the dialogue box and spoken prompts, as "Fairly Pleasant" or higher (points 5-7 on the scale) but only 9/15 of the consultations using the integral window as similarly pleasant, a One-way ANOVA with type of prompt as a between subject factor (3 levels: dialogue box, integral window and spoken prompts) found no significant difference between the styles (One-way ANOVA, $F[2,42]$: 1.176, $\underline{P}$: 0.318).

Similar results were found for the other two semantic differential scales which looked promising. Customer ratings on the liveliness scale, which ranged from "Very Still" to "Very Lively", included 2/15 ratings of the dialogue box and spoken prompt consultations which were "Fairly Still" or lower (points 1 to 3 on the scales) and 7/15 ratings of the integral window in the same range, but One-way ANOVAs with style as a between subjects factor (3 levels: dialogue box, integral window and spoken prompts) was not significant (One-way ANOVA, $F[2,42]$: 1.416, $\underline{P}$: 0.254). Customer ratings on the Strength scale, which ranged from "Very Weak" to "Very Strong", included 5/15 rating of the integral window consultations as "Fairly Weak" or lower, 3/15 of the dialogue box consultations in the same range, and 1/15 of the spoken prompts, but a One-way ANOVA with style as a between subjects factor (3 levels: dialogue box, integral window and spoken prompts) was not significant (One-way ANOVA, $F[2,42]$: 1.776, $\underline{P}$: 0.182). Analyses of the remaining responses to the Osgood scales with similar One-way ANOVAs showed no probability lower than those reported.

None of the rating on the CIQM scales even hinted at a difference between the styles, and statistical analyses using One-way ANOVAs showed no probability lower than 0.435.

On the scales which had previously been used, only the responses to the questions "How easy was it to make changes in your itinerary?" suggested a variation, with the customer rating 9/15 of the dialogue box, 8/15 of the integral window and 11/15 of the spoken prompt consultations as "Fairly Easy" or "Very Easy" on the 5 point scales, but a One-way ANOVA with style of prompt as a between subjects factor (3 levels: dialogue box, integral window and spoken prompts) was not significant (One-way ANOVA, $F[2,42]$: 1.102, P: 0.342). Statistical analyses of the remaining 3 scaled questions showed no lower probability than this. Chi-squared analysis of the 2 × 3 contingency table of customers' "Yes" or "No" responses to the question "Did you feel that you had lost touch with your partner at any time?" revealed no statistically significant relationship between the style of prompt and the number of positive answers (Chi^2 $(2, N=45) = 1.03$, P: >0.05).

3.2.2.2 PERFORMANCE

The global PERFORMANCE of the task was measured by how many airports the customer managed to fit into their itinerary and the time that the whole consultation took. The number of airports in the itinerary was essentially a measure of the customer's performance since their aim was to visit as many as possible, and were aware that a prize would be given to the customer who arranged the most. The time taken was more a measure of the shared success, since the companies employing agents see clear monetary benefit when calls are shorted and customers are also aware of costs they may incur. In a related fashion, the number of sales the agent made, which was reported in Section 3.2.1.2, was used as a direct measure of the agents success.

It was not hypothesized that there would be any significant variation in the number of airports which would be visited, but it was hypothesized that there would be a difference in the total time taken to perform the trials. The latter difference was expected to be either because the dialogue box and spoken prompts would lead to greater disruption of the ongoing travel game task, leading to loss of common ground which would require time to rebuild, or because the style of prompt is affecting the dynamic of the interaction between the agent and their computer causing the agent to take more or less time with their customers. This latter could either be because, as could be inferred from Wickens et al. (1983, see Section 3.0.1.1.1), the spoken prompts would generally encourage agents' speech or because, as could be inferred from Gaver (1989), Halliday (1989) and Williams and Simpson (1976) (see Sections 3.0.1.1.1 and 3.0.1.1.3), the spoken prompts cause the computer's role in the interaction between agent and computer to be one of commander whereas the integral window causes its role to be advisor. It was also recognised that the two effects are not mutually exclusive and could both have lead to global differences, with overall patterns of behaviour needing to be considered in order to distinguish whether one was more responsible than the other.

H7 The dialogue box and spoken prompting styles will lead to longer consultations than the integral window.

As was expected, there was no significant difference between the styles of prompt in the number of airports which the customers managed to include in their itinerary (One-way ANOVA, F[2,42]: 9.027, P: 0.936) (see Table 3/3).

Measure	Dialogue Box	Integral Window	Spoken Prompt	ANOVA	P
Airports Planned	9.4	9.6	9.7	F(2,42) 0.066	0.936

Table 3/3 Global PERFORMANCE Results: Airports Planned

Although the time limit imposed on the dialogues, even after pilot testing, proved to be too short for the majority of the subjects to complete the arrangement of the itinerary; 25/45 (56%) of the trials reached the time limit. Chi^2 analysis of the 2 x 3 contingency table (Table 3/4) displaying the number of consultations of each style which reached the maximum time limit, revealed a statistically significant association between the style of prompt and the number of consultations which reached the time limit, Chi^2 (2, N=45) = 7.74, P: <0.05.

Measure	Dialogue Box	Integral Window	Spoken Prompt
Time Limit Reached	11	4	10
Itinerary Completed	4	11	5

Table 3/4 Contingency table of number of trials with each style of prompt which reached time limit.

Individual comparisons of the number of consultations reaching the time limit in the dialogue box and integral window consultations, and of those reaching the time limit in the integral window and spoken prompt consultations using Chi^2 with Yates correction, showed that there was a statistically significant association between the dialogue box and integral window styles of prompt and the number of trials which reached the time limit (Chi^2 (1, N=30) = 4.80, P: <0.05), but only a trend in the association between the integral window and spoken styles of prompt and the number of trials which reached their time limit (Chi^2 (1, N=30) = 3.57, P: <0.1).

For two reasons it was considered appropriate to perform ANOVAs on the actual time taken on consultations (see Table 3/5). Primarily this was because the time limit was the time at which the agent's interface advised them to end the consultation, and the time at which no more airports could be added to the itinerary, but not an enforced termination. This meant that consultations often extended well beyond the time limit; in two cases for over a minute beyond. Additionally, it was recognised that the critical level of Chi for significance at the 5% level with 1 degree of freedom is 3.84, so the association between the integral window and spoken styles of prompt and the number of trials reaching their time limit only narrowly missed significance when Cohen (1996) notes that some statisticians consider Yates' correction overly conservative. Evaluation of the total time taken on the consultations also gives the ability to say how long the dialogues were and how much difference in time any effect of the style of prompt lead to.

Measure	Dialogue Box	Integral Window	Spoken Prompt	ANOVA	P
Time Taken	812	721	875	F(2,42) 3.523	0.039

Table 3/5 Global PERFORMANCE Results: Time Taken

A One-way ANOVA with style of prompt as a between subject factor (3 levels: dialogue box, integral window and spoken prompts) showed that there was a significant effect of the style of prompt on the time taken on the whole consultations (One-way ANOVA, F[2,42]: 3.523, P: 0.039). Post-hoc Tukey's HSD analyses showed that the only statistically significant pairwise comparison was between the integral window and the spoken prompt (Tukey's HSD, P: 0.031), with the spoken prompt consultations being on average 154 seconds (21%) longer.

In this case, the hypothesis that there would be a significant difference between the styles of prompt in the time consultations would take was supported. The more precise hypothesis that the consultations in which the dialogue box and spoken prompts were used would be longer than those in which the integral window were also supported. As was suggested during the building of the hypothesis (Section 3.0.3.3.2) the effect of the dialogue box was less clear cut.

Although more dialogue box consultations than integral window consultations reached the time limit, and on average dialogue box consultations were longer than integral window, there was no statistical support for a difference in the time the consultations took.

There was, however, strong support for the spoken prompt consultations being longer than the integral window with the average time difference of 154 seconds (21%) even though there was only a trend in the number of consultations which reached the time limit.

3.2.2.3 PROCESS

For the same reasons it was hypothesized that there would be a significant difference in the time taken over the whole dialogues, due to the variation of the style of prompts which the agent received. It was also expected that there would be a difference in the number of turns and words over the whole dialogues.

H8 More turns will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

H9 More words will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

It is expected that the dialogue box and integral window styles of prompt will cause more disturbance to the agent's consideration of the customer's desires in planning their itinerary, and because they will therefore lose some of the common ground they have built before the prompts are introduced. Increased discussion, and therefore an increase in words and turns, will be required to regain that common ground. It could be inferred that the measures of turns and words provide some indication of the varying cognitive loads which each style of prompt give the agent when they are involved in dialogues with customers.

To determine the effect of the different styles of prompt on the global number of turns and words in the consultations, the number of both in each consultation was counted and these counts were statistically analysed. A One-way ANOVA with style as a between subjects factor (3 levels: dialogue box, integral window and spoken prompt) was performed on the global counts of turns. A Two-way ANOVA with style as a between subjects factor (3 levels: dialogue box, integral window and spoken prompt) and speaker role as a within subjects factor (2 levels:

Agent and Customer) was performed on the global counts of words. Role was not taken into consideration in the analysis of turns, because the definition of turn employed at this stage, in which agent and customer turns alternated, implied that there could be no more than a one turn difference between them. The results of the between subjects comparisons are shown in Table 3/6.

Measure	Dialogue Box	Integral Window	Spoken Prompt	ANOVA	P
Turns	202	213	242	F[2,42] 1.326	0.276
Words	1719	1613	2124	F[2,42] 4.237	0.021

Table 3/6 Global turn and word counts.

No significant effect of the style of prompt on the number of turns in their dialogues was found (One-way ANOVA, F[2,42]: 1.326, P: 0.276), although the spoken prompt lead to 20% more on average than the dialogue box, which – contrary to the specific hypothesis – lead to slightly less on average than the integral window.

On the other hand, the style of prompt did have a significant main effect on the number of words (Two-way ANOVA, F[2,42]: 4.237, P: 0.021). Post-hoc Tukey's HSD tests showed that the only pairwise comparison which showed a significant difference was between the dialogues which used the integral window and spoken prompts (Tukey's HSD, P: 0.028). The spoken prompts lead to 511 (33%) more words being used over the whole dialogues than the integral window.

There was a consistent significant effect of role on the number of words used (Two-way ANOVA, F[2,42]: 49.382, P: <0.001), with the agent consistently using more words (see Table 3/7) but because there was no interaction between the effects of style of prompt and role (Two-way ANOVA, F[2,42]: 2.104, P: 0.135). It would seem likely that this is due to the particular functions each participant has to perform during the tasks, though it was noted that there was some variation in the ratio of agent words to customer words between the styles. When the dialogue box was used, there were on average 125 agent words to every 100 customer. When the integral window was used, there were on average 163 agent words to every 100 customer, and when the spoken prompt was used, there were on average 147 agent words to every 100 customer.

Measure	Dialogue Box	Integral Window	Spoken Prompt
Agent Words	956	999	1265
Customer Words	763	614	860

Table 3/7 Summary table of mean words by role of speaker and style of prompt.

To investigate the relationship between the turns and words further, a comparison of the number of words per turn was made (See Table 3/8). This would provide a measure reminiscent of Halliday's "lexical density"; which measures the ratio of content words to grammatical words in clauses and is capable of distinguishing between spoken and written language (Halliday, 1989).

Measure	Dialogue Box	Integral Window	Spoken Prompt	ANOVA	P
Words per Turn	8.64	7.65	9.09	F[2,42] 3.359	0.044

Table 3/8 Summary table of Words per Turn by style of prompt.

A Two-way ANOVA with style of prompt as a between subjects factor (3 levels: dialogue box, integral window and spoken prompt) and role as a within subjects factor (2 levels: Agent and Customer) showed that there was a significant effect of the style of the prompt on the number of words per turn (Two-way ANOVA, F[2,42]: 3.359, P: 0.044). Post-hoc Tukey's HSD tests showed that the only significant difference was between the integral window and spoken style (Tukey's HSD, P: 0.040). There were 1.44 (19%) more words per turn in the spoken condition than in the integral window.

As with the counts of words, there was a significant effect of role on the number of words per turn (see Table 3/9) (Two-way ANOVA, F[1,42]: 49.705, P: <0.001) with the agent consistently using more words per turn than the customer, but no significant interaction between the styles and role on the number of words per turn (Two-way ANOVA, F[2,42]: 1.486, P: 0.238); which again suggests that any effect of role is caused by the different functions each role has to perform during the tasks.

Measure	Dialogue Box	Integral Window	Spoken Prompt
Agent Words per Turn	9.7	9.6	11.0
Customer Words per Turn	7.6	5.7	7.2

Table 3/9 Summary table of Agent and Customer words per turn by style of prompt.

The occurrences of overlapping speech over the whole dialogue were analysed because they have received a great deal of attention in measuring the structure of conversation and formed a major part of previous investigation of tripartite communication between two humans and a computer (see Section 3.0.3.3.4). However, no hypotheses were made about the effect of the styles of prompt on the amount and sort of overlapping speech which would occur, because it was expected that small immediate effects of the style of prompt would be drowned out by noise over the whole dialogues. In fact, no sign of an effect of the style of prompt immediately after their presentation was found (see Section 3.2.1.2), making it seem even less likely that there would be a global effect of the style of prompt on the amount of overlapping speech. A summary of the counts of the overlapping speech, indicating the percentage of total turns which these counts represent, is presented in Table 3/10.

	Agent		Customer		Both Roles	
	INTERR-UPTIONs	OVER-LAPs	INTERR-UPTIONs	OVER-LAPs	INTERR-UPTIONs	OVER-LAPs
Dialogue Box	9.2 (8.8%)	5.1 (4.8%)	10.4 (9.5%)	6.0 (6.0%)	19.6 (9.2%)	11.1 (5.4%)
Integral Window	11.3 (10.3%)	3.7 (3.5%)	10.5 (9.3%)	7.3 (6.2%)	21.8 (9.8)	11.0 (4.9%)
Spoken Prompt	13.1 (10.2%)	6.1 (4.7%)	10.9 (8.7%)	8.1 (6.4%)	24.0 (9.4%)	14.2 (5.5%)
All Styles	11.2 (9.8%)	5.0 (4.35%)	10.6 (9.2%)	7.1 (6.2%)	21.8 (9.5%)	12.1 (5.3%)

Table 3/10 Summary table of number of occasions of overlapping speech by each speaker and combined.
The percentages given for agent and customer are the number of INTERRUPTIONs and OVERLAPs as a percentage of each role's turns. The percentages give in the "Both Roles" column are of the totalled agent and customer turns.

A Two-way ANOVA with style as the between subjects factor (2 levels: open and scripted) and role (2 levels: agent or customer) and type of overlapping speech (2 levels: INTERRUPTION and OVERLAP) was performed on the raw counts of overlapping speech.

There was no main effect of style (Two-way ANOVA, F[2,42] 0.544, P: 0.584), and no main effect of role (Two-way ANOVA, F[1,42] 1.270, P: 0.265) but a main effect of the type of overlapping speech (Two-way ANOVA, F[1,42] 39.417, P: <0.001). Since there was no interaction between the style of prompt and the type of overlapping speech, but an interaction between role and the type of overlapping speech, this is likely to be an artefact of the task and the counts would suggest that this is because the customers are giving more backchannel-like responses, indicated by an increase in OVERLAPs, to the agents, who are doing the majority of the talking.

3.3 Discussion

In summary of the results, although the hypotheses made about the local effects of the styles of prompt immediately after they were presented to the agent were not supported, several of those about the global effects were. Although there was no apparent effect on the customers' PERCEPTION of their consultations, and no effect on their effectiveness in performing the tasks, there were effects of the style of prompt on both the PROCESS of the agent-customer dialogues and on the time taken to complete their consultations.

In the case of the dialogue box style of prompts, there was a significant number more consultations which reached their maximum time limit than of the integral window style, but there was neither a concomittant increase in the time taken to complete the consultations nor an effect on the number of words each style of prompt lead to. In the case of the spoken prompts, there was only a trend in the number of consultations more than the integral window which reached their maximum time limit, but there was a large significant increase in the average time taken to perform the task (154 seconds, 21%), and in the average number of words produced (511 words, 33%) which amounted to an average 1.44 (19%) increase in the number of words per turn.

3.3.1 The Thesis and the Global Findings

The large global differences in the consultations, mostly evident in the differences between the integral window and spoken prompts, validate the proposal that differences in the interface the agent is using can permeate to the dialogue between the agent and customer. The ability to capture differences due to the nature of the interaction between the agent and computer in the dialogue between agent and customer indicates that it may be possible to inform the design of those interfaces through analysis of the dialogues.

In this particular study, the major finding has been that the presentation of the same information through spoken or through written language, even though for only a brief part of the total period of the consultation, can have a considerable impact on that agent's simultaneous dialogue with a customer. The four seven-word (5 lexical item) prompts from the computer to the agent, when presented through the spoken prompt, lead to an increase of over 500 words and 2 1/2 minutes over their written presentation through the integral window.

The use of speech by the agent's computer is either producing a major disruption in the agent-customer dialogues, promoting more verbose dialogues; or possibly both.

3.3.2 Literature and the Local Findings

The findings are also somewhat controversial however. Although there is ample evidence and suggestion that the spoken prompts should lead to faster (Byblow and Corbett, 1989; Gaver, 1989; Halliday, 1989; Williams and Simpson, 1976) and more accurate (Wickens et al., 1983; Roberts and Sikora, 1997) introduction of the products to the customer, particularly since the response had to be spoken (Wickens et al., 1983), there was no clear evidence that the spoken prompt was producing a benefit in this study.

Although on average the spoken prompt lead to 25% fewer words between the prompt and the mention of the opportunity by the agent, twice as many accurate descriptions of the product and 28% more successful sales than the integral window, the variance was such that not one of these measures was statistically significant.

Similarly, although Greatbatch et al. (1993) report micro-level co-ordination of a patient's conversation with their doctor's use of a computer, this study found no difference in the way turns were managed immediately after any of the styles of prompt.

3.3.4 Conclusions and Contentions

It has to be concluded that the spoken prompt is leading to less successful dialogues, since the large amount of extra time they require leads to no corresponding increase in PERFORMANCE nor in customer satisfaction. Even if the additional speech were to build a less tangible long term benefit such as customer loyalty, the size of the increase is unlikely to be economically viable in a customer service setting where time really is money; Gray et al. (1990) quoted per annum savings of 3 million US dollars per second reduction in call time to Directory Assistance operators.

Nevertheless the pattern of findings warrants further investigation because it is not clear what influence the spoken or written nature of the prompt is having on the tripartite communication which leads to the large gross effects detected.

The possibility that immediate effects of the prompts are the sole reason for the effect seems unlikely because of the large effect it would have to explain; indeed the lack of clear evidence for effects of the prompt on the number of words between the prompt and its mention, the amount of overlapping speech during the mention and the number of successful sales would suggest that the style is having next to no immediate effect.

Instead, it seems likely that the presence of the spoken prompt is affecting the dynamic of the relationship between the agent and computer, but the exact quality of that effect is impossible to discern with the blunt tools which have been employed so far. The work of Byblow and Corbett (1989), Cotton and McCawley (1983), Halliday (1989) and Simpson et al. (1987) would suggest that the spoken prompt may lead to the computer being considered more of a commander than an advisor; with concomittant effects on the way the two interact, but further evidence will be needed to confirm this.

In the next chapter, this evidence will be sought in a micro analysis of the way the agent and customer share information, using the Conversational Games Analysis dialogue coding scheme to capture the function of every utterance in their dialogues. Because the strongest differences appeared between the spoken prompted and integral window prompted dialogues, only these two styles were subjected to the further study.

By analysing patterns of functions over the whole dialogues, an indication may be found of how the style of prompt is affecting the agent's relationship with their computer. An overall aim is to show that closer analysis of the agent-customer dialogues can provide information about the interaction between the agent and computer in a tripartite communication and potentially influence the design of the computer interface.

Chapter 4

—

Dialogue Analysis

—

Further Analysis of the Impact of Sales Prompting Styles on the Process of Communication in Simulated Customer Service Interactions with Audio Only Communication

4.0 Introduction

In the previous two chapters it has been shown that in tripartite customer service consultations both large and small variations in the design of the agent-computer interface can produce detectable differences in the dialogues between the agent and their customer. In this chapter the inverse relationship will be investigated in that the extent to which the differences detected in the dialogue can inform the design of the interface will be tested.

The potential for this was seen in the last chapter, where brief prompts being spoken lead to an increase of more than 500 words (511 words, 33%) and 2 1/2 minutes (154 seconds, 21%) over consultations where the same prompts were written in an integral window. The high cost of customer service agent time is such that such large differences are unlikely to be defensible in the light of any benefits gained.

In other cases, interface design decisions may depend on less sizable differences however. To determine whether analysis of the dialogues could be capable of informing interface design in such cases, the spoken and integral window dialogues were subjected to a finer-detailed analysis in order to test alternative explanations for the observed differences.

The aim was to show that the finer analysis of the dialogues could provide more detailed information about how the interface affected the agent-customer dialogue, and addresses directly the second contention of the thesis; that more than the typical set of human-computer interaction evaluation tools are required when the computer interface is to be involved in a tripartite communication.

Two possible explanations for the global differences were proposed in the last chapter, and they can be heuristically labelled the **disruption** and the **promotion** explanations, neither of which are mutually exclusive. Increased verbosity may have lead to **disruption**, but **disruption** could equally well occur independently of any **promotion** of dialogue.

In the **promotion** explanation, change in the style of prompt is taken to have altered the perceived role of the computer within the agent-customer interaction, which in turn alters the dynamic of the tripartite communication.

In the light of such research as that reviewed in Section 3.0.1.1, such as Cotton and McCawley's (1983, referred to in Baber et al., 1992) suggestion that too much intelligence may be attributed to machines using human speech and Halliday's (1989) argument that speech is about action, whereas written language is static, it was inferred that the style of prompt could be making the agent regard the interface with spoken prompts as a commander but regard the interface with the written integral window prompts as an adviser.

The commanding role assigned to the computer could explain an agent's increased compulsion to interact with the customer, whereas no such compulsion might be felt if the computer's assigned role was advisory.

In the **disruption** explanation, the increase in words and time would be explained by the style of prompt leading to misunderstandings between the agent and customer which required the extra dialogue to be repaired.

This might be caused by the agent regarding the interface with the spoken prompt as demanding more reaction, so that they were compelled to split their attention between monitoring and reacting to their computer and fulfilling their requirements in the route planning task with the customer. To carry the anthropomorphism already employed one step further, if there was **disruption** the agent could be perceiving the computer's role to be an overseer, whose constant monitoring and potential admonishment they were always aware of.

The split of attention such a perception would entail could in turn lead to poor construction of the agent's utterances within the agent-customer dialogue, which would require increased discourse to be repaired. For instance, if the agent and customer lost track of the flow of the ongoing route planning task when the agent introduced the sales opportunity, more discussion would be required to regain their place in the planning task once the sale had been resolved.

4.0.1 Why use dialogue coding?

The single method chosen to achieve these diverse aims was a fine-grained coding of utterances in the agent-customer dialogues which would capture how the participants used their turns at speaking.

Such a fine-grained coding was decided upon because because such analysis has proven to be useful in the investigation of other effects of technology on communication. For instance, Doherty-Sneddon et al. (1997) used a dialogue coding scheme, Conversational Games Analysis, to investigate whether a video-conferencing system could provide the same dialogue support as face-to-face vision.

The closest similar approach which has been used in studying the effect of a computer on three-way communication between two humans and a computer was in Greatbatch et al.'s (1993) ethnographic study, where it was noted that patients could time their turns at conversation to coincide with the doctors' turns entering data into their computer. This suggests that effects of an interface can be seen in the structure of dialogues between humans it supports, but is unlikely to be able to provide insight into any more than elementary faults in the design of interfaces. For instance, it might capture poor use of closure (Shneiderman, 1980), which is the design of the units into which interface tasks are divided into, if the patient's co-ordination with the doctor's entering data lead to interruption of subsequent doctor-computer tasks. It could not, however, capture the effects of varying interface designs on the way in which information is transferred.

4.0.2 Chapter Overview

The chapter will begin with an evaluation of dialogue coding schemes which have already been used, with the intention of determining their benefits and failings, and a description of the design and application of the scheme used in this study. In section two, the hypotheses which were made about the models of the dialogues provided by the coding will be described. In section three, the analysis of the coding relevant to the hypotheses will be described. In the final discussion section the interpretation of the importance of the explanations for the large differences will be built and the effectiveness of dialogue analysis in building this explanation evaluated.

4.1 Coding Schemes

The first action which needed to be taken was the selection or design of a coding scheme by which the fine details of the dialogues could be extracted. In doing this, considerable attention was paid to the warning in Bakeman and Gottman's (1997) guide-lines for developing coding schemes – that borrowing someone else's coding scheme is like wearing someone else's underwear, because the coding scheme embodies the hypotheses under investigation (p. 15). At the same time, it was recognised that, without examining the theoretical background, structure and use of those schemes which have been used by other researchers, any misconceptions or failings within those schemes would likely be recreated in a new scheme. Indeed, it was hoped that successful aspects of these precedents could be incorporated deeply into any scheme to be developed so that the amount of untried structure within the scheme was kept to a minimum.

4.1.0.1 Overly specific coding schemes

There are many ways which have been suggested for the coding of language, and each presumes some intention or theoretical background. Among the first to be rejected were schemes which were never intended to be applied to spontaneous dialogue and so were unlikely to be able to encode all the activities in dialogue. For instance Austin (1962) and Searle's (1969, 1979) Speech Act Theory was only intended to account for utterances whose production was equivalent to performing an action in the world – for instance a declaration of war – would be equivalent to landing soldiers on a foreign nation's soil, and was only intended to be applied to constructed not spontaneous utterances (Schiffrin, 1994).

Others were rejected because they had been intended for an extreme micro-analysis of a small selection of dialogues, but were unlikely to be applicable to dialogues from other situations or to a larger set of dialogues. For instance Labov and Fanshel (1977) were primarily concerned with how dialogue was used to confirm and challenge social "power" during psycho-analytical sessions, and so created a multi-level and multi-category scheme which permitted the graphical representation of every utterance's propositional, dialogue and social function, but applied this scheme to only one therapeutic session between a counsellor and client.

4.1.0.2 General Dialogue Coding Schemes

Even those schemes which were intended to be used in the coding of a collection of real dialogues can have diverse bases and intentions, and may be applied in disparate ways. The intention in the following section will be to review a selection of these schemes to determine which aspects of the schemes could be suitable for reuse, and to reject aspects which were too dependent on the scheme's theoretical background, the style of dialogue it was intended to encode, or the questions the scheme was intended to answer.

With this in mind, four aspects of the scheme were paid particular attention; its theoretical background, the procedure for applying the scheme including its reliability, the dialogues to which it had been applied, and finally the questions the scheme had been intended to answer and the results which had been gathered.

4.1.0.2.1 Theoretical Background

If any scheme's theoretical background depended on a particular theoretical background or was intended to answer a specific research question it might be less likely to shed any light in this situation. For instance, where our intention is to explain why there was a large difference in time and words between two conditions, a coding scheme whose intention was to capture the way in which speakers demonstrate and maintain their social "power" (such as Labov and Fanshel, 1977) is unlikely to be as successful as one which could capture how common ground (Clark and Brennan, 1991; see Section 1.3.1.2.2) is built.

4.1.0.2.2 Application Procedure

The main consideration in evaluating a coding scheme's application was its reliability, but this involved several factors. Primary of course was the reported reliability of the scheme; but this was considered in two lights, both how the reliability had been measured so that if its replicability seemed doubtful it would be reconsidered, and secondly how likely this reliability was to be maintained over the coding of a relatively large set of dialogues.

If the scheme included a large number of mutually exclusive categories and had only been applied to a small set of dialogues, there was some danger that over a larger set of dialogues some categories would be ignored or used to the detriment of others (Carletta, pers comm).

Similarly, Bakeman and Gottman (1997) suggest that, where multiple levels of coding are employed to reduce the overall number of categories when some are not mutually exclusive, care should be taken that the levels encode features of behaviour which are on a similar conceptual level. For instance, they comment that a coder should not be asked to decide whether an utterance is a question and a response to a previous speaker; having to consider both the purpose of the utterance on its own and its role within the dialogue simultaneously.

4.1.0.2.3 Dialogues already coded

The reason for looking at what dialogues had been coded before was to see whether there was any face-value suggestion that the coding scheme could be effectively applied to the current dialogues. The primary consideration was the situation in which the previously coded dialogues had taken place but a further, less important though still valid, consideration was the amount of dialogues the previous users had considered it suitable to apply the scheme to, in case this suggested the scheme would be unsuitable for coding such a relatively large set of dialogues.

4.1.0.2.4 Previous hypotheses and results

Finally, the questions which the scheme had been used to answer previously were considered. In this study, a scheme was required which would be likely to explain the large time and word differences between the dialogues which had used the integral window and spoken styles of prompt. If a coding scheme had only been used to describe a single class of dialogue, or had not been shown to be able to distinguish between different classes, then it would have been overly ambitious to expect it to draw out suitable distinctions in this case. This will also include description of the categories in which differences have been found previously, so some consideration can be given to whether these categories are likely to display differences in the current situation.

4.1.0.3 Alternative dialogue measures

Once the major dialogue coding schemes have been described and evaluated, a number of other measures which have been derived from dialogues will be described so that their suitability for this analysis, or the possibility of their inclusion with a future coding scheme can be assessed.

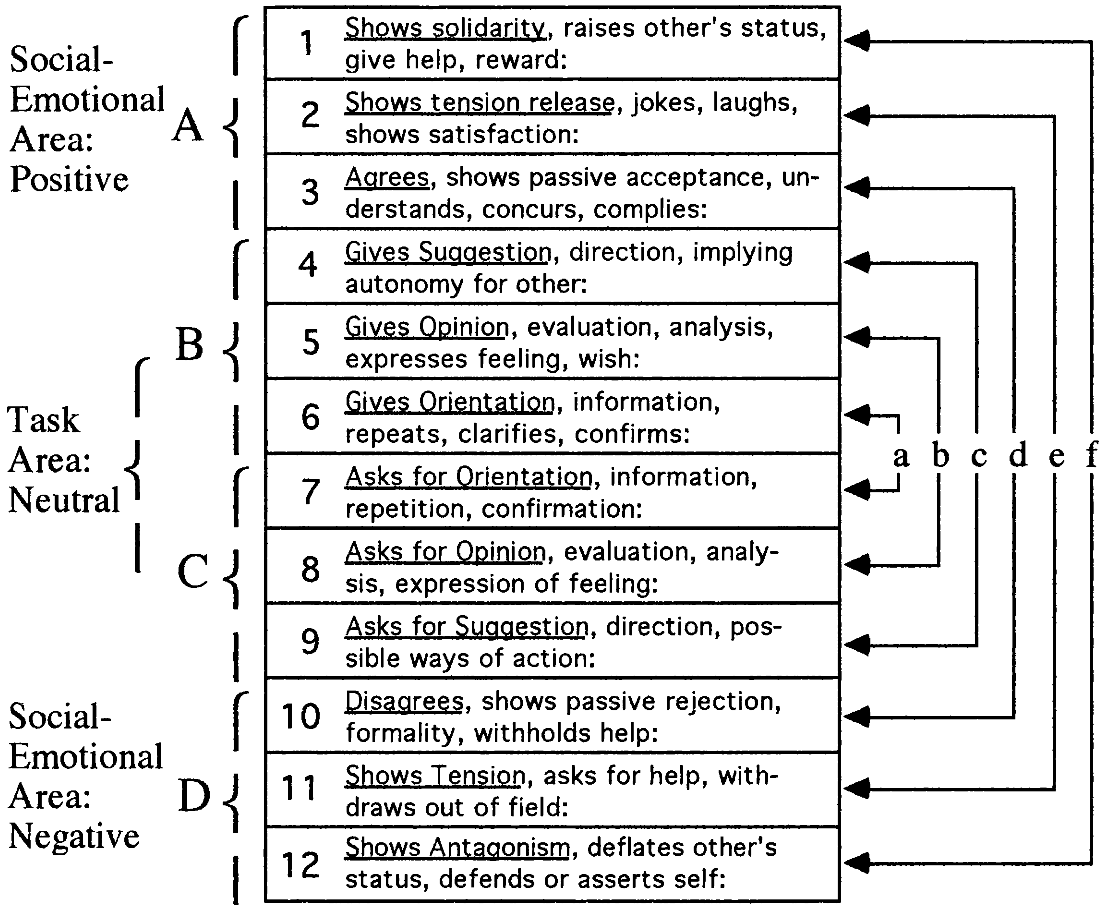
4.1.1 Four Dialogue Coding Schemes

The coding schemes evaluated were Bales (1950) Interaction Process Analysis (IPA), Morley and Stevenson's (1977) Conference Process Analysis (CPA), Olson, Olson, Carter and Storrøsten's (1992) scheme for coding design meeting activity, and Kowtko, Isard and Doherty-Sneddon's (1992) Conversational Games Analysis (CGA).

4.1.1.1 Interaction Process Analysis

Bales' (1950) Interaction Process Analysis (IPA) deserves mention because it is essentially the first scheme which was designed to code entire dialogues. Bales' intention was to build a general-purpose scheme for encoding dialogues in small groups, no matter what the make-up of the participants, their topics or the hypotheses under investigation. This scheme could then be used to specify how dialogue proceeded in all different circumstances. The scheme developed was created by an oscillation between real-time ad hoc categorisation of dialogues in various settings, with either no coding scheme or a preliminary version in place, and conflation and classification of the categories into theoretically delineated sets which were suggested by the empirical work.

The resulting scheme involved 12 categories which covered both the transfer of task-related information, both its donation and request, and the sharing of social-emotional information, whether positive or negative. The four broad classifications involved, *Positive Reactions*, *Attempted Answers*, *Questions* and *Negative Reactions*, each had three subsidiary categories (See Figure 4/1). *Positive Reactions* included utterances which SHOWED SOLIDARITY, SHOWED TENSION RELEASE and AGREED with something a partner had said. *Negative Reactions* include utterances which SHOWED ANTAGONISM, SHOWED TENSION or DISAGREED with something a partner had said. *Attempted Answers* included utterances which GAVE SUGGESTIONS, GAVE OPINIONS and GAVE ORIENTATION to things which had already been said. *Questions* included utterances which ASKED FOR A SUGGESTION, which ASKED FOR AN OPINION and which ASKED FOR ORIENTATION to things which had already been said. Utterances being classified according to how the coder feels the intended hearer of an utterance would interpret it.



Key:

A	*Positive Reactions*	a	Problems of Communication
B	*Attempted Answers*	b	Problems of Evaluation
C	*Questions*	c	Problems of Control
D	*Negative Reactions*	d	Problems of Decision
		e	Problems of Tension Reduction
		f	Problems of Reintegration

Figure 4/1 The system of categories in Interaction Process Analysis (IPA)

The unit to which these categories were applied was "*the smallest discriminable segment of verbal or nonverbal behavior (sic) to which the observer, using the present set of categories after appropriate training, can assign a classification under conditions of continuous serial scoring.*" and so the unit is defined as a psychological one rather than a temporal (i.e. defined in terms of some arbitrary time interval), transactional (i.e. defined in terms of an uninterrupted speech burst) or categorical (i.e. defined in terms of the system of categories to be used) (Morley and Stephenson, 1977). Although these units often roughly match a section of speech like a sentence, compound phrases could also be individually coded; for instance the utterance "This problem which we talked about for three hours yesterday / impresses me as very complicated / difficult / and perhaps beyond our power to solve." should be categorised as four sentences with the divisions being drawn at the diagonals.

Originally this system was applied in real time to current dialogues, partly because of the difficulties of recording during its development in the 1940s and partly because working from recordings or transcripts was expected to lead to a loss of content, and so it would be expected that the coding could be applied quickly, but Bales also notes that comparison of these initial codings with recordings was necessary to achieve good reliability. However Bales notes that long practice and frequent training was necessary to maintain consistency, and for the most part only eyeball tests of reliability seem to have been used. He does however report one study of reliability in which 2 observers independently coded a 12 minute segment of dialogue, and produced a level of reliability he considered suitable[1] and Heinicke and Bales (1953) report a Pearson's r of 0.86 for two coders scoring a number of group meetings. On the other hand, Waxler and Mishler (1966), whose primary interest was the effect of scoring of direct observations, tape recordings or typescripts on the scoring's reliability, report that there are difficulties in achieving good reliability between coders. Agreement between their coders produced a kappa of 0.619 (n=10910 [the number of units], k=2 [the number of coders]), which is outside the range Krippendorf (1980) considers suitable for drawing even tentative conclusions, and only barely in the range Landis and Koch (1977) consider "substantial". In addition to this, agreement on particular codes varied considerably, with percentage agreement at its lowest being 14% in displays of tension (category 11).

In his programmatic book on IPA, Bales (1950) suggested that the scheme could be used to find traces of differences in personality, social organisation, culture and in the type of problem and situation, and reports IPA being used for description of chess problem solving, student counselling sessions, academic thesis discussion groups, married couples' decision making, pre-schoolers' dialogues and ninth-

---

[1] Bales seems to have used an inverted Chi-square test, so considering the $\underline{P}$=0.97 his test produced for the difference between the two coders well above the 0.50 level he arbitrarily considered a suitable degree of similarity.

grade boys' discussion of two different explanations of a short story, but for only one experimental comparison; a study of non-directive and directive leadership styles in training groups.

In the study of leadership styles, the most noticeable difference was that the leader accounted for only 14% of the utterances when he was non-directive but 52% when directive. Beyond that, when non-directive, the majority of the leader's utterances were agreements (category 3) or donations of orientation (category 6) and when directive – although the most common utterance was still the donation of orientation (category 6) – the leader also gave many more suggestions (category 4) and opinions (category 5). Informally, Bales noted that there were considerable differences between the pre-schoolers' and academic discussion groups; with the academic groups involving very little negative social-emotional behaviour and much more dealing with information and analysis. Although the scheme had been applied in a number of different studies, none of the studies included many trials; often a single dialogue was coded and the most reported was apparently nine in the study of chess problem solving.

Bales' IPA can be criticised for several reasons. Firstly, there is the question of its reliability, which was noted by Waxler and Mishler (1966). Secondly, there is some doubt as to the usefulness of a general coding scheme; Bakeman and Gottman in their guide-lines on the use of coding schemes (1997) recommend that every coding scheme should be specifically designed to answer a hypothesis. On the other hand, a proven and sound general coding scheme, which could be standardised, might frequently outweigh the advantages of a less well tested one designed for a specific situation.

There are also two problems with the theoretical design of the coding scheme. A major disadvantage, noted by Morley and Stephenson (1977), is the conflation of two types of activity into one dimension of coding because no distinction is made between the function of the information being exchanged and the way in which that information is introduced. This conflation means that there are gaps in the scheme; for instance categories 3 (AGREES) and 10 (DISAGREES) do not recognise the nature of information accepted or rejected whereas categories 4 to 9 do.

Similarly, there is a conflation of task-related information and social information, leading to categories which are not mutually exclusive. Categories 1 and 12 are specifically related to the maintenance of social status relations, and categories 2 and 11 are related to the description of emotional states whereas the remaining categories encode task-related information sharing. It would be surprising to find circumstances in which an utterance did not transfer task-related information and social or emotional information simultaneously; for instance any donation of an opinion is likely to imply some transfer of social status information.

Some of these faults have been countered by variations on IPA mentioned by Waxler and Mishler (1966), such as the splitting of the coding scheme into the two instrumental and emotional dimensions, but overall this suggests that a bottom-up reconsideration of the coding scheme might be preferable

Finally, Bales' concentration was clearly on the social and cultural aspects of dialogues. How social status is maintained in dialogues was an important theoretical aspect of IPA, and Bales only mentioned noting broad differences between tasks and situations, such as that between pre-schoolers playing and academic thesis discussion groups. Because it is this sort of difference which the scheme is required to find in the current study, because of the theoretical design failings and questionable reliability, and because the scheme was primarily used for description not comparison, it was not expected that IPA would be useful in the current study.

4.1.1.2 Conference Process Analysis

The next scheme to be considered will be Morley and Stephenson's (1977) Conference Process Analysis (CPA). This is an important scheme because it attempts to resolve some of the faults in Bales IPA, and was also one of the first to be applied in determining the effects of technology on human communication. It was developed with reference to previous theory and empirical work on mother-child and patient-patient interaction, on negotiation and in clinical psychology, with particular influence being taken from Longabaugh's (1963) Resource Process Analysis (RPA); a general two dimensional category system for all human interaction which encodes *Resource*, the activity being performed, and *Mode*, how that activity was performed. Their intention was to create a coding scheme whose categories were mutually exclusive and at the same time as inclusive as possible; both faults they perceived in Bales' IPA.

The full set of categories are presented in Figure 4/2, but the following paragraphs describe the most relevant features.

Mode	*Resource*	*Referent*
1 Offer	*Structuring Activity*	0 No referent
2 Accept	1 PROCEDURE	1 Self
3 Reject	*Outcome Activity*	2 Person
4 Seek	2 Settlement Point	3 Other
	(A) INITIAL	4 Party
	(B) NEW	5 Opponent
	3 Limits	6 Both persons
	4 Positive consequences of proposed outcomes	7 Both parties
	5 Negative consequences of proposed outcomes	
	6 Other statements about outcomes	
	Acknowledgement	
	7 Acknowledgement +	
	(a) own and both sides,	
	(b) other side	
	8 Acknowledgement -	
	(a) own and both sides,	
	(b) other side	
	Other Information	
	9 Information	

Figure 4/2 Conference Process Analysis: a list of categories used (Reproduced from Morley and Stephenson, 1977)

Morley and Stephenson (1977) adapted Longabaugh's (1963) scheme to the coding of negotiation dialogues by paring the mode dimension to only those categories which were likely to occur, and specifying the categories which should be included in the resource dimension. The DEPRIVING category, and its two subcategories TAKING AWAY and WITHHOLDING, which they removed from Longabaugh's mode were included in his more general coding scheme for such cases as a mother's withdrawal of a toy and would not be expected to occur with any great frequency in a dialogue, but might need to be re-introduced in some cases. The specification of the resource was something which is necessary in any application of RPA since the definition of a resource was "anything anyone wants" (Longabaugh, 1963, p.321), though in human interaction this was expected to be things which the participants value and exchange. Their definitions of resource categories were based on having a high face-validity, balancing theoretical presuppositions and the common sense realities of negotiations.

The *Referent* dimension was added because it was expected that particular patterns of mention of your own self or group or other participants would reflect particular patterns of inter-relationship between and within the negotiating groups. For instance, if other groups were mentioned a lot, whilst mention of one's own group were small, this might reflect tension within or upon the group.

As with IPA, these categories were used to encode psychologically defined units of dialogues, which amounted to essentially a single proposition; the only difference being that Morley and Stephenson define the means to dividing transcripts into units more clearly than Bales. It was presumed in the case of CPA that coding would be done from tape and transcript rather than at the time of the dialogues, so the time to code dialogues was expected to be longer than the use of IPA (it was estimated at a full day per thirty-five minute transcript) but was also expected to lead to a better inter-rater reliability. The most dialogues the scheme was applied to in any one study was 24 (Stephenson, Ayling and Rutter's study, 1976).

Inter-rater reliability in coding appeared high, with typically over 90% reliability in the division of dialogues into units, over 90% agreement on the mode, between 67% and 96% agreement on resource and typically around 90% agreement on referent. The measures of reliability they mentioned included intra-rater reliability where one of the developers recoded a transcript after a passage of time, comparisons of the developers and trained coders, and comparisons of expert and trained coders including intra-rater reliability checks of the trained coders. The description of the coding scheme included an algorithm by which individual and it is likely this considerably aided the level of reliability.

Morley and Stephenson report that CPA had been used in the coding of several variations of union negotiations conducted either face-to-face or over the telephone. In their own work, the coding scheme only found that the weaker, management, side made fewer references to their opponents when they were face-to-face. This was not a failing in the category system per se however, because Stephenson, Ayling and Rutter (1976) found many more differences when they matched subjects' personal interests with the brief they were given. The majority of these differences involved the referent dimension, for instance there was less praise for the opponent and more in party references when the negotiations were audio-only, with the only differences which did not involve the referent dimension was the amount of offers and "information" acts in both case there being more by management face-to-face and more by union audio-only. In neither of the described applications of the coding was the scheme effective at confirming the experimenters' hypotheses, but Stephenson, Ayling and Rutter were able to use the observed differences to build a model of the effects which were leading to the differences; concluding that there was an interaction between a party's strength and conviction of case and intersocial factors with the medium a negotiation was taking place over, with intersocial factors being more important face-to-face and strength and conviction of case being more important audio-only.

Conference Process Analysis seems much more attractive than Interaction Process Analysis, but still has some flaws. Morley and Stephenson admit that the design was programmatic, but the programme does not seem to have progressed beyond their analyses of negotiations and it would seem likely that to cover other tasks several changes would need to be made. For instance, the lack of a command category among the possible mode categories would be a basic addition which would need to be made in any task which required one participant to perform an action for another. Although the scheme was designed to facilitate such adaptation, it is impossible to predict how much change would be required to fit the scheme to a new task. Since the travel game used in the current study involved both co-operation as the agent built an itinerary for their customer and negotiation as the agent tried to sell products in return a considerable amount of development would be predicted.

The strongest point of the scheme was the inclusion of the *Referent* dimension which was instrumental in capturing the majority of the differences in the studies Morley and Stephenson reported. This is unlikely to be useful in evaluating the current studies, however, because in Morley and Stephenson's studies the dimension was supposedly capturing inter- and intra-party tension which was particularly evident when the participants had long-term personal interest in the briefs they were arguing. In the current studies, although the participants were given some vested interest by the prizes available to them, this was at best short term and they were also aware that a reasonable amount was available to them just for taking part. In addition to this, any interpersonal tension was most likely to be displayed during the negotiation sections of the dialogue, during which no significant differences appeared after the initial analyses described in the last chapter.

Although CPA is attractive, the possibility that it would require considerable adaptation to be applied to the current set of dialogues, and doubt that the strong *Referent* dimension would be useful in the evaluation of the current set of dialogues, means that other schemes may be more suitable.

4.1.1.3 Design Meeting Activity

A more recently developed scheme which has been used for evaluating the effects of technology on dialogues is that developed by Judith and Gary Olson (Olson, Olson, Carter and Storrøsten, 1992; Olson, Olson and Meader, 1995) for categorising utterances in design meetings. This is based on previous schemes developed in management studies for improving the design process and for characterising group management processes. As such, its application to the answering of research questions seems to have been subsidiary to description of design meetings, and this leaves it open to the same criticisms, with caveats, levelled at Bales IPA.

The core coding scheme involved two dimensions, the first containing 10 categories which captured the type of activity performed with the utterance and the second was binary and captured whether the utterance was a CLARIFICATION or STATEMENT; where clarifications were questions and answers which cleared up others' misunderstandings of the initial presentation of subject matter.

The 10 types of activity were ISSUES, ALTERNATIVES, CRITERION, PROJECT MANAGEMENT, MEETING MANAGEMENT, SUMMARY, DIGRESSION, GOAL, WALKTHROUGH and a catch-all OTHER. The first three of these were intended to capture the presumed pattern of exchanges in the solution of design problems in which an issue would be raised to which alternatives would be presented, and criteria for selecting each given until a solution was selected. The rest were added to capture events which had been observed during design meetings while the scheme was being developed. Project and meeting management was captured in the MEETING MANAGEMENT, PROJECT MANAGEMENT and statement of GOALS. SUMMARIES and WALKTHROUGHS are two ways in which the current state of the product or design could be presented, the former being a list-like description and the latter being a step-by-step description of how the product would be used. The final two of these 10 categories of activity, DIGRESSIONS and OTHER. DIGRESSIONS captured jokes or discussion of side topics, and OTHERS were used to capture utterances which did not fit into any of the other codes. Initially time during which there was no discussion, for instance while answers were being looked up, was included in the OTHER category, but later a category called a PAUSE was included for this purpose.

When the coding scheme was applied to technologically mediated discussions involving a shared text editor, the core 22 category scheme was extended with 4 categories which accounted for the planning and writing of the document, PLAN and WRITE, and discussion of the technology, TECHNOLOGY CONFUSION and TECHNOLOGY MANAGEMENT.

There were two further categories which only existed as clarifications; CLARIFICATION ARTIFACT and CLARIFICATION GENERAL. The former were utterances which explained a drawing or other artifact, the latter were clarifications not covered by any other categories.

The scheme was intended to be applied to video and audio tape recordings of design meetings which lasted up to 1 1/2 hours, was applied to around 74 dialogues during one study, and during its development an inter-rater check between three coders produced correlations of between 0.68 and 0.99. One unusual feature in the Olsons' application of their scheme is that they not only look at the amount of time spent on each activity, but also at how transitions were made between categories, allowing some measure of how the dialogues were structured. Another unusual feature is that, although the categories are psychologically based, according to the definition provided by Morley and Stephenson (1977) and described in the discussion of Bales' IPA, the Olsons consider not only the amount of occurrences of each category but also the time spent on them.

The scheme has been applied both in the description of real world design meetings and experimental comparisons of design meetings performed either face-to-face or remotely, over audio or video links, and either using a computer-based text editor which supported all participants to work on the same document simultaneously or when face-to-face using whiteboard, paper and pencil.

In the studies of real design meetings, it was found that there was little difference between such meetings on the basis of the company at which they took place, though the two companies they looked at claimed very different methods, nor on the basis of the projects the meetings were discussing.

In the studies of real design meetings, they found that there was a clustering of activities into the three primary design activities of ISSUE, ALTERNATIVE and CRITERIA and into the other 9 categories because there were far more transitions within these two groups than between them. In their study of technologically mediated design meetings, remote meetings were found to spend more time in MEETING MANAGEMENT and in CLARIFICATION than the face-to-face which used the shared editor, and the video group was found to spend less time stating and clarifying ISSUES than the audio only.

The main reason for rejecting this scheme is that it was very clearly designed for the encoding of design meetings and is probably tailored too specifically to that task. The scheme has not been applied to anything but design meetings and it would seem likely that any such application would require require some re-definition of the core categories or the inclusion of more categories. Since even the core scheme was somewhat category heavy with 22 categories – 10 STATEMENT and 12 CLARIFICATION – the inclusion of more categories would seem undesirable, and the amount and definition of new categories would have to be carefully considered because the mere addition of the shared editor and remote meetings lead to the fairly unprincipled inclusion of four new categories of the type of activity. If a re-definition were to be considered, this would inevitably require the rebuilding of the ISSUE, ALTERNATIVE and CRITERION categories but since these were strongly inspired by previous work on design rationale, this would seem no more desirable than trying to include new categories.

4.1.1.4 Conversational Games Analysis

The final scheme which will be discussed is Conversational Games Analysis (Kowtko, Isard, Doherty-Sneddon, 1991; Kowtko, 1997; Carletta et al., 1996), which was designed from the first to be a general scheme for coding dialogues; "*It was intended to be a system which could represent conversational activity in an independent manner, free from the influence of any potential application*" (Kowtko, 1997; p.51). This alone sets it apart from the Olsons' scheme, but the effect is even more visible in the classification of categories within it and the structure of encoded dialogues which follows. Whereas the Olsons' scheme was almost entirely concerned with what is spoken about in the utterance, essentially Longabaugh's (1963) and Morley and Stephenson's (1977) resource dimension, Conversational Games Analysis is almost entirely concerned with how that information is shared[2], which is more closely related to Longabaugh's (1963) and Morley and Stephenson's (1977) mode dimension. An underlying, and occasionally explicit, aim seems to be capturing the means by which common ground on any discussed information is established. Even were it to have been

---

[2] The ALIGN, CHECK and CLARIFY moves (see below) implicitly recognise that there is content because the utterances they encode must refer to information which has already been presented in the dialogues, but there is no consideration of what that information might be.

applied to a limited range of dialogues, this would suggest that it should be context free and that its more general application would be possible.

At the heart of the scheme was the assumption that dialogues are not sequences of speakers presenting information to be interpreted by their hearers, but interactions between the participants in a dialogue trying to understand each other and make each other aware of that understanding. This is probably because the scheme was derived not from Sociological or Management Studies perceptions of dialogues, but from the fields of Artificial Intelligence and Psychology. It owed a considerable debt to Sinclair and Coulthard's (1975) Dialogue Analysis coding scheme for the recognition that an utterance's discourse function often differs from its syntactic form, the recognition of layers of dialogue and the suggestion of a three-step rather than two-step structure to individual interactions, but its closest ancestor was the work of Power (1979) and Houghton and Isard (1987) on computational models of software robots co-ordinating to achieve goals, for instance one trying to open a door when another had the key. Beyond these two, however, there is also strong influence from current psychological models of dialogue mentioned in Section 1.3.1.2 particularly visible in the ACKNOWLEDGEMENT category which explicitly captures three of Clark and Schaefers' (1989) five indicators that an utterance has been understood and common ground achieved.

The primary difference this made in the encoding of dialogues was that whereas the other schemes described here would lead to dialogues being encoded as linear strings of the schemes' categories, Conversational Games Analysis implies more of a syntactic structure to the dialogues. It recognises three levels of dialogue; the *transactional* level, the *conversational games* level and the *conversational move* level. *Transactions* are high level representations of dialogues, only rarely included when dialogues are coded, described as "*subdialogues that accomplish one major step in the participants' plan for accomplishing the task.*" (Carletta et al., 1996, p.1). *Conversational games* are explicitly represented in the scheme and encoded the sharing of one piece of information, but can include any number of utterances by any participants in the dialogue. For this reason, the *conversational moves* which encode individual utterances, are divided into two sets; *initiating moves* which "set up the expectation of a response" (Carletta et al., 1996, p.2) and often "introduce a new discourse purpose into the dialogue" (Carletta et al., 1996, p.2), and *response moves* which fulfil an *initiating move*'s expectations. In this way, *conversational games* resemble Clark and Schaefer's (1987) "contributions" (see Section 1.3.1.2.2) since they include not only the presentation of information but also reaction to it, and reflects Schegloff and Sacks (1973) concept of the adjacency pair, which is that neighbouring pairs of conversational turns are typically related.

Games / Initiating Moves		**Response and Feedback Moves**	
INSTRUCT	Communicates a direct or indirect request or instruction, to be done immediately or shortly, e.g. "You then go down two inches."	CLARIFY	Clarifies or rephrases what has previously been said; usually repeats given or known information; elicited by other person. e.g. "South, two inches."
CHECK	Checks self-understanding of a previous message or instruction by requesting confirmation directly or indirectly; makes sure that a complicated instruction is understood. e.g. "So you want me to go down two inches?"	REPLY-Y REPLY-N	Affirmative (REPLY-Y) or negative (REPLY-N), elicited response to QUERY-YN, CHECK, or ALIGN; also indicates agreement, disagreement, or denial; e.g. "Yes, I have."
QUERY-YN QUERY-W	Yes-No question (QUERY-YN) and open-answer Wh-question (QUERY-W); asks for new or unknown detail about some part of task; does not request clarification about instructions (that would be CHECK); e.g. "Do you have a rockfall?"	REPLY-W	An elicited reply to QUERY-W or check; can be response to QUERY-YN that is not easily categorisable as positive or negative (REPLY-Y/N). e.g. "Down"
EXPLAIN	Describes status quo or position in task with respect to the goal; freely offered, not elicited; provides new information. e.g. "I've got a cairn."	ACKNOWL-EDGEMENT	Vocal acknowledgement of having heard and understood; not specifically elicited but often expected before the speaker will continue; announces readiness to hear next move -- in essence a request to 'please continue'; may close a *game*. e.g. "Alright" or "Oh right, I see what you mean."
ALIGN	Checks the other participant's understanding or accomplishment of goal; elicits a positive response which closes a larger *game*; checks alignment of both participants' plans or position in task with respect to goal; checks attention, agreement, or readiness. e.g. "OK?", meaning *Are_you_with_me*?	READY	Indicates intention to begin a new *game* and focuses attention on oneself, in preparation for the new move; an acknowledgement that the previous *game* has just been completed, or leaving the previous level or *game*; consists of a cue-word. e.g. "Now" or "Right."

Figure 4/3 The Conversational Games Analysis categories, and their definitions.
(Reproduced from Kowtko et al., 1992).

The original definition of the categories is presented in Figure 4/3, but more recently it has been recognised that the READY is a unique *move* type which has been classed a *preparation move*. Carletta et al. (1996) provide an algorithm for the coding of utterances which recognises the unique nature of READYs and is reproduced in Figure 4/4.

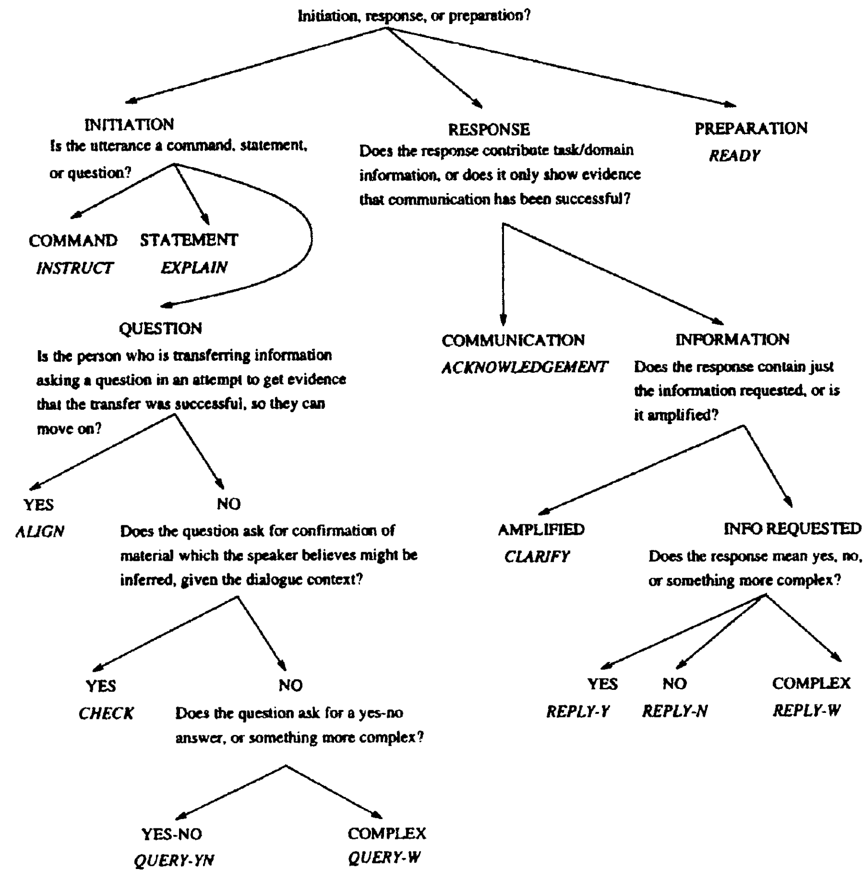


Figure 4/4 Algorithm for the coding of Conversational Games Analysis *moves* (reproduced from Carletta et al., 1996).

A representative *conversational game* might include a question (QUERY-W) followed by an answer (REPLY-W) or a command (INSTRUCT) followed by its acknowledgement (ACKNOWLEDGEMENT), but might equally have a further *conversational game* embedded within it if some discussion of a preceding *move* was deemed necessary. For an example of the latter, a statement of a piece of information (EXPLAIN) might require the hearer to question that they have fully understood it (CHECK) and some further explanation by the original speaker (CLARIFY) before the hearer shows their understanding (ACKNOWLEDGEMENT).

Only two potential additions to the scheme have been suggested. Newlands (1998) suggested the DIRECTIVE *initiating move*, which is similar to an INSTRUCT but does not demand a physical action of the hearer, instead more implicitly requiring an action while describing the speakers' desires. The

DIRECTIVE accounted for only around 1.5% of the *initiatingmoves* in Newlands' study and because of the functional similarity to the INSTRUCT is unlikely to be included in future versions of the coding scheme. The second potential addition was the OBJECT *responsemove* which is reported by Carletta et al. (1996) to have been suggested in Grice and Savino's (1995) coding of familiar Italian speakers performing the map task. This is a form of the ACKNOWLEDGE which has a negative slant similar to Bales' category 10. In Grice and Savino's study, it was used to indicate that the previous speaker's *conversational game* was failing to resolve a *transaction* and often preceded the initiation of a *game* trying to resolve the *transaction* in a different fashion. This function may be language specific.

This coding scheme has been applied to transcripts combined with audio- and video-tape recordings of dialogues, because function could not be captured without reference to the recordings, but no estimate of the time to code a dialogue has been reported. A day per 15 minute dialogue would be expected (Newlands, pers comm). The scheme has been applied extensively, most notably in the coding of the map task corpus (Anderson et al., 1991) of 128 dialogues in which one participant had to direct another around a series of landmarks on a map in various circumstances such as with the participants being friends or not or being able to see each other or being separated by a screen (Doherty-Sneddon et al., 1997). More recently, it has been used by Veinott, Olson and Olson (1999) to compare native and non-native speakers' performance of the map task. To assess the scheme's reliability in situations other than the map task, it has been applied to maze task dialogues, in which two participants had to co-ordinate to manoeuvre their individual tokens through a maze of cells with locked doors (Kowtko, Isard and Doherty-Sneddon, 1992), and to a transcribed conversation between a hi-fi sales assistant and a married couple (Carletta et al., 1997). Of particular relevance to the schemes' application in the current study, the scheme has already been applied in coding travel game dialogues (Newlands, 1998).

The reliability in all these cases was good. In the coding of the maze task dialogues, a percentage agreement of 85% in the coding of *moves* by category was reported (Kowtko et al., 1992). Coding the map task dialogue, four coders including one of the schemes' developer reached K=0.92 (N=4079, k=4) in the coding of *move* boundaries and K=0.89 (N=563, k=4) in the categorisation of *moves* (Carletta et al., 1997). Coding the hi-fi sales conversation, two coders reached K=0.95 (N=819, k=2) in the coding of *move* boundaries and K=0.81 (N=80, k=2) in the categorisation of *moves* (Carletta et al., 1997). In the coding of the travel games, two coders reached a K=0.94 (N=177, k=2) in the categorisation of *moves* (Newlands, 1998).

Statistical analysis of coding for deriving conclusions has concentrated on the *conversational games*[3] and *initiating moves* because the majority of *response moves* are not independent of their *game*'s *initiating move*. The coding of the map task showed that more ALIGN *games* were used when participants were separated by screens which prevented them seeing each other, with the largest part of these being produced by the direction giver. In Newlands' (1998) coding of the travel games, in which the travel agent and customer communicated either face-to-face or through a video link, showed differences in the number of EXPLAIN, QUERY-YN and ALIGN *moves* produced during the dialogues using each medium. When the participants were using the video link they produced more EXPLAIN *moves*, more QUERY-YN *moves* and more ALIGN *moves*, with the agent being responsible for the majority of the difference in EXPLAINs and ALIGNs while the customer was primarily responsible for the difference in QUERY-YNs.

The main query to be raised about Conversational Games Analysis is that its categories do not consider the content of utterances, only the way in which the content is shared. This may make make it more context-free than other coding schemes, but potentially means that the power to capture specific variations in specific dialogues is lost.

At the same time, evaluation of the schemes which have included measures of what is being spoken about, show that feature to be particularly likely to add confusion and uncertainty to the coding. The Olsons' scheme includes a lot of categories, and some are recognised by the authors to have been included only to fill in gaps in the definition of the more important categories. Morley and Stephenson's resource dimension contains more categories than any other, and the authors suggest ways in which it could be expanded further. The coding of Morley and Stephenson's resource dimension also seems to have debatable reliability since the percentage agreement between coders they report varies considerably and dips well below the agreement on the mode and referent dimensions.

[3] The category of a conversational game is inherited from the category of its first initiating move.

The benefits of Conversational Games Analysis far outweigh this disadvantage however. Its concentration on how topics are spoken about, not what those topics are, means that it has a close link to universal context free aspects of language production. The scheme is closely modelled on well respected models of dialogue and linguistics, such as Clark's common ground model (Clark and Brennan, 1991; see Section 1.3.1.2.2), Power's (1979) AI model of co-operative dialogue, and through Sinclair and Coulthard (1975) has links to Halliday's systemic functional linguistics, which makes it a simple matter to convert hypotheses about how variations in the situation of dialogues will affect their process into hypotheses about specific patterns of categories. Because the scheme is closely derived from strong models of dialogue, this has the added benefit that – unlike the other schemes mentioned – its description is not programmatic; basic research on communication has been applied in the definition of a structurally simple coding scheme. The scheme has a proven reliability, and the relatively small number of categories suggest that there will not be neglect of categories in coding a relatively large corpus; a suggestion which is empirically supported by the coding of the 128 map task dialogues. Finally, the scheme has been proven in the finding of useful statistical differences between controlled variations in the setting of dialogues.

For these reasons, it was decided to break with the strongest interpretation of Bakeman and Gottman's (1997) instruction not to reuse someone else's coding scheme, and use Conversational Games Analysis in its entirety. It was, however, also considered appropriate to consider how Conversational Games Analysis could be made more specifically targeted to the current study without losing its strengths and the development of these specifications are given in Section 4.1.3.

4.1.2 Alternative dialogue measures

This last subsection covers *backchannels* (Yngve, 1970), *breakdown* (Wright and Monk, 1989) and the LOGOS tool for evaluating the emotional tone of texts (Anderson and McMaster, 1986). These three are united in the fact that they work by selecting only particular features of the dialogues, typically words or non-lexical productions which have specific meaning or theoretical association to them, but the reason for including them is varied. *Backchannels* are included because they have become a staple means of measuring the process of dialogues, because they are taken to represent attempts by a listener to show that common ground is being maintained. *Breakdown* is included because it specifically addresses the issue of the technology used by a group of human conversants imposing itself in their dialogue; it will provide a basis for the additions to the Conversational Games Analysis scheme which are described in Section 4.1.3.1. LOGOS is included because emotional tone provides a measure of process far different from anything previously described. It has never been applied to spontaneous dialogues before, however, and although it was a relatively simple matter to adapt transcripts to be analysed by the LOGOS software, it was considered necessary to evaluate the system's usefulness in dialogue analysis before applying it more generally. Section 4.1.2.3, in which LOGOS is described and evaluated, is therefore considerably longer than the other two subsections, but performs a function more like the evaluations of the coding schemes in Section 4.1.1.

4.1.2.1 Backchannels

Backchannels (Yngve, 1970) are short feedback utterances which show that the producer is following, understands and often agrees with the primary speaker. They are typically non-lexical, "mhm" or "uh huh", but the definition is often stretched to weakened lexical items such as "yeah", "okay", "right" or "yes", and do not involve any effort to take the floor from the primary speaker. They are recognised by Clark and Brennan (1991) as the most obvious form of recognition that common ground has been established by the primary speaker.

O'Connaill, Whittaker and Wilbur (1993) have applied these as a tool for measuring the difference between two forms of video-mediated communication one of which included a time lag with face to face communication. They hypothesized that the lag would lead to fewer *backchannels* but found that as well as this being true, both video conditions lead to fewer than the face-to-face meetings.

For the current study such a measure is unlikely to be useful, however. Beyond the caution that must be employed because the definition of a *backchannel* is not rock solid, "yes" as a response to a question being difficult to distinguish from "yes" as a *backchannel* (Anderson, pers comm), there is no reason to expect that such a measure would have been influenced by the variation of the agents-computer interface being investigated in the current study.

It is also likely that a stronger related measure would be found in an analysis of the ACKNOWLEDGEMENT Conversational Games Analysis *move* because this represent not just part of one of Clark and Brennan's means to show grounding is being achieved, but fully encompasses three (Carletta et al., 1996).

At the same time, the use of *backchannels*, ACKNOWLEDGEMENTs or any of Clark and Brennan's positive forms of evidence for grounding should be used cautiously because they cannot be fully independent of the utterances whose grounding they are supporting. For instance, a large amount of positive evidence could accompany a large amount of well constructed utterances or a large amount of effort to overcome poorly constructed utterances. Positive evidence may be useful to show where verbal means of grounding are replacing non-verbal evidence, such as a *backchannel* being used in place of a nod in O'Connaill, Whittaker and Wilbur's (1993) study, but even then inferences must only be made in the light of how the discussions have otherwise changed.

4.1.2.2 Breakdown

Breakdown (Wright and Monk, 1989) is the entering of the technologies which are being used as a medium of communication into the communication itself. This is based on the contention by Winograd and Flores (1987) that transparency is a desirable feature in a human-computer interface, and that any recognition of the existence of the system implies a failure.

McCarthy and Monk (1994) contended that such *breakdown* during the use of a text-based computer-mediated communication system would be indicated by the participants mentioning the system explicitly in their dialogues. They found, however, that there were too few occurrences of *breakdowns* to permit their analysis.

Were there to be enough mentions of the technology, this might seem to be a valid measure to apply in the analysis of the process of the dialogues in the current study, because the variation in the agent-computer interface could lead to more mentions in one case than another. For instance, both the dialogue box or spoken prompting style was expected to persuade the agent to mention the opportunity before considering its exact message, and so might be expected to lead to the presentation of the prompt itself being included in the mention of the opportunity to the customer. On the other hand, no differences in the mention of the prompt were found in the initial analysis of the dialogues described in the last chapter and – although this may vary with the technology used – McCarthy and Monk's finding that there were too few occurrences to be useful must be considered.

4.1.2.3 LOGOS

The final of these measures is the LOGOS measure of the emotional tone of language developed by Cliff Anderson (Anderson and McMaster, 1986). LOGOS is a piece of software which searches a text for occurrences of words in its database, all of which are rated for their own emotional tone, and then produces a composite measure of the emotional tone of the text by averaging the ratings of the emotional tone of the individual words. For the most part this is entirely automatic, but if the database finds a known homophone or polyseme the operator is prompted to select which meaning of the word is the correct one.

The words in the database are the 1000 most frequent content words, and their emotional tone is represented by their scores on Osgood's (1962) three primary semantical differential dimensions which were rigourously measured by Heise (1965). These three dimensions are those of *Evaluation*, *Activity* and *Potency*. Evaluation represents the word's pleasantness, is measured by a scale such as Good-Bad and words such as "beauty" and "pleasant" are high in it whereas "fire" and "disease" are low. Activity represents the word's arousal, is measured by a scale such as Active-Passive and words such as "fire" and "attack" are high in it whereas "dead" and "rock" are low. Potency represents the word's dominance, is measured by a scale such as Strong-Weak, and words such as "steel" and "rock" are high in it whereas "kiss" and "baby" are low.

According to whether the combination of scores on these scales were high, low or medium, the words, and hence texts, can also be assigned a categorical emotional state. For instance in the eight state classification, which only takes account of whether scores were high or low, a word which was rated low in evaluation, low in activity and low in potency would be considered "depressed" whereas a word which was high in evaluation, low in activity and high in potency would be rated "relaxed".

Although this technique has not been applied to dialogues, it has been applied fairly extensively in analysing written texts. Anderson and McMaster (1982a, 1982b) validated the system by using it to assess the emotional tone of college students writing about the best and worst possible progression of their careers. In both cases, they found good inter-rater reliability and that the descriptions of the best progressions were higher in evaluation and activity than the descriptions of the worst. In one they found the best progression being lower in potency.

Most often LOGOS has been applied to literary works. Anderson and McMaster (1987) used LOGOS to chart the development of the Grimm Brothers' stories and compare them to critics' judgements. Anderson and McMaster (1992) report ratings of Dylan Thomas' poems which reflects closely literary criticism of the same poems. In the same paper, Anderson and McMaster (1992) also report charting the emotional progress of a single story by rating 100 word blocks in sequence.

It was hoped that the LOGOS system could be used to find any effect on the emotional content of dialogues occurring under different conditions; for instance, a medium of communication which impeded the dialogue could be expected to lead to lower ratings of pleasantness or activity than a better medium. Before adding the system to any battery of tests, it was felt necessary to evaluate whether the technique was suitable to be applied to dialogues at all. In particular, it was necessary to determine whether the 1000 words LOGOS searched for occurred frequently enough in dialogue to permit valid measurements, and whether the emotional tone was likely to vary as much in spontaneous dialogue as in more carefully crafted written texts.

To determine the answer to these two questions, LOGOS was applied to 4 dialogues generated at British Telecommunications in the development of the Customer Interface Quality Measurement (CIQM) system, which was intended for real time evaluation of the success of agents' dialogues with their customers (Marle and Coyle, 1995). Two of these dialogues represented the worst recognised features of customer service calls; the other two, the best and these representations were supported by experimentally gathered ratings using the CIQM system.

LOGOS hit an average of 23.7% of the words in the three dialogues, which compares favourably with the number of hits are reported in the literature; for instance 19.5% in Anderson and McMaster's (1992) study of poetry and 21.8% in Anderson and McMaster's (1993) study of children's stories. Caution dictated that the hits should be further investigated to determine the amount of words which LOGOS hit which were homophones or polysemes whose definition in LOGOS dictionary did not match their definition in the dialogue. For instance, the word "right" had four definitions in the LOGOS dictionary, but none of the them matched its use as an acknowledgement in the dialogue. Of the 82 hits in one of the dialogues, 28 were found to be homophones or polysemes, and of these only 8 were found to have definitions in the LOGOS dictionary which did not match their use in the dialogues, meaning that 90.2% of the hits were on words which matched the LOGOS dictionary definition. Therefore LOGOS hit 21.4% of the words in the total dialogue accurately, which again compares favourably with the number of hits which are reported to have been used for deriving conclusions in the literature.

To determine whether there was enough emotion expressed in the dialogues for LOGOS to find useful variation, the measurements of emotional tone which LOGOS produced for the CIQM dialogues were also evaluated. To aid this evaluation, the 27 state labels which have been assigned to each combination of low, medium or high scores in each of the semantic differential dimensions were used to give a heuristic for the emotional state in the dialogues. The 27 state label system was used because the 8 state system only considers whether the scores on each dimension were high or low and tends to accentuate differences which in the case of the dialogues were expected to be small.

Dialogue	State Breakdown (27 state system)			State Label
	Evaluation	*Activity*	*Potency*	
GOOD 1	Low	Med	Med	Unpleasurable
GOOD 2	Low	Med	Med	Unpleasurable
BAD 1	Med	Low	Med	Nonaroused
BAD 2	Low	High	High	Hostile

Table 4/1 State breakdown and labels for the four CIQM dialogues.

As Table 4/1 shows, the majority of the scores on the dimensions were indeed in the medium ranges; which not only suggests that use of the 8 state heuristic labels might inflate differences between dialogues, but also that there may only be small differences in emotional tone in spontaneous dialogue. This is further supported by the finding that, although the dialogues were supposed to represent the extreme qualities of customer service calls, only one "Bad" dialogue showed any sign of extreme emotion. The findings also show that the relation of the labels LOGOS assigns to the actual quality of the dialogues is somewhat dangerous because the two "Good" dialogues are both labelled "unpleasurable".

Given the findings about the rated emotional tone of even supposedly polarised spontaneous dialogues, the use of LOGOS in dialogue analysis was not pursued further.

4.1.3 The Developed Coding Scheme

As was explained in the discussions of the coding schemes in Section 4.1.1, it was decided to use Conversational Games Analysis for the analysis of dialogues in the current study. This was because the other schemes reviewed could require considerable redesign to be applicable to the specific dialogues under investigation, because Conversational Games Analysis was intricately bound to well respected general models of communication, and because of its proven capability as an experimental tool.

At the same time, it was considered appropriate to take note of the theoretical reasoning of Longabaugh (1963) and Morley and Stephenson (1977) – that the information being passed between participants in a dialogue was a fundamental part of them, the practical suggestions of Bakeman and Gottman (1997) – that coding schemes should be tailored to the dialogues they are coding, and personal observation of real customer service calls and discussion with customer service centre management – which high-lighted their interests and points where those interests coincided with methods suggested by other researchers.

To this end, a second dimension was added to Conversational Games Analysis which encoded the *task orientation* of the *conversational moves* and *games*. It was expected that, by recognising the task of an utterance in this second dimension, some further analysis of aspects particular to the dialogues' situation could be made without damaging the core benefits of Conversational Games Analysis.

4.1.3.1 The Task Orientation Dimension

The aim in designing this second, *Task Orientation*, dimension was to find out how the variation in the agent's software affected that software's perceived role in the dialogues. It was intended to do this by encoding the degree to which the presence of the software imposed itself either directly or indirectly in the dialogues between the agent and customer.

This was achieved by marking any *conversational moves* which were not directly related to the performance of the task, but made some reference either directly or indirectly to the presence of the tools the participants were using or to physical activities they had to perform. As a set these *moves* were referred to as OFF-TASK, but six subcategories were recognised; labelled ABOUT_TECH, ABOUT_SELF, ABOUT_PARTNER, ABOUT_LANGUAGE, TO_SELF and INTERJECTION whose definitions are provided in Figure 4/5 and will be further described in the following paragraphs.

OFF-TASK Category Label	Description
ABOUT_TECH	Comments on the technology which is being used. *The button doesn't seem to work.*
ABOUT_SELF	The speaker describes their own actions. Typically used to explain why there will be a period of silence, to prevent the partner interrupting. Usually EXPLAIN moves. *I'll just look that up for you.*
ABOUT_PARTNER	The speaker says something about their partner's actions. For instance, noting something the partner is or should be doing. Often QUERY moves. *Are you looking that up?*
ABOUT_LANGUAGE	Comments on the language being used, often produced when the pronunciation of a destination is unknown. *Is that Elly or Ealy?*
TO_SELF	Utterances which are not for the benefit of the speaker's partner. Usually remind the speaker of something which has been agreed with the partner. Often sotto voce.
INTERJECTION	A humorous observation. *The travel pass would be very useful.* *What like a spanner?*

Figure 4/5 Definitions of OFF-TASK *move* categories

ABOUT_TECH OFF-TASK *moves* bear a close relationship to Wright and Monk's (1989) *breakdowns.* They encode occasions on which the tools available to the participants are mentioned directly in the dialogues. In the agent's case this typically involves mention of their software, in the customer's it may include mention of the map or other resources available to them but also included reciprocal comments on the agent's software.

ABOUT_SELF and ABOUT_PARTNER typically relayed information about where the speaker's or the partner's attention was focused. These *moves* typically let the other participant know that the speaker was involved in some activity or what that activity was. For instance, an explicit ABOUT_SELF *move* by the agent would let the customer know that they were about to look something up. A less explicit *move* would be an utterance such as, "Just a moment." letting the other participant know that the speaker was engaged in a non-dialogue activity. ABOUT_PARTNER *moves* typically occurred after a period of silence and were queries about what the active partner was doing. These subcategories of OFF-TASK *move* reflect Morley and Stephenson's (1977) *Referent* dimension which encodes who is being talked about in an utterance, but is fundamentally different because it has to refer to an action by the person. One motivation for the inclusion of this category was personal communication with British Telecommunications customer service software design experts who reported that agents were trained to fill in time when software searches were holding up dialogues, by telling the customer what they were doing without indicating there to be a fault in the software (Rob Davis, pers comm).

ABOUT_TECH, ABOUT_SELF and ABOUT_PARTNER combined reflect the content Newlands (1998) found to be responsible for the majority of the increase in Conversational Games Analysis EXPLAIN *moves* she found when people performed the travel game over video-mediation (VMC) rather than face-to-face. She found that nearly 42% of the EXPLAINs produced by the travel agent when the VMC was used were explaining what the agent was doing or commenting on the technology whereas only 5 in total performed such functions when the tasks were performed face to face.

ABOUT_LANGUAGE *moves* indicated occasions where the language the participants were using itself failed and became explicit in the dialogues. This typically represented occasions when the agent or customer encountered names of airports whose correct pronunciation they did not know; and so represented to some extent occasions when the contents of the participants' software or off-line resources imposed on the dialogues. These were essentially the same as *moves* which would have been marked "meta" in the original Conversational Games Analysis scheme.

TO_SELF *moves* were included to recognise times when an utterances served no purpose in the dialogue, but was directed to the speaker themselves or to their tools. They typically encoded utterances in which the speaker reminded themselves of decisions which had been made between the participants before testing those decisions on the software in the agent's case, or making notes of them in the customer's case. Occasions in which the agent spoke to the experimenter, which reflect occasions in which an agent would ask a co-worker for assistance, were considered TO_SELF OFF-TASK *moves*.

INTERJECTION *moves* need not be directly related to the presence of the participants' tools, since they merely encoded occasions in which jokes were made, but these jokes often related to the presentation of the sales opportunity to the customer and may have reflected the social relationship between the participants, so were included among the marked OFF-TASK *moves*. In the original Conversational Games Analysis scheme, INTERJECTIONs were marked but were not assigned a *conversational move* category label; in consequence, their inclusion within the OFF-TASK subcategories also recognises that they may also give some indication of how the participants are communicating. For instance, in response to an agent's statement that the travel pass would be useful, one customer replied "What like a spanner?" subsequently indicating that they found the forced presentation of the travel pass an imposition which was to be laughed at.

Response moves to OFF_TASK *initiating moves* were considered to share their *initiating move*'s OFF-TASK subcategory.

Because many of the potential combinations of Conversational Games Analysis *moves* and OFF-TASK orientation did not occur in the dialogues (roughly 86%), and several of the combinations occurred in only isolated cases, it was decided to only consider whether a *move* was ON-TASK or OFF-TASK in the statistical analysis. This decision reflects Wright and Monk's (1989) discovery that *breakdowns* did not occur frequently enough in their set of dialogues to be used as an evaluative measure.

The encoding of the *task orientation* in another dimension compliments the original Conversational Games Analysis scheme well, because it potentially permits the capturing of topic information without weakening the functional information captured by Conversational Games Analysis.

4.1.3.2 Procedure

The dialogues in which the agent had received the Integral Window and Spoken prompts were coded according to the schemes described above, at a rate of approximately 1 1/2 days per dialogue.

All the dialogues were coded by the author, but one dialogue was coded by a second coder so that reliability of the coding scheme could be checked. This coder was experienced with Conversational Games Analysis, but had received only an hour's training in the application of the *task orientation* coding.

Because dialogues are not divided into the precise units which reliability statistics are designed to measure, the analyses of reliability have to consider both the reliability with which the raters agreed on turn boundaries and their reliability in coding *moves* on which the boundaries were agreed (Carletta et al., 1997). When the boundaries were agreed, reliability in coding both the category of *initiating move* and the *task orientation* of *moves* was good. The analyses which follow include occasions where there had been agreement on the segmentation of *move*, but one of the coders had classified the *move* as a *response*, not an *initiating move*. There was a 79% agreement in the classification of *initiating moves*, which produced a Kappa of 0.71 (N=173 [the number of units], k=2 [the number of coders]). There was a 92% agreement on whether *moves* were ON- or OFF-TASK which produced a Kappa of 0.67 (N=173 [the number of units], k=2 [the number of coders]). Both of these are within the range considered to show "substantial" agreement by Landis and Koch (1977) and within the range Krippendorff (1980) considers appropriate for drawing tentative conclusions from the α statistic, which is a generalised version of kappa.

The agreement on turn boundaries was disappointingly low, however. Carletta et al. (1997) used Kappa assess reliability of *move* segmentation as well, but admitted that in this case the definition of units for analysis by Kappa was difficult. Her answer was to use transcribed word boundaries as the units, but the measure of reliability then depends greatly on conventions in the transcription. She suggests that pairwise percent agreement is the best alternative when there is no reasonable independent definition of units, and that is the only measure which has been applied here. There was only 69% agreement between the coders in the segmentation of *moves*.

There were 74 occasions on which there was disagreement on the boundaries of turns, but 69 (93%) of these were occasions where the author defined a *move* but the novice did not. Of these, 42 were OFF-TASK *moves*, and 20 of these were ABOUT_SELF *moves* which were coded according to a convention the novice coder had ignored. On 56 occasions there was disagreement in segmentation because the novice coder did not recognise *moves* which the author had; these were not mutually exclusive from the OFF-TASK *moves* the author recognised, and 19 of the 20 conventionalised ABOUT_SELF *moves* were included in larger *moves* by the novice. It appears the novice coding is including a large amount of the OFF-TASK *moves* within ON-TASK moves.

The disagreement between the coders in *move* segmentation, and presumably some of the errors in *move* classification, appeared to be primarily caused by the novice coder poorly adapting their previous use of Conversational Games Analysis to the novel *task orientation* dimension. Because the author's reliability in coding dialogues with Conversational Games Analysis had been measured previously (Newlands, 1998), because the novice coder had received only limited instruction and had ignored conventions they had been informed of, and because an early intra-rater check the author had made of their own coding including both initiating moves and responses showed a 91% agreement in *move* segmentation and 89% agreement in the classification of *moves* whose segmentation was the same, it was considered appropriate to continue with the analysis of the author's coding.

4.2 Hypotheses

The hypotheses about the codings of the integral window and spoken prompted dialogues serve several goals. Their most direct aim was to test explanations for the large differences between the spoken and integral window prompted dialogues described in the last chapter, but beyond this are the thesis level aims of showing that the effects of the agent-computer interface can be captured in the agent-customer dialogue and that that which is captured can be used to provide insight into the design of the interface.

In the last chapter it was found that the presentation of four seven word (5 lexical item) prompts by spoken prompt rather than written in an integral window lead to an increase of over 500 words and 2 1/2 minutes in task-oriented consultations which took on average 1800 words and 13 1/3 minutes.

Two mutually inexclusive explanations were provided for this. Either the spoken prompts were producing a **disruption** in the dialogues, or the spoken prompt was **promoting** more verbosity in the agent-customer dialogues. More detailed descriptions of these explanations were given in this chapter's introduction but in brief they imply the following.

In the case of **disruption**, the style of prompt would be leading to confusion between the agent and customer, either immediately after the prompt if the sales opportunity was introduced poorly, or globally because the presence of the speaking computer was making the tripartite communication awkward.

In the case of **promotion**, the style of prompt would be leading to the agent and customer conversing more, whether that discussion was task oriented or social. If this occurred immediately after the presentation of prompt it could be because the speaking of the prompt made the agent try to sell the opportunity more actively or because the stimulus (prompt) being in the same modality as the response (mention and discussion of sale) facilitated that response. If the agent and customer conversed more over the whole dialogues, then the speaking would appear to be changing the dynamic of the interaction between the agent, customer and computer and it was proposed this would be because speech lead to the computer's role being seen as more commanding than advisory.

If both **disruption** and **promotion** were found to be happening over the whole dialogues, this could indicate that the role of the computer is instead being seen as an Orwellian Big Brother. In this case the agent would feel compelled to converse more with the customer, but at the same time be constantly aware of the the computer's monitoring, so that their conversation was uncomfortable and awkwardly constructed.

The success of the coding schemes in supporting the two thesis-level aims of this chapter were judged on how well the predictions can be represented in the coding scheme's terms, and how well the analyses of the coding test the various explanations.

4.2.1 Hypotheses

As has been indicated, the reason for the large global increase in time and words caused by the style of prompt was expected to be either **disruption** or **promotion**, and although the cause of these might be different either immediately after the presentation of the prompt or during the whole consultations the means of determining whether there was **disruption** or **promotion** could remain the same.

Hypotheses can be made about the patterns of particular Conversational Games Analysis *moves* and their *task orientation* which would occur if **disruption** or **promotion** were taking place.

4.2.1.1 CHECKs

The most easily built such hypothesis is about **disruption**, since the CHECK *initiating move* is directly designed to capture occasions when conversants are having trouble understanding each other.

CHECKs represent utterances in which one speaker asks another for clarification of something which has just been spoken. In Clark and Schaefer's (1989) terms they are negative evidence of grounding, which are signs that common ground has been lost, and this loss of common ground is something which would be expected to accompany **disruption** of the dialogues.

For instance, immediately after the agent has mentioned the opportunity to the customer, if the agent has introduced it poorly and the customer does not understand what this new topic is, they might produce a CHECK requesting the agent to enlighten them (see Example 4/1).

• Agent
• *Game* 29 QUERY-YN em
well, the thing is right, you see you're going to be staying in [uh] Cedar City for three days, right so you should, you should get a travel pass, man,
• *Move* QUERY-YN

-> • Customer
aye
• *Move* ACKN
-> • *Game* 30 CHECK em'
-> **travel pass**
-> • *Move* CHECK

• Agent
this, aye
• *Move* REPLY-Y
• End 30

-> • Customer
-> • *Game* 31 CHECK em'
-> **what's that**
-> • *Move* CHECK

• Agent
• *Game* 32 EXPLAIN em"
you can buy a travel pass,
• *Move* EXPLAIN

-> • Customer
-> **what's that**
-> • *Move* CHECK repeated

...

Example 4/1 Example of Customer using CHECK when confused the the agent's introduction of the sales opportunity. (CHECKs marked with arrows) (Spoken Prompt, Eighth Trial, First Prompt).

More CHECKs would imply more dialogue to rebuild the common ground, so it would be expected that were **disruption** to be a factor in the large differences in time and words it would be the spoken prompted dialogue, which took longer and contained more words, which would also contain more CHECKs.

The first hypothesis to be built, then, is that the spoken prompt will lead to more CHECKs.

H1 *The spoken prompt will lead to more use of CHECK moves.*

This hypothesis is directly concerned with testing the **disruption** explanation for the large increase in time and words in the spoken prompted dialogues.

4.2.1.2 ALIGNs

The ALIGN *initiating move* is similar to CHECKs in that its role can be to overcome misunderstanding within dialogues, but whereas the CHECK is produced by the person who does not understand, an ALIGN is produced when the speaker does not believe they have been understood (see Example 4/2). Although these do not directly represent one of the forms of negative evidence for grounding which Clark and Schaefer (1989) recognise, they can also be interpreted as a sign of poor grounding.

• Customer
• *Game* 24 QUERY-YN em'
so maybe one night in Colorado, two nights in Denver
• *Move* QUERY-YN

• Agent
yes you could do that.
• *Move* REPLY-Y
eh no sorry you can't
• *Move* REPLY-N

• Customer
• *Game* 25 CHECK em''
can't
• *Move* CHECK

-> • Agent
no because I can't find Denver listed, it's not listed I'm sorry to say it's not available,
• *Move* REPLY-W
-> • *Game* 26 ALIGN em'''
-> **ok**
-> • *Move* ALIGN

• Customer
• *Game* 27 CHECK em''''
what you just can't go to Denver at all
• *Move* CHECK

• Agent
no, [eh] yes I'm sure you could go to Denver but it'd probably be from other airports, from other States alright
• *Move* REPLY-W

• Customer
right that's fine,
• *Move* ACKN
• End 27, 26, 25, 24, 18,

Example 4/2 Example of agent using ALIGN to determine that a previous utterance has been understood. (ALIGN marked with arrows) (Spoken Prompt, Eleventh Trial).

Whereas negative evidence is typically given by the hearer of an utterance, an ALIGN used to question whether the hearer has understood would essentially be a form of self-correction (Schegloff, Jefferson and Sacks; 1977) in which the person who can solve misunderstanding initiates the correction of utterances they do not believe to be successful.

The second hypothesis is, then, that the spoken prompt will lead to more ALIGNs. This would support the proposal that **disruption** caused by the spoken prompt could be leading to the overall increase in time and words.

H2 The spoken prompt will lead to more use of ALIGN *moves*.

Because, as can be seen in the definition provided in Figure 4/3, ALIGNs may serve many more purposes than that exemplified above, it is less likely that ALIGNs will be as sensitive to **disruption** as CHECKs; but they have been found sensitive to changes in media of communication by both Carletta et al. (1997) and Newlands (1998).

4.2.1.3 EXPLAINs

The EXPLAIN *initiating move* is more likely to be capture **promotion** effects. EXPLAINs are unsolicited mentions of information which has not previously been made available in the dialogue, and as such an increase in their production would indicate that the speakers are voluntarily contributing to the dialogues.

Examples 4/3 and 4/4 provide an example of how **promotion** of EXPLAINs might occur immediately after the presentation of the prompt. In both cases, the examples are from the agent's first mention of the opportunity to the customer. In Example 4/3, which follows an integral window prompt, the agent simply asks the customer whether they would be interested in buying a travel pass without explaining the product. In Example 4/4, which follows a spoken prompt, the agent provides a considerable description of the product before asking whether the customer will buy it.

- Agent
- *Game* 10 QUERY-YN em (MENTION)
  [eh] would you like to buy a travel pass
- *Move* QUERY-YN (MENTION)

Example 4/3 Example of agent introducing sales opportunity without EXPLAIN.
(Integral Window, Fifth Trial, First Prompt).

```
->  • Agent
->  • Game 27 EXPLAIN em (MENTION)
->      now what you're going to [um], need there I would say
->      is probably the travel pass, and this gives you
->      unlimited travel while you're there [em] on the local
->      bus and commuter rail [em] for the duration of your
->      stay in the place, so obviously that's [em] fairly
->      indispensable for getting around [eh] a large city like
->      Seattle
->  • Move EXPLAIN (MENTION)

    • Customer
        yeah
    • Move ACKN
    • End 27

    • Agent
    • Game 28 QUERY-YN em
        so you you would like the travel pass then
    • Move QUERY-YN

    • Customer
        yeah I would, yeah
    • Move REPLY-Y
    • End 28
```

Example 4/4 Example of agent introducing sales opportunity with EXPLAIN. (EXPLAIN marked with arrows) (Spoken Prompt, Sixth Trial, First Prompt).

On the other hand, **disruption** could also be accountable for an increase in the use of EXPLAINs.

Newlands (1998) found that the majority of the increase in EXPLAINs she found when subjects performed the travel game by video mediation, rather than face-to-face, was due to the agent commenting on mistakes they made interacting with their computer or telling the customer what actions they were performing with their computer. These types of EXPLAIN represent occasions where the computer is imposing itself in the dialogue between the agent and customer and should be considered **disruptions**.

In the current coding scheme, such utterances are represented as the OFF-TASK categories of ABOUT_TECH, ABOUT_SELF and ABOUT_PARTNER; and Example 4/5 provides an example of such a **disruptive** EXPLAIN.

```
->    • Agent
->    • Game 170 OFF-TASK(ABOUT_SELF) EXPLAIN
->        and the time, oh dear, now I've got you up to three
->        days in Dallas, I'm afraid got it wrong, {laughs}
->    • Move OFF-TASK(ABOUT_SELF) EXPLAIN

      • Customer
          {chuckles}
      • Move OFF-TASK(ABOUT_SELF) ACKN

->    • Agent
->    • Game 171 OFF-TASK(ABOUT_SELF) EXPLAIN em
->        oh wait hang on a minute I think I can get it down
->        again, I presume
->    • Move OFF-TASK(ABOUT_SELF) EXPLAIN

      • Customer
          {chuckles} down to two
      • Move OFF-TASK(ABOUT_SELF) ACKN
      • End 171

->    • Agent
->    • Game 172 OFF-TASK(ABOUT_TECH) EXPLAIN em
->        no obviously not, no maybe not, it's just going to
->        keep on giving,
->    • Move OFF-TASK(ABOUT_TECH) EXPLAIN
->    • End 172
->    • Game 173 OFF-TASK(ABOUT_TECH) EXPLAIN em
->        you have to go back to the beginning,
->    • Move OFF-TASK(ABOUT_TECH) EXPLAIN
->    • End 173
```

Example 4/5 Example of OFF-TASK EXPLAINs showing agent's computer imposing itself in agent-customer dialogue (OFF-TASK EXPLAINs marked with arrows) (Spoken Prompt, Ninth Trial).

If OFF-TASK EXPLAINs were a major part of any increase in EXPLAINs that would indicate the style of prompt responsible for the increase was causing **disruption**.

Whether they were **promotional** or **disruptive**, additional EXPLAINs would be expected to contribute to the global increase in time and words when the spoken prompt was employed, therefore it would be hypothesized that the spoken prompt would lead to more EXPLAINs than the integral window.

H3 *The spoken prompt will lead to more use of EXPLAIN moves.*

If the hypothesis is supported, the amount to which ON- and OFF-TASK orientations of EXPLAIN contribute to the increase will have to be investigated to determine the degree to which the style of prompt leads to **promotion** of discourse or **disruption**.

4.2.1.4 QUERY-YNs

An increase in the use of QUERY-YN *moves* is also likely to occur if **promotion** is contributing to the global increase in time and words when the spoken prompt is used.

This is because QUERY-YNs, just like EXPLAINs, are *initiating moves* which introduce new information to the dialogue. QUERY-YNs allow the speaker to introduce that information in as much detail as they feel necessary, whilst only requiring a minimal "Yes" or "No" response from the hearer. This can be contrasted with QUERY-Ws which tend to be requests for new information to be added by the hearer; if the speaker wants to volunteer information to the dialogue QUERY-YNs are more likely but if they want information to be given, QUERY-Ws would be used[4].

If the spoken prompt is **promoting** more dialogue, then it would be expected that information would be volunteered rather than requested, and more use of QUERY-YNs would be expected. This could include both more occurrences of QUERY-YNs, in which more options are presented to the hearer, and longer QUERY-YNs, in which more information about the options are presented.

Examples 4/6 and 4/7 exemplify the contrasting uses of QUERY-YNs. In Example 4/6, the agent simply asks the customer if they would like to take a flight which they consider to be the best option. In Example 4/7 the agent is again asking the customer where they would like to go, but provides two options with separate QUERY-YNs and within the QUERY-YNs tells the customer why they consider the options desirable.

---

[4] In QUERY-W games, any compulsion to be verbose is more likely to appear in REPLY-W response moves to the QUERY-W initiating moves; because response moves are not independent of their initiating moves they are not being analysed here (see Section 4.1.1.4).

```
->    • Agent
      ...
->    • Game 177 QUERY-YN
->       and from Great Falls you'd be interested in travelling
->       to Billings
->    • Move QUERY-YN

      • Customer
      • Game 178 INSTRUCT em
         go to Billings
      • Move INSTRUCT
      • End 178
         bang on man
      • Move REPLY-Y
      • End 177
```

Example 4/6 Example of succinct QUERY-YN (QUERY-YN marked with arrows)
(Integral Window, Fourteenth Trial).

```
      • Agent
      • Game 162 EXPLAIN
         [um] so basically the time is your own, whether you
         want to spend another day in in the Great Falls, or,
      • Move EXPLAIN
      • End 162
->    • Game 163 QUERY-YN
->       I mean there's [eh] there's places like Colorado
->       Springs which you haven't actually touched yet in
->       Colorado,
->    • Move QUERY-YN
->    • End 163
->    • Game 164 QUERY-YN
->       [em,] or if you wanted to go to a state that you
->       haven't been in at all like Idaho, you know for for
->       the day, just to get a feel of it, might be an idea
->    • Move QUERY-YN

      • Customer
         yeah, ok, yeah,
      • Move REPLY-Y
      • End 164
```

Example 4/7 Examples of verbose QUERY-YNs (QUERY-YNs marked with arrows)
(Spoken Prompt, Sixth Trial).

Because QUERY-YNs represent occasions where the speaker has to introduce information, but may choose how verbosely that information is introduced, the fourth hypothesis is that the spoken prompt will lead to more use of QUERY-YNs than the integral window.

H4 The spoken prompt will lead to more use of QUERY-YN *moves*.

It is not foreseen that **disruption** would be a potential explanation for an increase in QUERY-YNs but as with all the other *initiating moves* the effect of the style of prompt on the *task orientation* of *moves* will also be analysed and this should indicate **disruption** in the same way it would for EXPLAINs.

4.2.1.5 QUERY-Ws and INSTRUCTs

No hypotheses were made about the effect of the style of prompt on the remaining QUERY-W and INSTRUCT *initiating moves*, because neither were expected to be sensitive to an effect. This is because neither is particularly designed to capture behaviours which would be expected to represent either **promotion** or **disruption**, nor has either been found to be sensitive to effects in previous studies.

QUERY-W *moves* themselves tend to be short, simple questions such as "Where would you like to go next?", which give the hearer more chance to expand than the speaker; some may provide options, such as "Would you like to go to Denver, Los Angeles or Las Vegas?", but even then are strongly constrained by the options available.

As was described in section 4.2.1.4, where the hypothesis about QUERY-YNs was built, any effect of **promotion** in QUERY-W *games* is more likely to occur in REPLY-W responses to the QUERY-W *initiating moves* or embedded EXPLAIN *games* because it is here that the speakers have more opportunity to expand on why a choice is being made.

INSTRUCTs like QUERY-Ws are also more constrained in how they are used than the other categories of *move*. Informal analysis of the coding showed that the majority of INSTRUCTs were produced by the customer (roughly 80% of the total) and these came at the end of long exchanges between the agent and customer trying to find a suitable airport available to be flown to when the customer finally told the agent where to take them (see Example 4/8).

• Agent
• *Game* 10 QUERY-W
and your next destination please
• *Move* QUERY-W

• Customer
it's Pocatello
• *Move* REPLY-W

• Agent
Pocatello,
• *Move* ACKN
• *Game* 11 EXPLAIN em
[um] sorry sir we don't have Pocatello from, from that airport
• *Move* EXPLAIN
• End 11

• Customer
ok then, [em]
• *Move* READY
I'd like to go to Ely
• *Move* REPLY-W cont

• Agent
• *Game* 12 EXPLAIN em
that's unavailable as well,
• *Move* EXPLAIN
• End 12
• *Game* 13 EXPLAIN em
we have eh Colorado, Dallas, Denver, Houston, Omaha, Phoenix, Portland, Salt Lake City, San Francisco and Seattle
• *Move* EXPLAIN

• Customer
oh
• *Move* ACKN
• End 13
-> • *Game* 14 INSTRUCT em
-> **can I go to Seattle then please**
-> • *Move* INSTRUCT

• Agent
yeah
• *Move* ACKN
• End 14, 10

Example 4/8 Example of typical INSTRUCT (INSTRUCT marked with arrows)
(Integral Window, Seventh Trial).

4.2.2 Local and Global Hypotheses

In the last chapter, when the initial analysis of the effects of changing the style of prompt used in an agent's interface were presented, both the hypotheses and the analyses were divided into local effects occurring immediately after the prompt, and global effects occurring over the whole dialogues. This was because different means of analysis were used in the two situations, and because different mechanisms were expected to be responsible for any discovered differences.

Different mechanisms could still be called upon to explain differences in the functional process of dialogues captured locally and globally, for instance following Halliday (1989) and to a certain extent Byblow and Corbett (1989) and Wickens et al. (1983) (see Sections 3.0.1.1.1 and 3.0.1.1.2), it could be predicted that local effects of the prompt could be stimulus-response reactions whereas predominantly following Cotton and McCawley (1983, referred to in Baber et al., 1992) it could be predicted that global effects were due to a more wide ranging effect on the perception of the agent's computer's role within the dialogue.

For a number of reasons, however, it was decided not to formulate local and global hypotheses nor to report local and global analyses separately.

Firstly, no predictable local effects were found to be supported. The most likely local effects to have occurred were the spoken prompt to have been spurring the agent into action more than the integral window, would have been more mentions of the opportunities and quicker mentions, but neither of these was supported.

This argument could have been countered by noting that the analysis of the coding was likely to capture very different effects, and indeed would be capturing those effects in a different section of dialogue since the area where analysis of the coding would be expected to be most fruitful would be between the agent's mention of the opportunity and the resolution of the sale[5].

---

[5] This section was only defined following the coding because without the objectiveness of the coding the specification of the sections boundaries was predicted to be too indistinct. It was finally defined as the period of dialogue between the beginning of the conversational game in which the opportunity was mentioned and the end of the conversational game in which the customer purchased or rejected the sale.

To redress this counter-argument, however, it was noted in preliminary evaluation of the dialogues that the size of the section of dialogues between the mention and resolution did not represent a large part of the whole dialogues, and the difference in the size of these sections following the integral window and spoken prompt did not represent a considerable part of the large global difference whose explanation the coding was intended to support.

On average, over the whole dialogue there were 228 turns, but between all mentions and resolutions in each dialogue there were only 28 (12.5%). Over the whole dialogue, there were 1819 words, but between all mentions and resolutions in each dialogue there were 274 words (15.1%), and over the whole dialogue there were 183 *initiating moves* but between all mentions and resolutions in each dialogue there were 20 *initiating moves* (11.0%). The coding was intended to explain the large 511 word (33%) statistically significant global increase in dialogues when the spoken prompt was used over those in which the integral window was used, but there was only a 50 word (20%) increase between the mention and resolution. Furthermore, this 50 word increase was not found to be indicative of a significant main effect of the style of prompt (Two-way ANOVA; Between Subjects Factor: Style, 2 levels: Integral Window and Spoken Prompt; Within Subjects Factor: Role, 2 levels: Agent and Customer; $F[1,28]$: 1.322, $\underline{P}$: 0.260).

Nevertheless, where local analyses may elucidate global findings they will be reported. For instance, this will take place if evidence of a "sit-up-and-listen" effect would support a global difference in the use of a *move*, or where one of the hypotheses is not supported over the whole dialogues but it is possible a stimulus-response reaction could be occurring immediately after the prompt. Although the latter findings would be unlikely to unequivocally explain the large global difference, they would demonstrate the ability of the coding scheme to capture fine-detailed effects of the prompt on the agent-customer dialogue, supporting the schemes usefulness in that regard.

4.2.3 Summary of Hypotheses

The four hypotheses made about the use of particular categories of Conversational Games Analysis *initiating moves* were as follows.

H1 The spoken prompt will lead to more use of CHECK *moves.*

H2 The spoken prompt will lead to more use of ALIGN *moves.*

H3 The spoken prompt will lead to more use of EXPLAIN *moves.*

H4 The spoken prompt will lead to more use of QUERY-YN *moves.*

At first these hypotheses may seem tautologous; given that there is a 511 word and 154 second increase in the size of the consultations when the spoken prompt is used over those in which the integral window is used, it could be expected that there will be a concomitant general increase in the use of all *initiating moves.*

The hypotheses are being used, however, to capture specific changes in the way particular conversational functions are being used, and with these findings it is intended to build a model of the interaction which will show how much each of the predicted explanations is responsible for the increase. Each of the hypotheses is intended to indicate either (1) direct support for one of the predicted explanations or (2) where further investigation could be targeted to find more support. Only those explanations whose related hypotheses are supported should be considered to be playing an important role in accounting for the increase in words and time.

The overall aim is to demonstrate that the coding scheme is capable of informing the design of the interface the agent in a tripartite customer service situation such as that being investigated, and tentatively argue that fine-grained analysis in general would be beneficial in such situations.

As stated in the Section 4.2.0, the first way in which the scheme had to be assessed was how well expectations about the way the different prompts would influence the agent-customer dialogues could be represented as hypotheses about the use of particular Conversational Games Analysis categories. In the last section, based on the definitions of the categories of *initiating move*, previously observed patterns of usage and preliminary observations of the agent-customer dialogues several such hypotheses were successfully built, including relevant reference to the newly introduced *task orientation* dimension of coding. Therefore, at least the initial proposal of fine-grained coding schemes such as Conversational Games Analysis to inform the design of customer service agents' interfaces is supported.

In the next section, where the statistical analysis of the coded dialogues is reported, the stronger proposal that these coding schemes can test alternative explanations for observed differences will be evaluated.

4.3 Analysis

In this section, the statistical analysis of the Conversational Games Analysis and *task orientation* coding of the dialogues will be reported.

The aim will be to test the hypotheses built in the last section, with the intention of concurrently testing the **disruption** and **promotion** explanations for the large 511 word increase in the size of consultation when the spoken prompt, rather than the integral window, was employed in a customer service agent's computer interface. The overall goal of this analysis is to show that these coding schemes can inform decisions to be made about the design of an agent's interface to be used in such a tripartite communication.

To achieve this several aspects of the use of each category of Conversational Games Analysis *initiating move* were measured.

The primary measure taken was counts of how many occurrences of each category of *initiating move* there were over the whole dialogue. This is the measure of coding which has been used in previous studies which have used Conversational Games Analysis (e.g. Newlands, 1998) though many studies have counted *conversational games* rather than *initiating moves* (Kowtko, 1997; Carletta et al., 1996). It is also the typical measure used when other coding schemes have been employed (Bales, 1950; Morley and Stevenson, 1977).

To augment this measure, counts were also taken of the amount of words which each category accounted for in the dialogue. This measure is more similar to that used by the Olsons (Olson, Olson, Carter and Storrøsten, 1992), who measured the amount of time spent on each category.

The two measures combined should give a clear picture of how the style of prompt is affecting the agent and customers' use of their dialogue.

Where it was expected that the *task orientations* of a category of *move* could support different explanations, as for instance more use of ON-TASK EXPLAINs supporting **promotion**, but more use of OFF-TASK EXPLAINs supporting **disruption** (see Section 4.2.1.3), or one of the *task orientations* could emphasize a relationship between the *initiating move* category and the style of prompt, the two classes of *task orientations* were analysed separately.

A final measure taken was the number of occurrences of each category of *initiating move* which occurred between the agent's mention of the sales opportunity to the customer and the resolution of the sale. For the reasons given in Section 4.2.2 this was only used as a supplementary measure, and so will only be reported where individual arguments for it can be built.

Before describing the analysis related to each hypothesis individually, the main effect of the style of prompt on all the categories will be presented to provide an overview of the general functional use of dialogue within each prompt's consultations. Further main effects such as those of role and *task orientation*, and simple effects, will be reported in the following subsections where the individual testing of the hypotheses is reported.

A Three-way ANOVA with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker (2 levels: Agent and Customer) and *task orientation* (2 levels: ON- and OFF-TASK) as within subjects factors was used to analyse the counts of the number of occurrences of each category of *initiating move*.

Although the effect of the style of prompt on each category did seem to vary somewhat, for instance there were on average 20.4 (30.7%) more EXPLAIN *moves* in Spoken Prompt dialogues than in Integral Window but only 0.3 (<1%) more QUERY-W *moves*, there was no main effect of the style of prompt on any of the six categories of *initiating move* (see Table 4/2).

Category	Integral Window	Spoken Prompt	ANOVA	P
EXPLAIN	66.5	86.9	F[1,28] 2.835	0.103
QUERY-YN	33.2	39.7	F[1,28] 1.979	0.170
QUERY-W	32.1	32.4	F[1,28] 0.004	0.951
CHECK	19.6	24.4	F[1,28] 0.974	0.332
INSTRUCT	11.1	12.9	F[1,28] 0.452	0.507
ALIGN	2.6	4.1	F[1,28] 1.284	0.267

Table 4/2 Mean occurrences of both *task orientations* of each category of *initiating moves* during the whole dialogues, by style of prompt.

This suggests that the major effects of the change in the style of prompt are purely in the verbosity of the speakers, not in the pattern of exchanges they make. The counts of the total number of words uttered within each category of *move* would indicate whether this verbosity was particularly evident in some categories of *initiating move* more than others and thereby indicate the functions of dialogue the participants were concentrating on.

Although at first sight increased verbosity in categories of *move* could be taken to indicate **promotion**, an increase in *moves* which had been hypothesized to indicate **disruption** it might also indicate that speakers were were forced to concentrate on resolving that disruption.

The counts of the amount of words uttered within each category of *initiating move* were analysed using Two-way ANOVAs with style as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer).

There was an increase in the mean amount of words spent on all categories of *initiating move*, but significant main effects of the style of prompt were only found on the total amount of words uttered during two (see Table 4/3).

There was a significant main effect of style on the total amount of words used to EXPLAIN (F[1,28] 5.083, P: 0.032) with the Spoken Prompt leading to 203.8

(45%) more than the Integral Window, a significant main effect of style on the total amount of words used to QUERY-YN (F[1,28] 11.722, P: 0.002) with the Spoken Prompt leading to 116.7 (45%) more than the Integral Window.

It is also worth noting a trend towards a significant main effect of the style of prompt on the total amount of words used to ALIGN (F[1,28] 3.344, P: 0.078) with the Spoken Prompt leading to 14.7 (163%) more than the Integral Window.

All told, these significant increases and the trend account for 66% (335 words) of the total 511 word increase for which the dialogue analysis is being used to find an explanation.

Category	Integral Window	Spoken Prompt	ANOVA	P
EXPLAIN	452.7	656.5	F[1,28] 5.083	0.032
QUERY-YN	256.9	373.6	F[1,28] 11.722	0.002
QUERY-W	276.4	294.5	F[1,28] 0.399	0.533
CHECK	92.7	131.8	F[1,28] 2.523	0.123
INSTRUCT	70.4	79.3	F[1,28] 0.250	0.621
ALIGN	9.0	23.7	F[1,28] 3.344	0.078

Table 4/3 Mean number of words uttered within all occurrences of each category of *initiating move* during the whole dialogues, by style of prompt.

These findings provide more support for the contention that the the spoken prompt is **promoting** the participants to converse more freely.

The two *initiating moves* which showed a significant increase in the amount of words spent on them were EXPLAINs and QUERY-YNs, both of which had hypotheses built about them (**H3** and **H4**) which were considered most likely to support such **promotion**.

As was described in Sections 4.2.1.3 and 4.2.1.4, both categories of *move* involve the introduction of new information to the dialogue, and an increase in the amount of words spent introducing new information most likely indicates that speakers feel it is suitable to volunteer more detail.

Although an increase in the use of ALIGNs was expected to support a **disruption** explanation for the global increase in the size of the agent-customer dialogues (Section 4.2.1.2; **H2**), the observed pattern of results may in fact hint at more support for the spoken prompt **promoting** free conversation between the agent and customer.

As was mentioned in the hypotheses, ALIGNs occur when speakers believe they have not been understood, and would therefore be expected to indicate when those speakers' utterances have been ill-formed; but they are also self-corrections and therefore attempts to solve potential mis-understandings before they cause subsequent problems in the dialogues.

No significant increase in the number of ALIGN *moves*, nor in the amount of CHECK *moves*, as an effect of the spoken prompt suggests that the spoken prompt was having no effect on the amount of mis-formed utterances, and so was not contributing to an increase in the amount of **disruption** in the dialogues. The trend towards an increase in the amount of words spent ALIGNing would suggest that instead the effect was on how mis-formed utterances were dealt with.

In this case, the self-correctional aspect of ALIGNs should be highlighted, and it could tentatively be inferred that the speakers are taking more pains to volunteer solutions to potential misunderstanding rather than waiting for solutions to be requested. This would hint that the spoken prompt is **promoting** agent-customer discourse.

Example 4/9 provides an example of a short ALIGN produced in a dialogue where the Integral Window was being used, and Example 4/10 provides an example of a similar but longer ALIGN produced in a dialogue where the Spoken Prompt was being used. ALIGNs are typically short, so the distinction is not large.

• Customer
can I go to Phoenix
• *Move* QUERY-W repeated

• Agent
yeah,
• *Move* REPLY-W
• End 71
• *Game* 78 QUERY-YN
ten o'clock in the morning
• *Move* QUERY-YN

• Customer
yeah
• *Move* REPLY-Y

• Agent
ok
• *Move* ACKN
• End 78
• *Game* 79 EXPLAIN
... [um], no problem
• *Move* EXPLAIN

-> • Customer
right,
• *Move* ACKN
• End 79
and then right,
• *Move* READY
-> • *Game* 80 ALIGN
-> **from there yeah**
-> • *Move* ALIGN

• TA
yeah
• *Move* REPLY-Y
• End 80

Example 4/9 Example of short ALIGN.
(Integral Window, First Trial).

```
        • Customer
        • Game 20 QUERY-YN em'
           am I explaining myself clearly enough here
        • Move QUERY-YN
        • End 20

        • Agent
        • Game 21 CHECK em'
           could you just repeat that again, fly to Denver
        • Move CHECK

->      • TC
           if if I could,
        • Move REPLY-Y
->      • Game 22 ALIGN em"
->         right I'm going to Colorado Springs first
->      • Move ALIGN

        • TA
           yes
        • Move REPLY-Y
        • End 22

        • TC
           then if I could fly from there to Denver
        • Move REPLY-Y cont

        • TA
           right
        • Move ACKN
```

Example 4/10 Example of lengthier ALIGN.
(Spoken Prompt, Eleventh Trial).

Of course, as was expected when the hypotheses were built, it will be necessary to look at the more precise use of all these *moves* to ensure these initial readings of the results are not masking alternative explanations or whether one speaker could be leading the more general increase.

4.3.1 Investigation of Hypotheses

Two aspects of the use of each category of *initiating move* were analysed; the number of occurrences of each category and the amount of words spent within them.

The number of occurrences of each category was first analysed with Three-way ANOVAs taking style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt), role of speaker as a within subjects factor (2 levels: agent and customer) and *task orientation* as a within subjects factor (2 levels: ON- and OFF-TASK). If appropriate, the number of occurrence of ON- and OFF- *task orientations* were individually analysed with Two-way ANOVAs taking style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: agent and customer). Simple effects were calculated when the previous analyses appeared to be supporting the hypotheses.

4.3.1.1 CHECKs

CHECKs represent occasions when a hearer of an utterance has asked the speaker to clarify what they have said, capturing negative evidence of grounding (Clark and Schaefer, 1989). As such, they would be expected to accompany **disruption** which has caused speakers to utter ill-timed or ill-formed contributions to the dialogues.

In this study, such **disruption** was posited as one possible explanation for the large global increase in the size of the dialogues when the agent's computer interface used a spoken prompt rather than an integral window. In consequence a hypothesis was built that the spoken prompt would lead to more use of CHECKs than the integral window.

> **H1** The spoken prompt will lead to more use of CHECK *moves.*

As reported in Section 4.3.0, there was in fact found to be no main effect of the style of prompt on either the number of CHECKs produced, or on the amount of words used in CHECKs, and so this measure does not immediately suggest that **disruption** is playing an important role in the global increase.

Nevertheless further investigation of the way CHECKs were used was performed to test whether evidence for **disruption** could be found within any aspect of the agent and customer's dialogue.

The Three-way ANOVA performed on the number of CHECKs generated, taking style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt), role of speaker as a within subjects factor (2 levels: agent and customer) and *task orientation* as a within subjects factor (2 levels: ON- and OFF-TASK), found no main effect of role (F[1,28] 0.931, P: 0.343) but a significant and sizeable main effect of *task orientation* (F[1,28] 90.936, P: <0.001) with there being on average 42.6 ON-TASK CHECKs but only 2.5 OFF-TASK CHECKs.

This would suggest that the use of CHECKs is not dependent on the different communicative activities the role of the speaker imposes on the participants in the travel game, and that by far the greater proportion of CHECKs were being used to acquire clarification of misunderstandings related to the organisation of the route around the US and the sales subtask.

None of the interactions between style of prompt, role of speaker and *task orientation* were found to be statistically significant at greater than the 10% level, but because of the large dichotomy between the number of ON- and OFF-TASK CHECKs each *task orientation* was independently statistically analysed.

Two-way ANOVAs on the ON- and OFF-TASK orientations of ALIGN, with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer), showed the only effect or interaction which was significant at more than the 10% level was the main effect of role on OFF-TASK CHECKs (F[1,28] 10.287, P: 0.003) (see Table 4/4).

Style of Prompt	Agent OFF-TASK CHECKs	Customer OFF-TASK CHECKs	Both Roles' OFF-TASK CHECKs
Integral Window	0.5	0.1	0.7
Spoken Prompt	0.7	0.3	1.0

Table 4/4 Mean occurrences of OFF-TASK CHECK *moves* during the whole dialogues, by style of prompt and role of speaker.

This small (0.3 *move*) main effect of role could provide no insight into the reason for the large increase due to the change in the style of prompt. Informal analysis of the dialogues suggested that the majority of the increase was due to

ABOUT_LANGUAGE OFF-TASK CHECKs uttered by the agent as a result of the customer mispronouncing airport names.

The Two-way ANOVA performed on the count of words spent on CHECKs during the whole dialogues, with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer), found that as well as there being no main effect of role (see Section 4.3.0) neither the main effect of role nor the interaction between the role and style of prompt was significant at more than the 10% level.

To test whether the style of prompt was having an effect on the **disruption** of dialogues immediately after its presentation, possibly as a result of more stimulus-response reactions to the prompt described in section 4.2.2, the use of CHECKs between the agent's mention of the prompt and the resolution of the sale was also analysed.

A Two-way ANOVA with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer) found no main effect of style or role, nor an interaction between the two which was significant at more than the 10% level. The *task orientation* of CHECKs between the mention and resolution was not analysed because only 2 OFF-TASK CHECKs were produced.

In summary, there is no evidence that the style of prompt is having an effect on the production of CHECKs; so the first hypothesis is false. The result is that no support has been given to the proposal that **disruption** is the cause of the increased size of dialogues when the spoken prompt rather than the integral window is employed in the agent's computer interface.

4.3.1.2 ALIGNs

ALIGNs were initially expected to be sensitive to **disruption** within the dialogues, showing occasions where people who could solve misunderstandings queried whether their solution was necessary, and since **disruption** was one of the potential explanations for the global increase in the size of dialogues when the spoken prompt was used, it was hypothesized that the spoken prompt would lead to an increase in the use of ALIGNs.

H2 The spoken prompt will lead to more use of ALIGN *moves*.

The initial analysis of the use of ALIGNs (see Section 4.3.0) actually provided a pattern of results which appeared to better support the spoken prompt **promoting** more free and verbose dialogue between the agent and customer. Although there was no increase in the number of ALIGNs, there was a trend towards an increase in the amount of words spent ALIGNing, hinting that speakers were investing more effort in volunteering solutions to those potential misunderstandings which occurred.

Further analysis of the way ALIGNs were used was carried out to test whether this alternative was valid.

The Three-way ANOVA performed on the count of the number of occurrences of ALIGNs, with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker (2 levels: Agent and Customer) and *task orientation* (2 levels: ON- and OFF-TASK) as within subjects factors, found no main effect of role (F[1,28] 0.077, P: 0.783) but a significant and considerable main effect of *task orientation* (F[1,28] 21.792, P: <0.001) with there being on average 6.5 ON-TASK and 0.2 OFF-TASK ALIGNs.

As with CHECKs, this would suggest that the use of ALIGNs is not dependent on the different communicative activities the role of the speaker imposes on the participants in the travel game, and that the majority of ALIGNs were related to the travel game and sales tasks.

Similarly again to CHECKs, none of the interactions between the style of prompt, role of speaker and *task orientation* on the use of ALIGNs were significant at more than the 10% level. Because so many more ON-TASK than OFF-TASK ALIGNs were produced, and to find indications of where further evidence for the use of ALIGNs being influences by **promotion** not **disruption** of the dialogues, the number of each *task orientation* of ALIGN was analysed independently.

Neither the Two-way ANOVA, with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer), performed on ON-TASK ALIGNs nor that performed on OFF-TASK ALIGNs showed either main effects or interactions which were significant at more than the 10% level.

To test whether there was any effect of the style of prompt on the amount of ALIGNs immediately after the prompt's presentation, which would indicate local

disruption of the dialogues, counts of the number of ALIGNs between the mention of the opportunity and the resolution of the sale were analysed.

There was on average less than one ALIGN per dialogue between the mention and resolution, so any effects could not be contributing an important amount to the global increase in the size of dialogues. Two-way ANOVAs with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and customer) found neither main effect nor their interaction to be significant at more than the 10% level.

These analyses suggest that the style of prompt is having no significant effect on the amount of **disruption** within the dialogues, supporting the similar conclusion from the analysis of CHECKs.

The Two-way ANOVA performed on the count of words spent on ALIGNs, with style as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer), found that beyond the trend towards a main effect of style (see Section 4.3.0) neither the main effect of role nor the interaction between style and role was significant at more than the 10% level.

To test whether the trend was lead by either the agent or customer, which might indicate how far the effect of the style of prompt had permeated into the agent-customer dialogue, even though there was no main effect of role, the amount of words spent on ALIGNs by each role was independently analysed.

Bonferroni T-tests on the amount of words both the agent and customer spent ALIGNing found that although on average both the agent and customer spent more words on ALIGNs when the spoken prompt was used than when the integral window was used, neither the effect of style on the agent (df: 28, t: -1.2, P: 0.223) nor on the customer (df: 28, t: -1.44, P: 0.162) was statistically significant (see Table 4/5).

Style of Prompt	Agent ALIGN words	Customer ALIGN words	Both Roles' ALIGN words
Integral Window	3.7	2.8	9.0
Spoken Prompt	9.9	13.8	23.7

Table 4/5 Mean number of ALIGN words during the whole dialogues, by style of prompt and role of speaker.

Neither the agent nor the customer, then, appears to be contributing more effort to ALIGNs than their partner.

No more evidence for or against the reinterpretation of the proposed reason for the increase in the use of ALIGNs has been discovered with this further analysis. Given the arguments presented in Section 4.3.0, it nevertheless seems likely that the trend towards an increase in the amount of words spent on ALIGNs coupled with no increase in the number of ALIGNs indicates that the spoken style of prompt is more actively **promoting** discourse between the participants than **disrupting** it.

4.3.1.3 EXPLAINs

When building the hypothesis that the spoken style of prompt would lead to more use of EXPLAINs than the integral window, it was recognised that there could be two potential explanations for an increase.

> **H3** The spoken prompt will lead to more use of EXPLAIN *moves*.

The most likely explanation was considered to be that an increase in the use of EXPLAINs would be due to the spoken prompt **promoting** the human participants in the consultation to converse more openly. This was based on the understanding that EXPLAINs represented occasions where new information was introduced to dialogues without having been solicited by other participants.

It was recognised, however, that **disruption** similar to the *breakdowns* suggested by Wright and Monk (1989) and independently discovered by Newlands (1998), in which the presence of technology used within a human-

human communication is explicitly mentioned in the that communication, could also lead to an increase in the number of EXPLAINs produced.

It was expected that analysis of the *task orientation* of EXPLAINs would help to test which of these explanations was contributing most to any increase in the amount of words spent EXPLAINing because several subcategories of OFF-TASK orientation of *move* were intended to capture such *breakdown.* If an increase in the use of EXPLAINs was lead by an increase in OFF-TASK EXPLAINs that would indicate **disruption** was making a major contribution.

The initial results reported in Section 4.3.0, showed that there was no effect of the style of prompt on the number of EXPLAINs, but a significant effect on the amount of words spent EXPLAINing. This effect was considerable; the spoken prompt lead to 203.8 (45%) more words being uttered within EXPLAINs than the integral window.

The Three-way ANOVA used to analyse the number of EXPLAINs, with the style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker (2 levels: Agent and Customer) and *task orientation* of *move* (2 levels: ON- and OFF-TASK) as within subjects factors, was used to investigate further patterns in their production.

This ANOVA found both a significant main effect of role (F[1,28] 82.648, P: <0.001) with the agent producing on average 33.2 (152%) more EXPLAINs and a significant main effect of *task orientation* (F[1,28] 58.783, P: <0.001) with there being on average 39.8 (70%) more ON-TASK than OFF-TASK, as well as a significant interaction between role and *task orientation* (F[1,28] 107.303, P: <0.001). None of the other interactions were significant at greater than the 10% level.

The interaction between role and *task orientation* was not investigated in detail because the primary interest was in the effects of the style of prompt on the dialogues and it was unlikely to increase understanding of that. This interaction was more likely to be caused by the different communicative activities forced on the participants as a result of their individual tasks within the travel game; for instance on average the agent produced 31.1 (364%) more ON-TASK EXPLAINs than the customer, but only 2 (15%) more OFF-TASK (see Table 4/6) and this is more likely to be caused by the travel game and sales tasks forcing the agent to give information about products to the customer (ON-TASK EXPLAINs) while

OFF-TASK EXPLAINs such as INTERJECTIONs or comments about the mediating technology (ABOUT_TECH) would show much less influence.

	Agent EXPLAINs	Customer EXPLAINs
ON-TASK	39.7	8.55
OFF-TASK	15.2	13.2

Table 4/6 Mean occurrences of ON- and OFF-TASK orientations of EXPLAIN *moves* during the whole dialogues, by role of speaker.

To test whether the style of prompt was having any particular effect on the use of each *task orientation* of EXPLAIN, because there was such a large effect of *task orientation* even though there was no significant interaction between the style of prompt and *task orientation* (F[1,28] 1.766, P: 0.195), each *task orientation* of EXPLAIN was statistically analysed independently.

Two-way ANOVAs on the the ON- and OFF-TASK orientations of EXPLAIN, with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role as a within subjects factor (2 level: Agent and Customer), found neither of the main effects, nor their interaction to be significant at more than the 10% level for OFF-TASK EXPLAINs, but several notable effects on ON-TASK EXPLAINs.

There was a trend in the main effect of of the style of prompt on the number of ON-TASK EXPLAINs produced (F[1,28] 3.729, P: 0.064) with 13.7 (33%) more being produced when the prompt was spoken rather than being presented through the integral window. As would be expected due to the interaction between role and *task orientation* described above, there was a statistically significant main effect of role on the number of ON-TASK EXPLAINs (F[1,28] 122.218, P: <0.001); as was mentioned above, the agent produced 31.1 (364%) more ON-TASK EXPLAINs than the customer. There was no interaction between the style of prompt and the role of the speaker (F[1,28] 0.705, P: 0.408).

Although there was only a trend in the effect of style of prompt on the number of ON-TASK EXPLAINs, post-hoc Bonferroni T-tests were performed on the agents' and customers' productions to ensure that the difference was not predominantly lead by one speaker or the other.

No significant effect of the style of prompt was found in either case, though there was a stronger trend in the effect on customer ON-TASK EXPLAIN *moves* (Agent: df: 28, t: -1.546, P: 0.133) (Customer: df: 28, t: -1.897, P: 0.068). As can be seen in Table 4/7, although the increase in the number of ON-TASK EXPLAINs produced by the agent is larger than that produced by the customer, the percentage difference is greater in the customer's case. The agent produces 9.2 (26%) more ON-TASK EXPLAINs when their software employs a spoken prompt rather than integral window. The customer produces 4.5 (71%) more ON-TASK EXPLAINs when the agent's software employs the spoken prompt.

	Agent ON-TASK EXPLAINs	Customer ON-TASK EXPLAINs	Both Roles' ON-TASK EXPLAINs
Integral	35.1	6.3	41.4
Spoken	44.3	10.8	55.1

Table 4/7 Mean occurrences of ON-TASK EXPLAIN *moves* during the whole dialogues, by style of prompt and role of speaker.

This means that both agent and customer tend to be affected by the effect of the style of prompt on the production of ON-TASK EXPLAINs. In other words, the style of prompt affects the agent's conduct of their dialogue with the customer strongly enough that increased production of ON-TASK EXPLAINs is reciprocated by the customer.

With no major effect on the number of EXPLAINs produced having been found, interest has to be concentrated on the amount of words spent EXPLAINing.

The Two-way ANOVAs, with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer), which found there to be a significant effect of the style of prompt on the amount of words spent EXPLAINing accounting for 203.8 (40%) of the global 511 word increase (see Section 4.3.0) also found a significant main effect of the role of speaker (F[1,28]

76.461, P: <0.001) with the agent producing 300.7 (237%) more EXPLAINing words than the customer. The interaction between the two did not reach a level of significance greater than 10%.

The effects of role are not unexpected since the travel game task imposes different communicative requirements on the participants; for instance, the agent uses EXPLAINs to give information about potential destinations whereas the customer predominantly uses them to confer OFF-TASK information (see Table 4/6).

Even though there was no interaction between style of prompt and the role of the speaker, Bonferroni t-tests were carried out on the amount of words each role spent on EXPLAINs. A similar pattern to that in the number of ON-TASK EXPLAINs was discovered; although the style of prompt leads to a larger numerical difference in the amount of words the agent spends on EXPLAINs than the customer spends, the percentage difference is larger for the customer (see Table 4/8).

	Agent EXPLAIN words	Customer EXPLAIN words	Both Roles EXPLAIN words
Integral	365.7	87.0	452.7
Spoken	489.6	166.9	656.5

Table 4/8 Mean number of EXPLAIN words during the whole dialogues, by style of prompt and role of speaker.

There was no statistically significant effect of the style of prompt on the amount of words the agent spent on EXPLAINs (df: 28, t: -1.7, P: 0.10) even though the agent produced 123.9 (34%) more when using the spoken prompt than the integral window, accounting for 60% of the total 204 word difference due to EXPLAINs.

There was, on the other hand, a statistically significant effect of the style of prompt on the amount of words the customer spent on EXPLAINs (df: 28, t: -2.37, P: 0.025) with the customer producing 79.9 (92%) more when using the spoken prompt.

These findings about the amount of words each role spends on EXPLAINs, because the effect is more defined in the customer's use, again suggest that the effect of the prompt is permeating well into the agent-customer dialogue.

To test whether the effect of the style of prompt on the production of EXPLAINs was more extreme immediately after the prompt's presentation, one possible "sit-up-and-listen" effect mentioned in Section 4.2.2, counts of the number of EXPLAINs between the mention of the opportunity and its resolution were analysed with Two-way ANOVAs using style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer). This analysis found that only the main effect of role was significant at more than the 10% level ($F[1,28]$ 24.447, $\underline{P}$: <0.001), with the agent producing 4.4 (157%) more EXPLAINs than the customer.

Unusually for between the mention and resolution, a sizeable amount of the EXPLAINs were OFF-TASK. On average 5.4 (27%) of the EXPLAINs between the mention and resolution were OFF-TASK with the majority of these (3.6, 67%) being produced by the customer. For this reason, ON-TASK EXPLAINs between the mention and resolution were studied independently to test whether the style of prompt was having a particular effect on them.

A Two-way ANOVA, with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker (2 levels: Agent and Customer) as a within subjects factor, performed on the counts of ON-TASK EXPLAINs between the mention and resolution also found that only the main effect of role was significant at more than the 10% level. In this case, the agent on average produced 10.6 (515%) more than the customer

Both effects of role were presumed to be due to the different communicative tasks imposed on each role by the travel game, and without interactions with style of prompt were unlikely to point to any effects of the style, so were not investigated further.

Based on these results, it was concluded that there was no apparent "sit-up-and-listen" effect of the style of prompt.

In summary of the total effects of the style of prompt on the production of EXPLAINs, although the style of prompt had only a small and highly specific effect on the number of EXPLAINs produced, the effect on the amount of words spent EXPLAINing was more pronounced and widespread.

The only effect on the number of EXPLAINs was a trend towards more ON-TASK EXPLAINs being produced when the spoken prompt was used than when the integral window was used.

Since there was no statistically significant increase in the number of OFF-TASK EXPLAINs, this would indicate that there was no increase in the amount of **disruption** caused by the computer directly imposing itself in the dialogues (see Section 4.2.1.3; Example 4/5).

More importantly, there was a statistically significant effect on the number of words spent EXPLAINing which accounted for 203.8 (45%) of the global 511 word increase due to the spoken prompt. This increase, when there is no corresponding increase in the number of *moves*, suggests that the speakers are lead to be more verbose when volunteering information.

Taken in combination, these results suggest support for the explanation that the spoken style of prompt is **promoting** more conversation between the participants than the integral window.

Furthermore, in both cases, although the increase in the use of EXPLAINs was numerically larger for the agent, the customer's increase was more pronounced. This means that the effect of the prompt is permeating fully into the dialogue between the agent and customer even though it is only the agent who hears it.

4.3.1.4 QUERY-YNs

QUERY-YNs, even though they are used to gather information from the hearer, also serve to introduce new information to dialogues. For this reason, they were considered to be sensitive to the style of prompt **promoting** dialogue both because, if speakers prefer to volunteer information, there would be an increase in the number of QUERY-YNs, possibly to the detriment of the more demanding QUERY-Ws, and because QUERY-YNs provide more opportunity for verbosity as options are explained than *moves* such as QUERY-Ws or INSTRUCTs.

Consequently, the hypothesis was built that there would be more use of QUERY-YNs if the spoken prompt was **promoting** dialogue between the agent and customer.

H4 The spoken prompt will lead to more use of QUERY-YN *moves.*

In section 4.3.0, it was seen that there was no effect of the style of prompt on the number of QUERY-YNs produced despite an average 6.5 (20%) increase with the spoken prompt, but there was an effect on the amount of words spent QUERY-YNing, which accounted for 116.7 words (23%) of the global 511 word difference these analyses were being used to explain.

As with the other *moves*, these findings were investigated further to discover whether there were any more fine-detailed effects on the number of QUERY-YN *moves* used, either over the whole dialogues or between the mention of the prompt and its resolution, and to test whether any aspect of the agent-customer dialogues was leading the increase in words. As with EXPLAINs, the *task orientation* of *moves* was analysed because it might indicate whether **disruption** was contributing to any increase.

The Three-way ANOVA, with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker (2 levels: Agent and Customer) and *task orientation* (2 levels: ON- and OFF-TASK) as within subjects factors, performed on the counts of QUERY-YN *moves* produced in each consultation found several notable effects.

Apart from the lack of a significant main effect of the style of prompt (see Section 4.3.0; Table 4/2, there were significant main effects of both role ($F[1,28]$: 4.376, $\underline{P}$: 0.046) and of *task orientation* ($F[1,28]$: 255.268, $\underline{P}$: <0.001). The customer produced on average 4.75 (30%) more QUERY-YNs than the agent, and of the average 36.4 QUERY-YNs there were only 1.2 OFF-TASK.

On the other hand, none of the interactions were statistically significant at more than the 5% level, but those between the style of prompt and *task orientation* ($F[1,28]$: 3.533, $\underline{P}$: 0.071) and the role of speaker and *task orientation* ($F[1,28]$: 3.401, $\underline{P}$: 0.076) displayed trends.

The main effects of role of speaker and *task orientation* and their interaction are not interesting for the same reasons the effects on EXPLAINs were not investigated. The effect of role is likely to be due to the different communicative demands placed on each participant by their role; for instance it is likely the customer is producing more QUERY-YNs asking whether they can get to particular destinations. The effect of *task orientation* is due to there being so few

OFF-TASK QUERY-YNs, which also suggests that the trend in the interaction is almost entirely generated by the main effect of role.

Because of the trend in the interaction of the style of prompt and *task orientation*, each of the *task orientations* of QUERY-YN was analysed with Two-way ANOVAs taking style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer).

Although there were consistent effects of role, with the customer producing more than the agent in both cases, neither of the interactions between style and role was significant at more than the 10% level, and – more importantly – neither of the main effects of the style of prompt was significant at more than the 10% level (ON-TASK – F[1,28] 2.692, P: 0.112) (OFF-TASK – F[1,28] 2.386, P: 0.134). This would suggest that the trend in the interaction is being lead by the large main effect of *task orientation.*

The most important finding is that there were only 1.2 OFF-TASK QUERY-YNs on average, so **disruption** symbolised by the *breakdowns* Wright and Monk (1989) described and Newlands (1998) found, and which could be captured by OFF-TASK orientations of *move*, is apparently having little effect on the number of QUERY-YNs produced.

This suggested that the main effect of style on the amount of words spent QUERY-YNing strongly supported the explanation that the increase was due to the spoken prompt **promoting** more free, verbose dialogue between the agent and customer.

The Two-way ANOVA with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker as a within subjects factor (2 levels: Agent and Customer), which discovered the main effect of style of prompt on the amount of words spent QUERY-YNing (see Section 4.3.0, Table 4/3), found neither a main effect of role (F[1,28] 0.065, P: 0.801) nor an interaction between style and role (F[1,28] 1.404, P: 0.246) (see Table 4/9).

	Agent QUERY-YN words	Customer QUERY-YN words	Both Roles QUERY-YN words
Integral	113.5	143.4	256.9
Spoken	196.5	177.1	373.6

Table 4/9 Mean number of QUERY-YN words during the whole dialogues, by style of prompt and role of speaker.

Although there was no main effect of role or interaction between role and style of prompt, it was noted that, although both agent and customer spent more words QUERY-YNing with the spoken prompt than with the integral window; the difference in amount the agent spent was much larger than the difference in the amount the customer used.

The agent spent 83 more words on QUERY-YNs when the spoken prompt was employed and the customer spent 34, meaning that the increase in the amount of words the agent spent on QUERY-YNs was over twice the increase in the customer's.

In trying to explain this difference, unlike in the testing of the other hypotheses, it was analysis of the section of dialogues between the mention of the prompt and the resolution of the sale which proved most elucidating.

Because there were only 4 OFF-TASK oriented QUERY-YNs between the mentions and resolutions of all 30 dialogues and 3 of these were produced by one customer, which suggested they were a feature of an unusual ideolect, the analysis concentrated on ON-TASK QUERY-YNs.

A Two-way ANOVA with style of prompt as a between subjects factor (2 levels: Integral Window and Spoken Prompt) and role of speaker (2 levels: Agent and Customer) as a within subjects factor found there to be a in the main effect of style on the number of QUERY-YN *moves* produced ($F[1,28]$ 3.929, $\underline{P}$: 0.057) with the spoken prompt leading to on average 2.1 (44%) more ON-TASK QUERY-YNs (see Table 4/10).

Category	Integral Window	Spoken Prompt	ANOVA	P
QUERY-YN (ON-TASK)	4.7	6.7	F[1,28] 3.929	0.057

Table 4/10 Mean number of ON-TASK QUERY-YN *initiating moves* between the mention and resolution, by style of prompt with main effect.

There was also a main effect of role (F[1,28] 93.944, P: <0.001) with the agent producing on average 3.9 (433%) more ON-TASK QUERY-YNs than the customer, who on average produced less than one, and a minor trend in the interaction between the style of prompt and the role of the speaker (F[1,28] 3.026, P: 0.093) (see Table 4/11).

	Agent ON-TASK QUERY-YNs	Customer ON-TASK QUERY-YNs	Both Roles ON-TASK QUERY-YNs
Integral	3.9	0.7	4.7
Spoken	5.7	1.1	6.7

Table 4/11 Mean number of ON-TASK QUERY-YN *moves* between the mention of an opportunity and its resolution, by style of prompt and role.

Since all the agent's QUERY-YNs were ON-TASK and the difference in the number they produced accounted for roughly 90% of the effect of the the style of prompt, explaining the interaction trend, the simple effect of the style of prompt on the agent's production of QUERY-YNs was calculated.

This showed that there was a significant simple effect of the style of prompt on the number of QUERY-YNs the agent produced (F[1,28] 5.335, P: 0.028), with on average 1.74 (44%) more occurring when the prompt was spoken.

This in itself may seem like a small difference, particularly since the average length of a QUERY-YN was 9.15 words, which would not account for an important proportion of the large global differences observed, but closer analysis of the agent's use of QUERY-YNs showed such an assumption to be false.

Most importantly, it was found that, whereas the agent spent on average 85 words on QUERY-YNs between the mention and resolution when given spoken prompts, they used only 47 words when prompted through the integral window.

This means that their use of QUERY-YNs between the mention and resolution of sales opportunities accounts for 38 (78%) of the 49 word difference between the agents' and customers' increase in use of QUERY-YN words over the whole dialogues (see Table 4/9).

Although both the agent and customer spend more effort on QUERY-YNs over the whole dialogues, the agent's effort is highly concentrated immediately after the prompt.

One implication of this finding is that the conclusion that there was no "sit up and listen" effect, arrived at after the analysis of EXPLAINs between the mention and resolution, needs to be reconsidered.

Analysis of the way in which the agent used QUERY-YNs between the mention and resolution showed that the QUERY-YNs after the spoken prompt concentrated more on description of the products than on requesting a response from the customer.

This was most clearly demonstrated in patterns of the additional tags which are included during Conversational Games Analysis coding.

When a *move* has the strong form of one Conversational Games Analysis category but is clearly performing the function of another, the category of the utterance can be marked with an equals sign preceded by the form category and followed by the function category[6].

Example 4/11 provides an example of an EXPLAIN=QUERY-YN, i.e. an utterance functioning as a QUERY-YN which has the syntactic and intonational form of an EXPLAIN. That this sentence is constructed as, and understood to be, a QUERY-YN is indicated by the customer's REPLY-Y response.

---

6 This marking has been rejected in recent versions of Conversational Games Analysis because the function is paramount, but was used in this study because early coding showed it to be fairly common. When counting the number of moves produced, only the functional aspect was included.

```
->    • Agent
->    • Game 56 EXPLAIN=QUERY-YN
->       so again I think you'd be wanting the travel pass to
->       to make the most of your stay
->    • Move EXPLAIN=QUERY-YN

      • Customer
      • Game 57 OFF-TASK(ABOUT_SELF) EXPLAIN em
         [um]
      • Move OFF-TASK(ABOUT_SELF) EXPLAIN
      • End 57

      • Agent
      • Game 58 EXPLAIN em
         which again it's it's just a further ten pounds, ten
         pounds for the two of you, and that's for the duration
         of the three days that you you say you're going to be
         spending there
      • Move EXPLAIN

      • Customer
         right,
      • Move ACKN
      • End 58
         ok then, yeah
      • Move REPLY-Y

      • Agent
         ok,
      • M ACKN
      • End 56
```

Example 4/11 Example of agent EXPLAIN=QUERY-YN between mention and resolution of opportunity. (EXPLAIN=QUERY-YN marked with arrows) (Spoken Prompt, Sixth Trial, Third Prompt).

Between the mention and the resolution, on average 1.4 (29%) of the agent's QUERY-YNs were EXPLAIN=QUERY-YNs; but there was a higher percentage after the spoken prompt (1.8 *moves*; 32%) than after the integral window (1.0; 25%). This accounts for 48% of the increase in agent QUERY-YNs after the spoken prompt.

Further combing of the dialogues showed that a considerable amount more of the agent's QUERY-YNs between mentions and resolutions after the spoken prompt which were not EXPLAIN=QUERY-YNs had considerable explanatory portions, but narrowly avoided classification as EXPLAIN=QUERY-YNs by the inclusion of a short querying phrase.

Example 4/12 shows a QUERY-YN which is largely explicatory, saying why the product is being mentioned and why the customer should purchase it, but whose central function is questioning why whether the customer wants a travel pass in Dallas. This *move* would have been counted as having 32 words, which means the explanation has lead to it being over three times longer than an average QUERY-YN.

> *Agent
> **Game* 65 QUERY-YN (MENTION)
> **[em,] seeing as I've just looked you've had you've got two days in Dallas, would you like the travel pass for Dallas since you'd like to see a lot of that town**
> **Move* QUERY-YN (MENTION)

Example 4/12 Example of explanatory agent QUERY-YN between mention and resolution of opportunity. (32 words) (Spoken Prompt, Fifth Trial, Second Prompt).

The implication of these observations is that there is a "sit up and listen" effect of the spoken prompt in which the agent relates more information about the benefits of products to the customer.

At the lowest level, this finding supports the hypothesis that the spoken prompt will lead to more use of QUERY-YNS, and finding that the increase was entirely within ON-TASK QUERY-YNs reinforces the contention that this increase would be due to **promotion** of dialogue rather than due to **disruption** caused by badly handled utterances.

More interestingly, because almost half of the increase is due to EXPLAIN=QUERY-YNs this relates well to the finding that there was an increase in the verbosity of EXPLAINs over the whole dialogues. The prompt being spoken rather than written seems to **promote** the more open donation of information within the dialogues, which eyeball observation would indicate was due to the agent and customer describing the arguments for their decisions.

Finally, it is also worth noting that the increase in the use of QUERY-YNs was confined solely to the agent's use immediately after the presentation of the prompt.

Coupled with the information that this was the only *move* which showed any effect of the style of prompt between the mention and resolution of the sale[7], this suggests that the global increase in EXPLAINs which permeated both the agents' and customers' production, may have been lead by the agent's increased openness when reacting to the spoken prompt.

4.3.1.5 QUERY-Ws and INSTRUCTs

To ensure that no unforeseen effects of the style of prompt not predicted in the hypotheses were missed, the use of QUERY-Ws and INSTRUCTs was subjected to the same analysis as the other *moves*.

The only effect of the style of prompt on QUERY-Ws and INSTRUCTs which was notable was on the number of OFF-TASK QUERY-Ws produced over the whole dialogues.

A Two-way ANOVA with style of prompt (2 levels: Integral Window and Spoken Prompt) as a between subjects factor and role of speaker (2 levels: Agent and Customer) as a within subjects factor, which was carried out subsequent to finding a main effect of both role and *task orientation* but not of style on the production of QUERY-Ws, found a significant main effect of style on the number of OFF-TASK QUERY-Ws produced ($F[1,28]$ 5.814, $\underline{P}$: 0.023).

This effect was not statistically analysed further because so few OFF-TASK QUERY-Ws were produced and because the one *move* difference was unlikely to be contributing considerably to the overall 511 word increase due to the spoken prompt. (see Table 4/12)

---

[7] This has already been stated for CHECKs, ALIGNS and EXPLAINS. It is also true for the number of QUERY-W and INSTRUCT moves between the mention and resolution. When they were submitted to the same two-way ANOVA analysis as the other moves, neither showed a main effect of the style of prompt significant at more than the 10% level.

	Agent OFF-TASK QUERY-Ws	Customer OFF-TASK QUERY-Ws	Both Roles' OFF-TASK QUERY-Ws
Integral	0.3	0.3	0.5
Spoken	0.7	0.7	1.5

Table 4/12 Mean occurrences of OFF-TASK QUERY-W *moves* during the whole dialogues, by style of prompt and role of speaker.

Nevertheless there was a considerable ratio of difference between the two styles, with around 3 times as many being produced when the spoken prompt was used than when the integral window was used, so the way in which the OFF-TASK QUERY-Ws were used was evaluated.

Investigation of the subcategories of OFF-TASK *moves* which were being produced showed that 57% of them were TO_SELF, of which 82% were produced in the spoken prompted dialogues with each speaker producing exactly half. The agent and customer appeared to be using the *moves* differently, with the agent vocalising attempts to remind themselves what they were doing for the customer, and the customer vocalising their attempts to find destinations on their maps (see Examples 4/13 and 4/14).

• Agent
• *Game* 116 OFF-TASK(TO_SELF) QUERY-W em'
[um,] let me see where were we,
• *Move* OFF-TASK(TO_SELF) QUERY-W
• End 116

Example 4/13 Example of agent OFF-TASK(TO_SELF) QUERY-W (Spoken Prompt, Twelfth Trial).

• Customer
• *Game* 187 OFF-TASK(TO_SELF) QUERY-W em'
where's Portland, bollocks where's Portland,
• *Move* OFF-TASK(TO_SELF) QUERY-W

Example 4/14 Example of customer OFF-TASK(TO_SELF) QUERY-W (Spoken Prompt, Eighth Trial).

It would seem that the style of prompt leads to the agent and customer vocalising otherwise internal questions, but since the mean difference in the number of words spent on TO_SELF *moves* was only 3.27 this is contributing little to the overall 511 words increase.

4.4 Discussion

At the end of Chapter 4, it was speculated that the reason for the increase in the length of spoken prompted dialogues over the integral window was a change in the dynamic of the relationship between the agent and their computer. Making inferences from such varied work as Cotton and McCawley (1983, referred to in Baber et al., 1992) and Halliday (1989) it was proposed that the spoken prompt lead to the agent assigning their system a commanding role, whereas the integral window lead to more of an advisory role.

To test this proposal, a system for coding the dialogues was developed and applied. This system owed a large debt to Conversational Games Analysis for its origins, but included second dimension of coding which captured the *task orientations* of Conversational Games Analysis *moves*.

The most important findings from the application of this coding scheme were that the increase in words was strongly lead by the participants' use of two-functions of utterances; the description of new information and the asking of yes-no questions. There were 203.8 (45%) more words spent on EXPLAINs in spoken prompt consultations than in integral window consultations and there were 116.7 words (23%) more words spent on QUERY-YNs in spoken prompt consultations than in integral window consultations.

Although there was no interaction between the style of prompt and the amount of words each participant produced over the whole dialogues, the effect was found to be particularly concentrated in the agent's queries whether the customer wanted to buy the travel pass. These queries accounted for 38 (78%) of the 49 word difference between the agents' and customers' increase in use of QUERY-YN words over the whole dialogues.

The spoken prompt seemed to be compelling the agent to volunteer more information about their products, and this was affecting the whole nature of information sharing within the dialogues.

The use of in depth functional analysis of the process of the human-human dialogues had indicated what the influence of the spoken prompt was that lead to the longer dialogues, suggesting that such analysis could be used to inform the design of the interfaces.

Chapter 5

—

Experiment 3

—

A Comparison of Sales Prompting Styles in Simulated Customer Service Interactions with One-Way Video Communication

5.0 Introduction

In previous chapters it has been shown that the variation in the design of an interface used in tripartite communication can exert effects on the dialogues between humans involved in that communication and that the dialogues can be examined to determine what those effects are. The studies have only involved remote audio-only communication however, and the aim of the third study will be to investigate the effects of improved visibility between the two human participants.

This was lead by the observation that within previous work on tripartite communication, the highest degree of co-ordination between the three parties had been recognised in studies of doctor's practices where both humans are covisible and the patient can see the doctor interacting with their computer. In Greatbatch et al. (1993, see Section 1.1.2), for instance, patients were observed to attend to movement of doctors' hands over their keyboard using these to infer the occurrence of turn transition points (Sacks, Schegloff and Jefferson, 1974; see Section 1.3.1.1.1).

It was decided to investigate this by extending the findings from the second study (see Chapters 3 and 4). In that experiment it was found that in simulated customer service calls the presentation of sales prompts through speech rather than through lasting written means lead to more voluntary donation of information and longer consultations.

This would have both design and methodological implications. For design, it would indicate how differences in visibility between the human participants would interact with the two written styles of prompt (dialogue box and integral window) being presented on the computer's screen, potentially interfering with the agent-computer interaction, and the spoken prompt being presented through the agent's headphones, potentially interfering with the agent-customer dialogue. For methodology, it permitted a test of the value of analysis of the human-human dialogues in tripartite communication in isolation from the other interactions as the relationships between the three parties alter.

To determine the exact nature of the visibility which would be used, British Telecom was again returned to and discussions were held with members of their Future Customer Handling Systems group and Human Factors division. These discussions were used to make predictions about the way in which customer service was developing and how increased visibility would fit into the customer service relationship.

The most widely occurring observation was that there was a trend towards more use of remote customer service. Naturally, this was the basis of British Telecom's own customer service, but their call centres were also supporting other companies' customer service, independent agencies were being set up to provide similar services and an increasing number of companies were appearing who had no high street presence but conducted all their business over the telephone or internet. This suggested that the study of face-to-face customer service while still holding ecological validity, might have less future application than a more innovative remote situation.

On the other hand, this remote customer service was seen to have the disadvantage of distancing the company from their customers. One aspect of this was a loss of the ability to tailor service to particular customers, and various steps were being taken to counter this by keeping more detailed customer records and facilitating more personalized services. The second was a loss of the subjective feeling that agents, hence company, and customer were working together and not that the agent and company were playing a more dictatorial role. In psychological terms there was a loss of "social presence" (Short, Williams and Christie, 1976) or an increase in "psychological distance" (Rutter, 1987) between the agent and their customer (see Section 1.3.2.2.1).

Since there is some indication that video-conferencing can increase the feeling of social presence or decrease psychological distance (see Section 1.3.2.1.1) its deployment was an obvious way in which the latter could be overcome. Observation of various prototype video-conferencing systems developed within BT's human factors division or already in use by their ISDN customers indicated, however, that these were often targeted at very specific users, for instance corporate clients or distance learning, and were unlikely to be directly convertible to customer service.

Further discussion with the Future Customer Handling Systems group and management personnel suggested that – with the increased unification of telephone, cable television and internet service – one future possibility might be interaction with customer service agents through domestic television. In this situation, when contacting a customer service agent, that agent might be visible over the home television as well as audible. At the same time, it seemed likely that privacy concerns would mitigate against customers having a camera in place in their living rooms.

For the thesis this was attractive because it matched the asymmetry of computer use in the tripartite customer service situations with an asymmetry of visibility, whilst also providing an opportunity to observe the effects of variation in the visio-spatial and social relationships between the agent, computer and customer on their interactions.

The idea of using one-way video in domestic customer service was born.

5.0.1 Visibility in human-human dialogue

In Chapter 1, a brief review was made of the ways in which differences in media of communication interacted with synchronous human-human dialogue (Section 1.3.2). This concentrated on the effect of visibility, covering the known differences between audio-only, video-mediated, screened and face-to-face communication and the models which have been built of the continuum of these effects but little attention was paid to the functions the different degrees of visibility paid in this continuum.

Recently the functions of vision in dialogue have been taxonomized along two dimensions; the co-ordination of the process and the co-ordination of content (Whittaker and O'Conaill, 1997; see Figure 5/1). The co-ordination of process ranges from recognising when it is appropriate to begin and end dialogue, to selecting which speakers will follow each other and how the transitions are made. The co-ordination of content is even more wide-ranging, covering the monitoring of collocutors' displayed emotion and attitude, the confirmation of mutual understanding and common belief and the establishment of common reference to objects and events.

	Conversational Mechanism	Communication Function
Process Coordination	Turn-taking Cues	Determine who will speak, who will listen, and how transitions are made between these roles
	Availability Cues	Determine when to initiate or end a communication episode.
Content Coordination	Reference	Allow participants to identify the objects and events the jointly want to talk about
	Feedback Cues	Inform speaker of listener's understanding; contribute to maintenance of mutual beliefs and common knowledge between speakers and listeners
	Interpersonal Cues	Allow participants to infer the emotional stance, affect, and motivations of other people to what is being discussed and to other conversational participants

Figure 5/1 A Taxonomy of Conversational Mechanisms and Their Communicative Functions.
Reproduced from Whittaker and O'Conaill (1997; p. 26)

The means in which visibility is used are varied, covering such behaviours as direct eye-contact (often called "mutual gaze"), observation of facial expression, posture and body movement and observation of the shared environment. Whittaker and O'Conaill (1997) provide numerous examples of the ways visibility supports each conversational mechanism in their taxonomy; a sample follows.

Turn-taking is one of the longest recognised functions of visibility in dialogue. The psychologist Kendon (1967), after on an "exploratory" study, referred to it as the regulatory function and noted that listeners would maintain visual attention on speakers, but that speakers would look away from their partners when speaking for longer than a few seconds but tended to return their eyes to their partners when approaching a turn transition point, and that if their gaze did not return there would typically be a delay in transition[1]. From the listener's point-of-view, Duncan and Fiske (1977) suggested that listeners may monitor the speaker's gestures since their presence will prevent turn taking attempts and cessation will indicate a turn transition point.

[1] Further examples of various media's effects on turn-taking were provided in Sections 1.3.2.1.3.2.

The ethnomethodologists Schegloff and Sacks (1973) first recognised the ways people used their observation of partners to decide whether to initiate and close conversations.

Kendon (1967) also recognised visibility's role in sharing interpersonal and feedback cues which he termed "expression" and "monitoring". It was often difficult to distinguish the influence of interpersonal cues from feedback in early studies however (Rutter, 1987), and since Clark's proposition of the importance of common ground (Clark and Brennan, 1991; see Section 1.3.1.2.2) the majority of work has concentrated on feedback and reference.

The best evidence that visibility can play a role in the transmission of emotion, and that the exchange of affective information can play a role in dialogue, comes from Morley and Stephenson (1969, 1970) who found that discutor's own beliefs played an increased role in simulated conflictual negotiation when discussions were face-to-face than when they were audio-only, and that this was achieved through more personalization of the dialogues (Morley and Stephenson, 1977)[2]. Even these studies could be flawed, however, if domain knowledge was not equally shared between the subjects; and it might be expected that partisan subjects would have more domain knowledge than non-partisan ones.

Feedback, which is used to show that the listener understands or agrees with what the speaker is saying, can be conveyed visually through head nods (Kendon, 1967) or even just showing continued attention with gazes at the speaker (Clark and Brennan, 1991).

Co-reference is supported by visibility in two ways, interpretation of partners' attention and direct observation of the environment.

Interpretation of a co-conversant's focus of gaze permits the regarding partner to infer which objects are salient to the observed party. This can be used to ensure that the speakers are talking about the same objects (Clark and Marshall, 1981).

Direct observation of the shared physical environment in itself reduces the ambiguity of dialogue, greatly simplifying the task of discussing or manipulating complex objects (Clark and Marshall, 1981; Gaver, Sellen, Heath and Luff, 1993, see Section 5.0.2.2). For instance, a speaker saying "his dogs" could be inferred

---

[2] See Sections 1.3.2.1.2 and 1.3.2.1.3.3 for more detail.

to be talking about his shoes if the speaker can be identified as one with a particular dialect and there are no animals present.

5.0.2 One-way video

One-way visibility between partners and locations has been used to research all of these functions. The earliest experimental work concentrated on how conversants used what they could see of each other to coordinate their dialogues and excluded reference to physical objects, more recent work has included controlled access to one conversant's physical environment.

5.0.2.1 Personal Visibility

Early studies of gaze in dialogue used one-way communication to discover the relative importance of its various functions, but it was soon judged to be too artificial for such a task (Rutter, 1987) and until recently there has been little interest in its effects.

These studies were all carried out by Michael Argyle at the Social Skill research group at Oxford (Argyle, Lalljee and Cook, 1968; Argyle, Ingham, Alkema and McCallin, 1973)[3]. He and his collaborators used one-way communication in a number of studies which attempted to separate the three functions of visibility which they recognised; feedback, turn-taking (their "synchronisation") and interpersonal cues (their "affiliative balance").

In one study reported in Argyle, Lalljee and Cook (1968)[4], subjects were given seven brief interviews by a researcher in each of which the visibility of the face was altered. In one interview neither face was obscured but in the remainder either the subject or the interviewer wore dark glasses, a large white cardboard mask which covered the whole face apart from the eyes or both. Their subjects rated each condition for comfort and their ability to judge the interviewer's reactions.

---

3 Goldberg and Mettee (1969) also used one-way visibility, with speakers addressing their partners through a slit and not always being visible. Their experiments only involved monologues.

4 Three studies were reported, only two of which involved one-way visibility. The third involved both subjects having only their eyes visible, only their face visible, only their body visible or nothing visible.

The subjects' ratings declined in the following order: open face, dark glasses, mask and both mask and dark glasses. This order remained the same whether it was the subject or the interviewer who wore the glasses, but was more marked when the interviewer wore them[5].

In the second study of one-way visibility reported in Argyle, Lalljee and Cook (1968), both conversants were subjects and they were asked to discuss five topics openly so that they were on an equal footing. In this case, the visibility of only one of the subjects was varied but the conditions used in the first study were altered; the use of mask and dark glasses together was removed, but conditions with a one-way screen and with no vision were added. The distance between the subjects was also varied, but in all cases the subjects were in adjacent rooms and audio communication was via an intercom system. As well as being asked to rate their comfort and ability to judge their partner's reactions they were additionally asked to rate their ability to communicate their own ideas and feelings and how easy conversation was.

Essentially the same pattern of obscurity discovered in the first study was replicated; dark glasses were rated less favourably than an open face, the mask less favourably than dark glasses and the one-way screen less favourably than the mask. The situation was more complex when neither speaker could see each other, however, and after extensive consideration of the subjects' ratings, in the light of each rater's visibility, Argyle et al. (1968) concluded that subjects preferred to be visible when the other was visible, but invisible when the other was invisible. Being on the wrong side of the one-way screen was rated least favourably of all. A briefly reported analysis of the dialogues found that obscured subjects spoke more than visible ones, producing more words per minute, but no more.

It could be concluded from these one-way studies that the asymmetry of visibility lead to discomfort for the fully visible party, and that they felt it was harder to coordinate with the obscured partner, but the limited analysis of the process of the dialogues were unable to assess the relative impact on different coordination mechanisms. They depended on the third, two-way, study in the article to make inferences about these mechanisms.

---

5 Throughout their work, sex of speaker was found to interact with the conditions of visibility, but these findings will not be reported because this may have been influenced by sexual role stereotyping at the time of the experiments.

In a later study, Argyle, Ingham, Alkema and McCallin (1973), dialogues using the one-way screen were gathered and analysed in closer detail, in an attempt to separate the effects of process and content coordination . In this case, two subjects delivered monologues or discussed various topics with each spending half the trials obscured and the other half visible, but in no case both being visible or obscured at the same time. This replicated an earlier study (Argyle and Ingham, 1972) in which visibility was normal.

The amount of looking between the subjects was analysed,with looking being rated as gazing at the partner's face, and several important findings were made. In the Argyle and Ingham (1971) study it had been found that subjects spent around 50% of their time looking at their partners, and comparison were drawn with this finding.

The first finding worth mentioning, which validated the study's other findings, was confirmation that listeners spent more time looking than speakers. The more relevant findings related to the effects on the amount of looking due to the different conditions of visibility.

The primary finding was that subjects spent more time looking at their partners when they could see them. However, even those whose partners were obscured spent around 25% of the trials looking at their partners[6]. On the other hand, subjects who could see their partners spent around 67% of the trials looking at their partners; more than Argyle and Ingham (1971) found covisible conversants to spend. Subjects spent around three times as much time looking at visible partners than invisible partners whether they were listening or talking.

The reason they provide for more gazing at visible subjects by non-visible is a reduction in feeling that this will make partners uncomfortable, an "avoidance" force in their terms. It has to be presumed that the extra gaze is used for increased

---

6 They spent 23% during dialogues and 27% during monologues. This findings must be considered in the light of the exact experimental setup -- the area of the partner's face was marked on the one-way screen so that subjects who could not see their partners knew where to look. Argyle et al (1973) report that their subjects "definitely, deliberately and normally" (p. 27) which may suggest that the marked face encouraged them to look.

content gathering by the subject who can see their partner[7]; the third, fourth and fifth mechanisms in Whittaker and O'Conaill's (1997) taxonomy (reproduced in Figure 5/1).

Another pertinent demonstration of the use of gaze for content coordination was its use by subjects who could not see their partners. This gaze was used more while listening (35%) than while talking (19%) and it was inferred that the subjects were aware of their visibility and used it to support feedback; the fourth mechanism in Whittaker and O'Conaill's (1997) taxonomy (reproduced in Figure 5/1).

The use of monologues and discussion was intended to cancel gazing for synchronisation of turn-taking, because no such synchronisation is necessary in monologues. The gazing of subjects who could not see their partners was not affected by monologue or discussion, but those who could see produced more in conversations than monologues. Most of this was produced while the subjects who could see were talking.

Argyle et al. (1973) concluded from this that turn-taking, the first mechanism in Whittaker and O'Conaill's (1997) taxonomy (reproduced in Figure 5/1), was primarily supported by looking while talking to gather information about the listener's gaze behaviour. They did not, however, provide any evidence that turn-taking was more fluent by visibility and it is possible that interaction with the different demands on content coordination in monologues and dialogues may have played a part in the increased use of gaze.

As was said in the introduction to this section, after Argyle's studies little attention was paid to one-way video's use in collaboration until recently. Its reappearance has not been driven by efforts to build theories of communication however, but has followed the development of new technologies; particularly "media spaces".

These media spaces are "*computer-controlled networks of audio and video equipment intended to support synchronous collaboration*" (Gaver, 1992). In its

---

7 Gaze can be used purely to elicit an avoidance reaction, and the work of Ellsworth (eg Ellsworth, Carlsmith and Henson, 1972, reported in Rutter, 1987) shows that unattributed gaze, or at least staring, can achieve this, but it would seem unlikely that the extra gaze was intended to make the observed feel uncomfortable in these studies! On the other hand, the camera appears to have been perceived as a stare.

most basic form, every member of a collaborating team has a video monitor and camera and a microphone and loudspeaker in their offices as well as their desktop computer, all linked to their collaborators through a high speed network. The broad intention is to support informal communication between physically separated co-workers, not just focused collaboration.

Heath and Luff (1992b) observed one-way connections being used in a media space at Xerox EuroPARC for making "glances" into collaborators' offices to see whether they are available for communication. This was used both to see whether people were in their offices and to avoid interrupting them if they were busy, and is an instantiation of the use of visible availability cues; the second mechanism in Whittaker and O'Conaill's (1997) taxonomy (reproduced in Figure 5/1).

5.0.2.2 Visibility of Environment

Further work on the EuroPARC media space saw the development of the Multiple Target Video (MTV) system (Gaver, Sellen, Heath and Luff, 1993), which turned the asymmetry of video-mediated communication noted by Heath and Luff (1992b) into a feature. The system permitted gross anisotropy of vision so that what one person saw need bear no relation to what their collaborators saw. Although the system did not preclude sharing of mirror image views, these were rarely used, for instance mutual face-to-face contact was only used for around 2% of the time[8], and it can therefore be considered as the use of two independent one-way views.

The MTV system permitted each subject access to four views of their collaborator's office. Both could choose to look at their partner's face, their partner's desktop and an "in-context" view which showed the relationship between their partner, their desk, their video monitors and the dollshouse if it was present. The final view differed. One subject could choose a bird's eye view of their partner's office while the other could choose to look through a camera focused on a dollshouse which was present in their partner's office. Both subjects also had a smaller feedback monitor on which they could see what their partners were looking at.

---

8 This low a figure was undoubtedly influenced by the highly visual nature of the task their subjects were performing.

The multiple cameras were intended to improve collaboration by giving the ability to share objects and to explore the collaborator's environment, so placing the collaborator and task within a context, overcoming the constraints of the fixed field of view of a single fixed camera. This recognised Gibson's (1979) concept of affordances; particularly his contention that movement was fundamental to perception (Gaver, 1992; Gaver, 1996).

In an experiment, the subjects were given two tasks; one cooperative and one competitive. In the cooperative task, they had to draw each others' rooms. In the competitive, the subjects were given different goals in arranging furniture around the dollshouse.

Their intention was both to investigate the MTV system's ability to support these visually based activities, which would not be supported by fixed face-to-face view video cameras, and to investigate the impact of choice of view in the competitive task because of visibility's well recognised impact on negotiation (see Section 1.3.2.1.0).

Their subjects used all four views available to them, but tended to use one view for long periods but make short connections to others and switched more in the cooperative task than in the competitive. The location specific views, bird's eye and dollhouse, were clearly used when most effective; the remote subjects used the dollhouse view for 71% of the time during the competitive task, and the subjects with the bird's eye view used it for 50% of the drawing task. Of the shared views, by far the most used were the "in-context" and "desk view" which accounted for 48% of the total time on average. As was commented earlier, face-to-face views were used very infrequently, each subject used the other's face view only around 11% of the time.

Even though the visual nature of the task makes direct comparison between this and earlier studies of negotiation impossible, it is worth noting that there were differences in the ways the two tasks were accomplished. Some subjects reported that they avoided the face-to-face view when trying to persuade their partners and some were seen to hide furniture out of the field of view of the cameras if they did not want it to be used.

Although the primary use for the alternative views seemed to be facilitating co-reference, the interviews with subjects and unstructured analysis of dialogues suggested that the alternative views were being used for more than just that, and indeed that even the four views available were far from ideal for that purpose. The clearest conclusion was that people preferred to use video to gather information related to their tasks rather than for mutual visibility per se. Unfortunately the analyses used do not permit definitive statements about the building of co-reference with video, but this is largely corrected by a study by Kraut, Miller and Siegel (1996).

Kraut et al. (1996) compared an unusual completely one-way video system to audio only communication, to investigate its effect on co-reference.

They gave unskilled subjects several repair tasks to do on a bicycle, which they either had to do on their own referring to a manual or with the help of an expert. When the expert was available, they either communicated by audio-only or the subject wore a helmet-mounted camera which allowed the expert to see what the subject's head was directed towards; this helmet was worn in all conditions and had a miniature screen in front of the right eye on which the manual or the output from the video camera was displayed.

The expert's aid halved the time to complete the task and improved the quality of the repairs. The use of the helmet video did not improve either the time or the quality[9], but did affect the way in which the subject and expert coordinated their activities. Using a task-specific coding scheme, they found that without the video, subjects described the state of their task more[10], and that when there was video, agents were more likely to follow customer descriptions of their task with help. The explicit descriptions appear to have been treated as requests for help when both expert and subject could see the task. Closer analysis showed that with video, the experts volunteered more help and were more likely to elaborate on their instructions, presumably seeing when the subjects were approaching difficulties and how the subjects had interpreted their instructions. The expert was also more explicit in their descriptions when video was not used.

---

9 Given that there was no conflict this may not be surprising, even though this task is very different from those in which differences between cooperation and negotiation have been found.

10 They also found that the expert acknowledged the worker's utterances more when video was not used, but this appears to have been a side-effect of the increase in the customer's descriptions; there were slightly more "speech-acts" when there was no video, the difference in the number of "speech-acts" used by the agent and the customer was greater when they used the video; the customer produced 10 more "speech-acts" and 15 more descriptions while the agent only produced 6 more acknowledgements.
(For more discussion of the use of acknowledgements in such analyses see Section 4.1.2.1)

Although the experimenters note that the cameras are not a substitute for real co-presence, it can still be concluded that suitable video-mediation does improve co-reference to physical objects; the third mechanism in Whittaker and O'Conaill's (1997) taxonomy (reproduced in Figure 5/1).

5.0.3 Hypotheses

The aim of this study was to see how the one-way video link which afforded the customer vision of the agent interacted with the effects of the style of prompt.

Review of the predictions which lead to the hypotheses built in Chapter 3 (see Section 3.0.3) gave no indication that those predictions were flawed, so the majority of the hypotheses were inherited into this study.

On the other hand, no indication of empirical support was found for a number of those hypotheses; and they will not be included. For instance, the only indication that there was a short term effect of the style of prompt was found after the extensive Conversational Games Analysis described in Chapter 4, and not during the testing of such features as perceivability (number of prompts mentioned to the customer) and impetus (words between prompt and mention). This suggested that the style of prompt's effects in a tripartite communication were not those which would be predicted from previous studies of stimulus-response reactions in human factors, but instead those which depended on more psychologically derived arguments.

The only hypothesis about local effects of the style of prompt which will be repeated is that about local PERFORMANCE, which measured the number of prompts which were converted into sales and was intended to provide a direct indication of the agent's success.

The prediction made about this were that the integral window would lead to more sales than the other two styles because the agent would find it easier to introduce the sales opportunity at a more beneficial time and in a more considered manner.

This will be numbered the first hypothesis in this chapter, but was the fifth in Chapter 3[11].

H1 {5} The integral window will lead to more sales than the dialogue box and integral window prompts.

The majority of the global hypotheses were supported, and therefore they will all be tested again.

For ease of reference, the predictions on which they were based will be repeated. These were essentially based on the reasoning that the style of prompt could have two types of global effect; technical and interpersonal.

Technically, it was predicted that, if the style of prompt caused disturbance to the ongoing travel task, this would lead to longer dialogues overall because that disturbance would have to be overcome. For instance, if the style of prompt immediately interrupted the agent's interaction with their system, as the dialogue box did, there would be immediate delay as the agent overcame this interruption and potentially more as they regained their original position in their tasks. Similarly, because the spoken prompt made no attempt to synchronise its presentation with the agent and customers' turns, it could interfere with the agent's audition causing him to have to ask for information to be repeated before or after the sales' resolution.

Predictions about the style of prompt's interpersonal impact were based on the findings of Cotton and McCawley (1983; referred to in Baber, 1992) and the analysis of Halliday (1989). The former had predicted that giving overly human speech to a machine could lead to the machine being attributed too much intelligence. The latter had suggested that speech conveyed information in a different way from writing; implying activity rather than stasis. The prediction from these was the agent's whole conception of their computer system's role might change, with the spoken prompted systems being considered commanders and the written prompted systems being advisors, and that this could lead to a sea-change in the interaction between agent, customer and computer.

The hypotheses based on these predictions were as follows:

---

[11] Throughout this chapter when hypotheses are repetitions of those from Chapter 3 their original number will be indicated within curly brackets in their label.

H2 {6} The customers will have a more positive perception of the consultations in which the agent uses the integral window than the consultations in which the other two styles of prompt are employed.

H3 {7} The dialogue box and spoken prompting styles will lead to longer consultations than the integral window.

H4 {8} More turns will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

H5 {9} More words will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

Two other measures are worthy of comment; the number of airports included in the itinerary and the turn-taking behaviour.

The number of airports included in the itinerary was a global measure of customer PERFORMANCE about which no hypotheses were made. This was inferred from previous researchers' findings that the PERFORMANCE cooperative tasks such as the itinerary building were insensitive to technologically based variation in the interpersonal conditions under which they took place. The measure was taken again, because of its relationship to customer PERFORMANCE, but as before no hypotheses were built.

Turn-taking behaviour was also coded, but again no hypotheses were made. As was discussed in Sections 1.3.2.1.3.2 and 3.0.3.3.4, turn-taking has been found to be sensitive to variation in the situation of dialogues as well as the participants both detecting facilitation and disruption. Although it has been suggested that the presence of a computer in a tripartite communication is considered during the human participants' coordination of turn transitions (Greatbatch et al., 1993), because the measure taken was the well-respected counting of overlapping and interrupting speech which only occurs during around 10-20% of spontaneous fluent conversation it was considered unlikely that the variation in the style of prompt would lead to such a gross difference in the dialogues as is generally required to be detected.

The presence of the video was expected to alter the degree of any effects, but not their causes, permitting the customer to overcome difficulties by seeing the agent's visual distraction but causing more frustration as the presence of otherwise invisible distractions became evident.

5.1 Task and Apparatus

5.1.1 Task and Apparatus

The experimental setup exactly duplicated that used in the second experiment (see Section 3.1), with the addition of a one-way video link between the agent and customer.

The task remained to be for the customer to organise a trip around the United States attempting to reach as many airports as possible within the bounds of the "Travel Game", and for the agent to sell a local travel pass to the customer when alerted by either a dialogue box, integral window or spoken prompt. No changes were made to the software the agent used, nor to the agents' and customers' handouts and questionnaires.

The video link between the agent and customer permitted the customer to see the agent but not the agent to see the customer. This simulated a future customer service situation where a "call" to the service agent could initiate a simultaneous video link to the agent's office through the initiator's home television.

In addition to the same audio-link used in the audio-only study, subjects were also provided with a direct video feed between a camera placed just above the agent's VDU and the customer's monitor. No digital processing of the signal took place, to this more closely simulated a broadcast quality transmission than current computer network based video-conferencing systems, removing any effects of delay or lost packets which have been shown to have effects on coordination in a number of studies (e.g. O'Conaill, Whittaker and Wilbur, 1993; see Sections 1.3.2.1.3.1 and 1.3.2.1.3.2)

The agent's camera was positioned at their eye level and captured their head and torso against the background of a neutrally coloured sheet of cloth. The image was sent to a small 5" monitor placed to one side of the camera, so the agent could see how they appeared, and to the customer's larger 15" monitor. The customer's monitor was directly in front of them and raised in such a way that if the agent looked directly into the lens of their camera they would appear to be looking directly at the customer's face.

The head and torso view was used because it has been argued to permit more awareness of the viewed person's environment and attentional focus, what Daly-Jones, Monk and Watts (1998) have called "Gaze Awareness". In this study, it was obvious to the customer whether the agent was directing their gaze towards the camera, their computer screen or their off-line resources.

Audio recordings were made of the participants' dialogues. Video recordings were made of the images from both the agent's camera and a second camera set off to the side of the customer. This second camera was intended to capture the angle of the customer's head, giving some indication of whether they were looking at their desktop or monitor.

5.1.3 Subjects

Ninety subjects, all students in the Strathclyde region, completed the task. 45 of these were trained and served as agents and 45 served as customers. All subjects were paid £3 for taking part and were aware that a £15 prize was available to them. The agent's prize was available to that agent who sold the most travel passes and the customer's prize was available to the customer who fitted the most airports into their itinerary. Although agents were aware of the winning criteria for both themselves and their customers, the customer was only aware of the prize available to themselves.

Two additional pairs of subjects participated but were not included in the analysis. One pair did not manage to trigger four sales prompts. One pair did not complete training because the agent found the one-way video connection too uncomfortable.

5.1.4 Procedure

Training for the agent and customer was the same as in experiment 2 (see Section 3.1) with the addition of a brief introduction to the video link.

This was given to the agent once they had been trained in how prompts would be given and how to respond to them.

They were shown how their monitor would show them how they appeared to the customer, but that they would have no corresponding view of the customer, and that if they wanted to appear to be paying attention to the customer they should look into the camera's lens. They were given no more guidance in when to look into the camera than "if you want to make eye contact with the customer."

Customers were introduced to their monitor after the agent had been trained, during the period when questions about their instructions were answered. They were introduced to the video link with the line "If you want to see what your agent's doing, you can see him in the screen.".

During the majority of the training the video link was inactive but immediately before the trials began subjects were involved in a test of the orientation of the agent's camera, the clarity of the customer's picture and both participants' audio levels. This served as limited familiarisation with the video link.

5.1.5 Vignette

As with the previous experiments, this one was also found to be immersive, with both agent and customer subjects fitting into their roles well (see Example 5/1).

*Agent
And where would you like to go next, Sir

*Customer
[em] down to Portland please

*Agent
And the departure times for Portland your travelling time's twelve hours, and departure times are ten am seventeen hundred hours or eighteen hundred hours

*Customer
[uh] eighteen hundred hours please

*Agent
Ok and you're so that's three days in Seattle and then you depart for Portland

*Customer
right

Example 5/1 Vignette demonstrating participants' immersion in their roles during video-mediated prompting task. (Dialogue Box Prompt, Fourth Trial)

The agents were not coached to refer to the customer with "sir" or "madam" yet many did, and the repetition of instructions to the customer before they are finalised is standard customer service practice.

5.2 Results

As in previous chapters, measures of PERCEPTION, PERFORMANCE and PROCESS were taken. Apart from there being no analysis of local effects of the prompt, these were the same as those used in the audio-only comparison of prompting styles.

Multiple measures were taken, but the specific hypotheses to be tested were as follows.

H1 {5} The integral window will lead to more sales than the dialogue box and spoken prompts.

H2 {6} The customers will have a more positive perception of the consultations in which the agent uses the integral window than the consultations in which the other two styles of prompt are employed.

H3 {7} The dialogue box and spoken prompting styles will lead to longer consultations than the integral window.

H4 {8} More turns will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

H5 {9} More words will occur over the whole dialogues when the dialogue box and spoken prompting styles are used than when the integral window is used.

PERCEPTION was measured using the same 25 item questionnaires given to subjects in the audio-only comparison of prompting styles (see Section 3.2.2.1.1). These were presented immediately after the consultations had finished.

PERFORMANCE was measured by the criteria of how many airports the customers managed to fit into their itinerary, how many of the four prompts the agent received they converted into sales and the time taken to complete the consultations.

PROCESS was measured by the number of words, turns and words per turn and by the amount and type of overlapping speech between turns during the agent-customer dialogues.

For more details about these measures, including the reasoning behind their inclusion, see Section 3.2.2.1.1 for PERCEPTION questionnaires, Section 3.0.3.3.2 and 3.0.3.2.5 for PERFORMANCE and Sections 3.0.3.3.3 and 3.0.3.3.4 for PROCESS.

5.2.1 PERCEPTION

The questionnaires used had three sections measuring attitude towards the consultation; one containing 7 point scales based on Osgood's semantic differential scores, one containing 5 point scales based on scales of customer satisfaction derived during British Telecom's Customer Interface Quality

Measurement ("CIQM") project (Marle and Coyle, 1995), and one containing questions of varying styles directly addressing features of the consultation and communication with their partner.

The Osgood section asked for ratings on the following scales:

- Unpleasant -- pleasant
- Still -- lively
- Tender -- Tough
- Bad -- Good
- Passive -- Active
- Weak -- Strong

The CIQM section asked for ratings on the following scales:

- Impolite -- polite
- Inefficient -- efficient
- Unfriendly -- friendly
- Unassuring -- assuring
- Incompetent -- competent
- Unsympathetic -- sympathetic
- Unhelpful -- helpful

The other questions demanded scalar or binary responses to the following questions:

- How easy was it to hear your partner?
- How easy was it to make changes in your itinerary?
- How satisfied were you with the final outcome of the consultation?
- Did you find the conversation natural?
- Did you feel that you had lost touch with your partner at any time?

The customers' responses to these scalar questions were analysed using Analysis of Variance, which was considered appropriate for the same reasons given in Section 2.2.1.

Several scales showed a statistically significant effect of the style of prompt; the Osgood "Bad-Good" scale and the CIQM "inefficient-efficient" and "unfriendly-friendly" consultation scales. None of the others were significant at more than the 10% level.

On the Osgood "Bad-Good" scale, all 15 of the consultations in which the integral window prompt was used were rated "Fairly good", "Good" or "Very Good" (points 5 to 7 the scales), whereas the consultations in which the other two styles of prompt were used both received only 8/15 ratings in this segment of the scale. A One-way ANOVA with style of prompt as a between subjects factor (3 levels: dialogue box, integral window and spoken prompts) found this to be statistically significant beyond the 5% level (One-way ANOVA, $F[2,42]$ 5.573, $\underline{P}$: 0.007) and post-hoc Tukey's HSD tests showed significant pairwise comparisons between the dialogue box and integral window styles (Tukey's HSD, $\underline{P}$: 0.040) and between the spoken prompt and integral window styles (Tukey's HSD, $\underline{P}$: 0.008).

On the CIQM "inefficient-efficient" consultation scale, 14/15 of the integral window consultations were rated "Efficient" or "Very Efficient" (points 4 and 5 on the 5 point scale), 10/15 of the dialogue box prompted consultations were rated in this segment and 7/15 of the spoken prompt consultations were rated in this segment. A One-way ANOVA with style of prompt as a between subjects factor (3 levels: dialogue box, integral window, spoken prompt) found this to be significant at the 5% level (One-way ANOVA, $F[2,42]$ 4.536, $\underline{P}$: 0.016) and post-hoc Tukey's HSD tests showed that the only significant pair-wise comparison was between the spoken prompt and integral window consultations (Tukey's HSD, $\underline{P}$: 0.012).

On the CIQM "unfriendly-friendly" consultations scale, all of the consultations where the integral window prompt was used were rated "Friendly" or "Very friendly" (points 4 to 5 on the scale), 14/15 of the dialogue box prompted consultations were rated in the same segment but only 9/15 of the spoken prompted consultations received such a complimentary rating. A One-way ANOVA with style of prompt as a between subjects measure (three levels: dialogue box, integral window and spoken prompt) found this to be significant at the 5% level (One-way ANOVA, $F[2,42]$ 3.234, $\underline{P}$: 0.049) and post-hoc Tukey's HSD tests showed that the only significant pair-wise comparison was between the spoken prompt and integral window consultations (Tukey's HSD, $\underline{P}$: 0.045).

From the analysis of the questionnaire data it can be concluded that the customers do detect the effects of the different styles of prompt, and as per the hypothesis (**H2 {6}**) perceive the integral window prompted consultations more favourably than the other two. The spoken prompt seems to be affecting their PERCEPTION more strongly than the dialogue box, being considered generally worse, less efficient and less friendly than the integral window consultations. This is slightly contrary to the expectations that the dialogue box would be considered less favourable than the spoken prompt but in line with the findings from the audio-only study of prompting style even though present in PERCEPTION not PERFORMANCE or PROCESS.

The fact that it is present in this one-way video study – but not in the audio-only – would suggest that the customer is better able to perceive the agent's distraction from the customer's intended building of an itinerary when able to see them, and that this distraction has a detrimental effect on the customers' PERCEPTION of their consultations. Although it is impossible to draw conclusions from null results, it is worth commenting that the statistically significant effects of the style of prompt were found in customers' ratings of the *consultation* on the CIQM scales not of the *agent*. This hints that the customer is able to distinguish between the quality of their dialogue with their agents and of the tripartite interaction as a whole. Third-party disturbance of the agent-customer dialogue seems to be more strongly felt than disturbance within the visual written interaction between the agent and customer.

5.2.2 PERFORMANCE

PERFORMANCE was measured by three means; the number of prompts to sell the customer a local travel pass which the agent converted into sales, the number of airports the customer and agent managed to fit into their itinerary, and the time that the whole consultation took. These are, respectively, indicators of the agent's success the customer's success and the overall success and each participant knew the value of each to themselves if not to their partner (see Section 3.1.2). The customer was to receive a prize if they included more airports than any other, the agent was to receive a prize if they sold more than any other, and both were limited to fifteen minutes in which to achieve this, though it was not impressed on either as strongly as it would have been in contemporary real world telephone customer service.

Specific hypotheses were made about two of these measures (**H1 {5}** and **H3 {7}**) and the analysis of all three are presented in Table 5/1 and discussed in the following paragraphs.

Measure	Dialogue Box	Integral Window	Spoken Prompt	ANOVA	P
Sales Made	1.5	2.1	1.9	F[2,42] 0.5976	0.5548
Airports Planned	10.7	10.2	10.7	F[2,42] 0.3636	0.6973
Time Taken	717.7	767.3	797.4	F[2,42] 0.9902	0.3800

Table 5/1 Summary table of PERFORMANCE measures by style of prompt.

Analysis of Variance showed no statistically significant effect of the style of prompt one any measure of PERFORMANCE.

Unlike in the audio-only study, the integral window was producing on average more sales than the other two style of prompt, but the small increase was not found to be statistically significant. A One-way ANOVA on the number of prompts converted into sales, again with style of prompt as a between subjects measure (3 levels: dialogue box, integral window and spoken prompt), also showed no significant main effect of the style of prompt (F[2,42] 0.5976, P: 0.5548). The hypothesis that the style of prompt would lead to an effect on the number of sales made (**H 1 {5}**) has been rejected in both the audio-only (Chapter 3) and this one-way video study.

No hypothesis had been made about the effect of the style of prompt on the number of airports included in the itineraries and, just as in the audio-only study, no main effect was found. A One-way ANOVA on the number of airports planned with style of prompt as a between subjects measure (3 levels: dialogue box, integral window and spoken prompt) found no significant main effect of the style of prompt (F[2,42] 0.3636, P: 0.6973).

Finally, a similar One-way ANOVA performed on the time taken to complete the consultations with style of prompt as a between subjects factor (3 levels: dialogue box, integral window and spoken prompt) also found no significant main effect (F[2,42] 0.9902, P: 0.3800). This lead to the hypothesis that the style of prompt would lead to an effect on the time each consultation took (**H 3 {7}**) being rejected, a finding which was striking given the large effect of the spoken style discovered in the audio-only study (Chapter 3). The measures of PROCESS may give some indication of why this turnaround has happened.

5.2.3 PROCESS

The PROCESS of the dialogue was measured using counts of the number of turns, words the agent and customer used as well as the amount of overlapping speech. Hypotheses were made about effects of the style of prompt on the number of turns and words which would be produced (**H4 {8}** and **H5 {9}**) with in both cases fewer being expected when the integral window was employed, but none were made about the amount of overlapping speech.

The counts of the number of turns and words were analysed using Analysis of Variance to determine whether the style of prompt was exerting an effect on their production and the results of these analyses are summarised in Table 5/2.

Measure	Dialogue Box	Integral Window	Spoken Prompt	ANOVA	P
Turns	176	202	198	F[2,42] 1.173	0.319
Words	1487	1803	1751	F[2,42] 1.460	0.244

Table 5/2 Summary table of turn and word counts by style of prompt.

A One-way ANOVA, with style of prompt as a between subjects factor (3 levels: dialogue box, integral window and spoken prompt), showed no significant difference between the styles in the number of turns of dialogue the agent and customer needed to reach their goal (F[2,42] 1.173, P: 0.319).

A Two-way ANOVA with style of prompt as a between subjects factor (3 levels: dialogue box, integral window and spoken prompt) and role of speaker as a within subjects factor (2 levels: agent and customer) showed no significant difference between the styles in the number of words used (F[2,42] 1.460, P: 0.244), but as in previous studies found a consistent effect of role (F[1,42] 34.020, P: 0.001).

	Dialogue Box	Integral Window	Spoken Prompt
Agent Words	954	1104	1091
Customer Words	532	699	660

Table 5/3 Summary table of mean words by role of speaker and style of prompt.

The agent consistently used more words than the customer (see Table 5/3) but there was no interaction between the role and style of prompt (F[2,42] 0.12, P: 0.988) so as in previous studies it was presumed that this was primarily an effect of the different communicative behaviours each role required.

These analyses indicate that neither **H4 {8}** nor **H5 {9}** is supported. The lack of support for an effect on the number of turns is not overly notable because none was found in the audio-only study either. On the other hand, the lack of support for an effect of style of prompt on the number of words produced is just as striking as the lack of support for an effect on the time taken for the consultations because there was a similarly sizeable effect of the style of prompt on both in the audio-only study. Since the effect in that study was discovered to be one of verbosity, to ensure that no effect is being missed, the number of words per turn was calculated here just as it was in Chapter 3.

Measure	Dialogue Box	Integral Window	Spoken Prompt	ANOVA	P
Words per Turn	8.41	8.87	8.83	F[2,42] 0.373	0.691

Table 5/4 Summary table of mean words per turn by style of prompt.

The results of this analysis also found no effect of the style of prompt (see Table 5/4). A Two-way ANOVA with style of prompt as a between subjects factor (3 levels: dialogue box, integral window and spoken prompt) and the role of the speaker as a within subjects factor (2 levels: agent and customer) showed no significant effect of the style of prompt (F[2,42] 0.373, P: 0.691).

	Dialogue Box	Integral Window	Spoken Prompt
Agent Words per Turn	10.9	10.6	11.2
Customer Words per Turn	5.9	7.1	6.5

Table 5/5 Summary table of mean words per turn by role of speaker and style of prompt.

As with the analysis of the word counts, there was a significant main effect of role with the agent producing consistently more words per turn (F[1,42] 38.237, $\underline{P}<0.001$) but no interaction between role and style (F[2,42] 0.411, $\underline{P}$: 0.411) (see Table 5/5) which was again interpreted to be an effect of the different communicative behaviours required of each role.

The remaining measure of PROCESS, that of counts of occurrences of overlapping speech, had no hypothesis made about it because it was not expected to be sensitive to the kind of effects the style of prompt would have. It was included because of its ubiquity in the study of mediated communication and the findings are presented in Table 5/6.

	Agent		Customer		Both Roles	
	Inter-upts	Over-laps	Inter-upts	Over-laps	Inter-upts	Over-laps
Dialogue Box	8.3 (9.4%)	3.0 (3.4%)	6.9 (7.8%)	4.7 (5.3%)	15.2 (8.6%)	7.7 (4.4%)
Integral Window	10.3 (10.1%)	5.3 (5.2%)	11.2 (11.1%)	9.3 (9.2%)	21.5 (10.6%)	14.6 (7.2%)
Spoken Prompt	9.1 (9.2%)	5.0 (5.1%)	12.1 (12.2%)	6.9 (7.0%)	21.2 (10.7%)	11.9 (6.0%)
All Styles	9.2 (9.6%)	4.4 (4.6%)	10.1 (10.5%)	7.0 (7.3%)	19.3 (10.1%)	11.4 (5.9%)

Table 5/6 Summary table of the number and percentage of overlaps by each speaker and combined.
The percentages given for agent and customer are the number of interruptions and overlaps as a percentage of each role's turns. The percentages give in the "Both Roles" column are of the totalled agent and customer turns.

A Three-way ANOVA with style of prompt as a between subjects factor (3 levels: dialogue box, integral window and spoken prompt) and role of speaker (2 levels: agent and customer) and type of overlapping speech (2 levels: interruption and overlap) as within subjects factors found no significant main effect of the style of prompt on the number of occurrences of overlapping speech produced ($F[2,42]$ 1.342, $\underline{P}$: 0.272).

There were significant main effects of role ($F[1,42]$ 5.147, P: 0.028) and type of overlapping speech ($F[1,42]$ 25.228, $P<0.001$) but no significant interactions between any of the factors. Because there were no interactions, these within subjects main effects were not considered more closely. The effect of type of overlapping speech was presumed to be due to a natural distribution of interruptions and overlaps particularly since it was similar to that found in the audio-only study. The effect of role was slightly tougher to classify as one of the communicative behaviour required of each participant because no such significant effect was found in the audio-only study, nevertheless because there was no interaction with the style of prompt any difference in distribution of overlapping speech between the speakers is more likely to be due to a interaction between the affordances of the asymmetrical one-way video and each role's communicative requirements rather than being relevant to the issue of variation in the computer interface.

Such an assumption is leant extra credence by the customer producing more overlapping speech than the agent, which matches expectations which could be inferred from previous work that the customer's vision of the agent is permitting them more fluid turn-taking. For instance, Argyle et al. (1968) found that subjects considered communication with one-way video easier than that with no vision, Argyle et al. (1973) found that subjects who could see their partners gazed in ways that suggested they were using it to support turn-taking and studies of two-way video mediation (see Section 1.3.2.1.3.2) have shown it to support more fluent turn-taking than audio-only.

5.3 Discussion

5.3.1 Summary of results

The PERFORMANCE and PROCESS measures which detected considerable effects of the computer-agent interface's style of prompt on the agent-customer dialogue in the audio-only study (Chapter 3) detected no such effects in this one-way video study. In particular, an expected increase in the time taken to complete the consultations and in the speakers' verbosity in spoken prompted consultations over integral window consultations were not found.

The first question these findings raise is whether the findings in the audio-only study were the product of chance rather than the independent variable. There are two possible reasons how chance could have come into play; firstly, the style of interface might really have been having no effect, and secondly, the variance amongst the consultation might have been different.

The possibility that the style of prompt is having no effect seems unlikely, not only was the effect of the style of prompt sizable, particularly the 154 second (21%) difference in time and 511 word (33%) difference in words between the integral window and spoken prompt consultations, but it had also been confirmed in the first study (Chapter 2) that the design of an agent's interface could have an effect on their dialogues with customers.

That study was looking at broad differences in the interface style rather than what appear to be such small differences as those introduced in the comparisons of sales prompting styles, however, and its similar findings would not provide a contraindication against a difference in the variance. This is, however, countered by a simple observation of the mean time taken for consultations and words produced during the individual style of prompts' consultations in each study (see Tables 5/7 and 5/8).

	Dialogue Box	Integral Window	Spoken Prompt
Audio-only	812	721	875
One-way Video	718	767	797

Table 5/7 Summary table of mean time taken for consultations by medium of communication and style of prompt.

	Dialogue Box	Integral Window	Spoken Prompt
Audio-only	1719	1613	2124
One-way Video	1487	1803	1751

Table 5/8 Summary table of mean words by medium of communication and style of prompt.

There is a paradigm shift in the length of the consultations with the dialogue box and spoken prompted consultations which are both longer on average the integral window in the audio-only study, becoming shorter than the integral window consultations in the one-way video study.

Even further indication that there is in fact an interaction between the medium of communication and the styles of prompt comes from the discovery in the one-way video study of statistically significant effects of the style of prompt on the customers' PERCEPTION of their consultations when no such discoveries had been made in either of the earlier audio-only studies. In the one-way video study, the integral window consultations were rated better, more efficient and more friendly than the spoken prompted and better than the the dialogue box prompted.

5.3.2 Conclusions and contentions

This suggests both the value of using multiple measures in the analysis of tripartite communication, as was inferred from Monk et al.'s (1994) similar comments about mediated human-human communication, and that the presence of the one-way video connection is fundamentally affecting the interactions within the tripartite communication. Its presence appears to be altering the effects of the variation in the design of the computer-agent interface in ways which could not be predicted from what is known about their methods of presentation from the human factors, human computer interaction and psychology studies reviewed in Chapter 3, nor from the empirical findings in the same chapter.

In the audio-only comparison of prompting styles, the customers appear not to have been subjectively aware of a difference between the styles, but the styles did exert large effects on the PROCESS of the dialogues, with the spoken prompt leading to much longer dialogues than the integral window. In this one-way video comparison of prompting styles, the opposite is true. The customers appear to be aware of the different means of presentation and find the spoken prompt more disturbing to the consultations than the integral window, but deleterous effects of the style of prompt on the PERFORMANCE and PROCESS of the consultations do not appear.

One explanation for this pattern would be that the one-way video gives the customer increased awareness of the agent-computer interaction, allowing them to perceive the dialogue boxes' and spoken prompts' more disturbing introductions of the sales opportunity and overcome the damaging effects on the agents' and their own patterns of speech.

To build a better model of how the visibility is altering the effects of the style of prompt on the consultations, particularly investigating the observed effects on the PROCESS of the agent-customer dialogues, the results from both the audio-only and one-way video studies were compared statistically, with new hypotheses being built which took into account what is known about the affordances of one-way video (see Section 5.0.2). The remainder of this chapter is concerned with the findings of those analyses.

5.4 Comparative Analysis

This section details the statistical analysis of the results from the audio-only and one-way video comparisons of sales prompting styles with the intention of finding evidence to support arguments about the different interactions between the styles of prompt and the co-visibility of the human participants in the tripartite communication.

The independent analysis of each study has suggested that with no co-visibility, customers are not aware of the different means of presentation but those means affect the structure of the agent-customer dialogue considerably, whereas when the customer can see the agent, they are more aware of the different styles of prompt but are able to counter negative impacts of those styles on the PROCESS of the consultations. The aim in statistically comparing the results from both studies is to provide more firm evidence for these contentions.

A statistical comparison is fully valid since the only difference between the studies was the medium of communication.

5.4.1 Hypotheses

The comparative hypotheses will be divided into two classes; hypotheses about the effects of the means of communication, and hypotheses about the interaction between the means of communication and the style of prompt by which the agent is alerted to sell products to the customer.

The hypotheses about the effects of the media of communication are intended to provide a background within which the interaction between the style of prompt and the media can be placed. These hypotheses depend directly upon the work reviewed at the beginning of this chapter.

The hypotheses about the interaction between the media of communication and the style of prompt used in the agent's computer interface is of the strongest relevance to the thesis. The suggestion that the effects of the style of prompt depend on the visibility between participants bears directly on the thesis' contention that the relationship between the three parties in tripartite communication involving one computer is triadic, with each dyadic interaction affecting each other.

5.4.1.1 Predictions

As in previous chapters, predictions will first be made about the ways in which the one-way video would be expected to affect the dialogues and then hypotheses will be made about how these effects will influence the measures taken.

In the case of all the predictions, it is likely that there will be a different effect on the agent and customer because of the asymmetry of vision.

The primary expected effect of the medium of communication will be an increase in the customer's awareness of the agent. This is based on such findings as Daly-Jones et al.'s (1998) that video communication can increase the salience of the partner, Root's (1988) that simple awareness of one's colleagues is a strong predictor of their collaboration and Short, Williams and Christie's (1976) belief that increased visibility, and the access to visual cues it brings, increases conversants' "social presence". This increased awareness would be expected to show up in customers' subjective feelings about their consultations.

More fluent synchronisation of turns has to be inferred rather than argued from previous evidence. Although Argyle et al. (1968) found subjects to rate communication with one-way vision easier than communication with no vision, and Argyle et al. (1973) showed that conversants who could see their partners gazed in ways that suggested they were using the visibility to support turn-taking, there has been no published study of the effects of one-way visibility on turn-taking behaviour. This evidence comes only from studies of two-way video, which have shown it to support more fluent turn-taking than audio-only communication (see Section 1.3.2.1.3.2).

Because only the customer can see the agent, and therefore only the customer can gather visual indicators of turn transition points, it would seem likely that there will be more fluent turn taking by the customer than by the agent.

There is more evidence that there would be coordination of content. This comes both from studies of one-way and two-way video communication.

Argyle et al. (1973) showed that one-way visibility through screens supported feedback, since their visible subjects appeared to use their gaze to signal attention while they were listening. Sellen's (1992) subjects believed they were more aware of their partners' attention in multi-party discussions when using either of two two-way video-conferencing systems than when using audio-only. Bull and Aylett (1998) have suggested that longer inter-speaker intervals when people can see each other while performing the cooperative map task may be because they are able to see when their partners are involved in physical tasks such as drawing routes on their maps or checking directions.

These would be expected to have effects on the PROCESS of communication more than other measures. The one-way video would be expected to lead to shorter turns by the customer because they will be able to gather visual signals of the agent's understanding as well as such verbal feedback as backchannels or other positive evidence of grounding (see Section 1.3.1.2.2). This does not necessarily contradict Argyle et al.'s (1973) finding that subjects obscured by screens spoke more than visible ones, because the tasks are different and the measure they used, words per minute, may have been influenced by their tightly controlled time-limit.

Fewer vocal backchannels by the agent would be expected because they will be able to give gestural indication of understanding instead.

For the interaction between the one-way video and the style of prompt, improved feedback could mean that there will be a customer lead reduction in the time and word increase due to the spoken prompt. The spoken prompt lead to both an increase in the agents' and customers' verbosity, but the Conversational Games Analysis indicated this to have been lead by the agent. It would appear that when conducting the novel task over audio-only the customers take their lead as to how information should be transferred from the agent. With one-way video, the customer will be able to use their ability to gather visual feedback to structure their utterances more naturally, and therefore be empowered to overcome the agent's lead. It would therefore be expected that the customer will not be as strongly influenced by the agent's increased verbosity after spoken prompt and will maintain a level closer to that with the other prompting styles.

Better coordination of content would be expected if the task were visual and the cameras were used to display the objects (Gaver et al., 1993; Kraut et al., 1996), but this is not the case, so it is unlikely that the video will be of great benefit. On the other hand, being able to judge the focus of the agent's attention may provide some content coordination. For instance, the customer may be able to predict topic introductions by seeing whether the agent's eyes are inclined towards their computer or off-line resources, they may accept visually represented responses such as head shakes or nods in answer to yes-no queries, and may be able to perceive agents' distraction during prompts' presentations. These behaviours have been termed "gaze awareness" by Daly-Jones et al., 1998) and they may permit the customer to pre-empt agents' utterances which would have been required in audio-only communication leading to fewer turns. For instance, on seeing the agent turning towards their off-line sales information, the customers may indicate that they are not interested before the agent has mentioned the opportunity.

Previous work would indicate that the effect of visibility will be related to the task the participants are involved in, with it exerting effects on the

PERFORMANCE of conflictual negotiation tasks but not of cooperative tasks. The version of the travel game developed for this study includes both. The task of building the itinerary is essentially cooperative, with both agent and customer benefiting from including more locations in the customer's itinerary; the agents knew that the number of airports would be used as a tie-breaker in the case of two or more selling the same number of travel passes. The sales task is more conflictual because the travel pass gives no objective benefit for the customer. The discussion of the product uses time they could be spending on building the itinerary, the product itself uses funds they require to make backtracks, and there is no association between the travel pass and the criteria they have for winning a prize. The only reason they have for purchasing it is the subjective one that it relates to the instructions they are told their friend gave them, and that it would save them money in the long term.

For this reason it was predicted that the inclusion of one-way video would have no effect on the PERFORMANCE of the itinerary building task, but would affect the PERFORMANCE of the sales task.

Typically, the inclusion of visibility increases subjective aspects of conflictual tasks (Short, 1974; Stephenson, Ayling and Rutter, 1976; see Section 1.3.2.1.0) and for this reason more sales would be predicted. The customer is expected to have an increased awareness of the agent's social presence and this effect may lead to the agent having comparable persuasive abilities as when the negotiators are covisible.

The effects of the one-way visibility are hard to predict however, and the following may also influence the results. The discomfort visible subjects feel may lead to more awkward presentation by the agent and fewer sales. The inability of the agent to observe the customer may mean that they do not adjust their negotiation appropriately to the customer's reactions. Since the sales task distracts the customer from building their itinerary, it is also possible that they will have more vested interest in the sales task; they will not want to waste time on it and this may alter the whole nature of the competition.

5.4.1.2 Hypotheses about media of communication

The hypotheses about the effects of the media of communication are not directly related to the thesis' main concern, which is the nature of the tripartite communication between computer, agent and customer, but serve to provide a general evaluation of effects of one-way video within which the interaction with each style of prompt can be grounded. Because the second study provides an example of one-way visibility in a believable if not natural, setting the findings about the independent effects of the medium of communication also have the potential to further our general knowledge about mediation in human-human communication.

5.4.1.2.1 PERCEPTION

The first hypothesis to be made about the effect of the one-way video is that the customer will have a more positive PERCEPTION of the video-mediated dialogues than of the audio-only because they will feel less psychological distance (see Section 1.3.2.2.1) from the agent .

HC1 The customer will have a more positive perception of the one-way video-mediated consultations than the audio-only.

Of particular interest will be the effect of the medium of communication on those scales in which was discovered a main effect of the style of prompt in the one-way video study, though it is during the analysis of the interaction between the medium of communication and the style of prompt that the most telling findings about these are expected.

5.4.1.2.2 PERFORMANCE

PERFORMANCE was measured with the number of prompts to sell the travel pass which were converted into sales, the number of airports included in the itinerary and the total time taken to complete the consultation.

Sales of the travel pass was a conflictual task which was primarily a measure of the agent's success. Building the itinerary was a cooperative task which was primarily a measure of the customer's success. The time taken to complete the consultation was a measure of joint success (see Sections 3.0.3.2.5 and 3.0.3.3.2 for more discussion of these measures).

In building the hypotheses about the effects of medium a major consideration was the distinction drawn in Section 1.3.2.1.0 between negotiation and cooperative tasks. It is a well respected finding that the PERFORMANCE of cooperative tasks is not seriously affected by variation in the media over which they are performed, but that conflictual negotiation tasks are.

Because the sales task involves such negotiation, it is expected to be susceptible to the change in medium. Decreased psychological distance between the agent and customer would be expected to permit the agent to be more persuasive, and more sales should ensue.

> **HC2** The one-way video will lead to more sales than audio-only.

The building of the itineraries is, on the other hand, a cooperative task which comprised 84.9% of the consultations during the audio-only study, which would lead to the expectation that neither the number of airports nor the time taken to perform the task will be affected by the medium.

An eyeball observation of the mean amount of airports in the audio-only and one-way video studies contradicts this however. There were on average 9.6 airports in each trial during the audio-only comparison of sales prompting styles and 10.6 during the one-way video.

If there is in fact an effect on the PERFORMANCE of the itinerary building, task this would suggest that the presence of the sales task has led to conflict within the building of the itinerary; potentially because the customer is fighting for the time which they would be wasting in discussing the sales prompts and using the time they win in the one-way video consultations to improve their itineraries. In essence, there would be conflict between the customer's desire to fit as many airports as possible into their itinerary, and the agent's desire to sell the travel pass.

The second comparative hypothesis about PERFORMANCE is that the medium of communication will affect the number of airports included in itineraries, with the one-way video leading to larger itineraries.

> **HC3** The one-way video will lead to larger itineraries than audio-only.

Although support for both hypotheses is not impossible, it seems more likely that support for one will preclude support for the other. If support is found for **HC2** but not **HC3** that would suggest that the agent's bargaining position is strengthened by being visible. If support is found for **HC3** but not **HC2** that would suggest that the customer's bargaining position is strengthened by vision of the agent.

5.4.1.2.3 PROCESS

The measures of PROCESS taken included the number of turns and words during the dialogues and the composite words per turn, all of which measure the density of the dialogue, and a count of overlapping speech which is used to indicate how well turn-taking is synchronised.

First, it was hypothesized that there would be fewer words and turns in the one-way video, and that this would be lead by the customer.

The decrease in words was predicted for two reasons, firstly, the agent would be able to replace many of their vocal feedback turns with visual signals, and secondly the customer would be able to read the agent's visual feedback to produce more succinct utterances and use their "gaze awareness" (Daly-Jones et al., 1998) to preempt agents' vocal utterances.

HC4 Fewer words will be produced in the one-way video consultations than in the audio-only, and this will be lead by the customer.

Support for this, particularly if it is found to be lead by the customer, would provide further suggestion that the customer's bargaining ability is strengthened by the one-way visibility of their agents.

The decrease in turns was hypothesized for essentially the same reason. In the audio-only dialogues, speakers often appeared to leave short gaps in longer turns which their partners could fill with short feedback utterances . With the one-way video, the agent's feedback could be produced visually and this could reduce the number of turns considerably.

HC5 Fewer turns will be produced in the one-way video consultations than in the audio-only.

At the same time it was recognised that other studies which had compared audio-only with two-way video had not shown any effect of the medium on density unless there were delays in audio (O'Conaill, Whittaker and Wilbur, 1993) or direct eye contact was possible (Anderson et al., 1994) (see Section 1.3.2.1.3.1).

In the absence of technological influences such as delay, "*short frequent turns and overlapping speech can be regarded as indications of verbal fluency*" (Daly-Jones et al., 1998, p. 36), so evidence for such behaviours will be sought where appropriate.

Finally, making inferences from the previous work of Argyle et al. (1968, 1973) on one-way video and studies of two-way video (see Section 1.3.2.1.3.2), it was predicted that one-way video would lead to increased fluency in the customer's turn-taking. It was hypothesized, therefore, that the one-way video would lead to more overlapping speech and that this would lead produced by the customer's production.

HC6 There will be more overlapping speech when the one-way video is used, and this will be lead by the customer.

5.4.1.3 Hypotheses about the interaction between medium of communication and style of prompt

For the thesis, it is the hypotheses about the interaction between the medium of communication and the style of prompt which are most important, because it is here that a triadic nature of the communication between the computer, agent and customer could most directly be observed.

In particular, evidence is being sought to support the observations that the one-way video permits the customer to overcome the significantly detrimental effects of the spoken prompts on the PERFORMANCE and PROCESS of their consultation which was discovered in the audio-only study, presumably by making them more aware of those prompts' disturbance of their agent's cooperation.

The hypotheses will concentrate attention on those measures in which effects of the style of prompt were found in the individual studies.

5.4.1.3.1 PERCEPTION

The hypotheses about the interaction between medium of communication and the style of prompt on the customer's ratings of their consultations were largely based on the previous empirical findings.

Because the one-way video appeared to make the customer more sensitive to the disturbance caused to the agent's cooperation by the more demanding dialogue box and spoken prompt styles, it was predicted that there would be an interaction between the medium of communication and style of prompt with the dialogue box and spoken prompt receiving lower ratings and the integral windows higher.

HC7 The one-way video will lead to lower ratings of the dialogue box and spoken prompt consultations, but higher ratings of the integral window, than audio-only.

Particular interest will be paid to the three scales which showed an effect of the style of prompt in the one-way video study (Osgood "Good-bad", CIQM "Inefficient-efficient" and CIQM "Unfriendly-friendly").

5.4.1.3.2 PERFORMANCE

Because the one-way video appears to reduce the damaging effects of the style of prompt on the dialogues between the agent and customer, the following hypothesis was made about the PERFORMANCE of the consultations.

HC8 There will be an interaction between the medium of communication and the style of prompt on the time taken to complete the consultation.

In particular, it is expected that the one-way video will lead to shorter spoken prompt consultations than audio-only, since it was the spoken prompt dialogues which showed the largest damaging effects in the audio-only study.

Support for this hypothesis would provide further indication that the customer is able to overcome damaging effects of more disruptive styles of prompt on their primary task of building the flight itinerary.

5.4.1.3.3 PROCESS

The work reviewed at the beginning of this chapter would indicate that with one-way video the customer will be able to coordinate better with the agent's visual feedback. From this, it would be expected that the bilateral increase in the verbosity of the dialogues caused by the spoken prompt in the audio-only study could be overcome by the customer in the one-way video study. The increase in the agent's verbosity of EXPLAINs in the audio-only spoken prompted consultations (see Section 4.3.1.3 and 4.3.1.4) was matched by an increase in the customer's verbosity of EXPLAINs (see Section 4.3.1.3). It is expected that with the ability to perceive the reason for the increased verbosity, the spoken prompt's interruption of their dialogue, the customer will also recognise the agent's increased verbosity and be able to return a more normal level of verbosity of explanation to the dialogues.

The only hypothesis made about the PROCESS is that the medium of communication will affect the number of words produced in each style of prompts' consultations to a different degree.

HC9 The medium of communication will affect the number of words produced in each style of prompt's consultations to different degrees.

In particular it is expected that the one-way video will lead to spoken prompt dialogues which require fewer words than audio-only, and that the effect will be lead by the customer, who uses their vision of the agent to overcome damaging styles of prompts' effects.

5.4.1.4 Summary of Comparative Hypotheses

For ease of reference the hypotheses have been summarised here:

HC1 The customer will have a more positive perception of the one-way video-mediated consultations than the audio-only.

HC2 The one-way video will lead to more sales than audio-only.

HC3 The one-way video will lead to larger itineraries than audio-only.

HC4 Fewer words will be produced in the one-way video consultations than in the audio-only, and this will be lead by the customer.

HC5 Fewer turns will be produced in the one-way video consultations than in the audio-only.

HC6 There will be more interruptions when the one-way video is used, and this will be lead by the customer.

HC7 The one-way video will lead to lower ratings of the dialogue box and spoken prompt consultations, but higher ratings of the integral window, than audio-only.

HC8 There will be an interaction between the medium of communication and the style of prompt on the time taken to complete the consultation.

HC9 The medium of communication will affect the number of words produced in each style of prompt's consultations to different degrees.

5.4.2 Analysis of the effects of communication medium

5.4.2.1 PERCEPTION

The first hypothesis was that there would be an effect on the customers' subjective PERCEPTION of their consultations due to increased social awareness of the agent.

HC1 The customer will have a more positive perception of the one-way video-mediated consultations than the audio-only.

All the scaled questions were analysed with Two-way ANOVAs, taking both medium of communication (2 levels: audio-only and one-way video) and style of prompt (3 levels: dialogue box, integral window and spoken prompt) as between subjects factors[12]. Although none of these analyses showed a significant effect of the medium of communication, there was one trend.

In answer to the question "How satisfied were you with the final outcome of the consultation?", there was a main effect of the medium with one-way video being rated more satisfactory than the audio-only (Two-way ANOVA, $F[1,84]$ 3.489, $\underline{P}$: 0.065). After the one-way video consultations only 8/45 customers were "dissatisfied" or "very dissatisfied", whereas in the audio-only this ratio was 16/45. Given how broad-brushed a measure questionnaires this gives an acceptable suggestion that the customer did in fact feel there was some benefit to the one-way video, and consequently that the one-way video was leading to decreased psychological distance.

---

12 The ANOVA was considered appropriate for analysing the scalar questions for the same reasons described in Section 2.2.1.

None of the other main effects of medium were significant at more than the 10% level. Specifically worth noting is that there was no statistically significant main effect of the medium of communication on the customers' ratings of the consultations on the three scales which showed a significant main effect of the style of prompt in the one-way video study (Osgood "Bad-good", $F[1,84]$ 0.057, $\underline{P}$: 0.812, CIQM "Inefficient-efficient", $F[1,84]$ 1.106, $\underline{P}$: 0.296, and CIQM "Unfriendly-friendly", $F[1,84]$ 0.000, $\underline{P}$: 1.000).

5.4.2.2 PERFORMANCE

The hypotheses made about the effect of the medium of communication on the PERFORMANCE of the task were that there would be an increase in the number of sales and in the number of airports included in itineraries during one-way video consultations. Support for the sales hypothesis would indicate that the agent's ability to bargain had been improved, and support for the airport hypothesis would indicate that the customer's ability to bargain had been improved. If one were improved, but not the other, the asymmetric benefits of the one-way video might be perceived.

HC2 The one-way video will lead to more sales than audio-only.

HC3 The one-way video will lead to larger itineraries than audio-only.

All three measures taken were analysed with Two-way ANOVAs which took both the medium of communication (2 levels: audio-only and one-way video) and the style of prompt (3 levels: dialogue box, integral window and spoken prompt) as between subjects factors.

The third measure, about which no hypotheses were made, was the time taken to complete the task, and although the one-way video consultations were slightly shorter (41.6 seconds, 5%) than the audio-only, there was no main effect of the medium of communication (Two-way ANOVA, $F[1,84]$ 1.556, $\underline{P}$: 0.216) (see Table 5/9).

Measure	Audio-Only	One-way Video	ANOVA	$\underline{P}$
Time Taken	802	761	F(1,84) 1.556	0.216

Table 5/9 Effect of medium of communication on Time Taken.

As for the measures about which hypotheses were made, although there were on average more sales during one-way video than audio-only consultations (see Table 5/10), this difference was not statistically significant (Two-way ANOVA, F[1,84] 0.046, P: 0.830).

Measure	Audio-Only	One-way Video	ANOVA	P
Sales Made	1.7	1.8	F(1,84) 0.046	0.830

Table 5/10 Effect of medium of communication on Sales Made.

On the other hand, there was a significant main effect of the medium of communication on the number of airports in itineraries (Two-way ANOVA, F[1,84] 5.450, P: 0.022), with there being 1 (10%) more on average in one-way video consultations than in audio-only (see Table 5/11).

Measure	Audio-Only	One-way Video	ANOVA	P
Airports Planned	9.6	10.6	F(1,84) 5.450	0.022

Table 5/11 Effect of medium of communication on the number of airports planned.

Support for the hypothesis about the number of airports has several implications. Firstly, it suggests that the presence of the sales task has added a negotiative dimension to the otherwise cooperative Travel Game since effects of the medium of communication on PERFORMANCE have not previously appeared in other than negotiative tasks. Secondly, it suggests that the customer's ability to bargain is improved by the presence of the one-way video, and because there was no concomittant significant increase in the agent's ability to bargain, indicates that being able to see one's opponent in negotiations is more important than being seen. This is supported by such findings as Gaver et al.'s (1993) that subjects avoided face-to-face views when trying to persuade partners to go against their wishes.

Reading of Whittaker and O'Conaill's (1997) taxonomy of conversational mechanisms (Reproduced in Figure 5/1) and consideration of Daly-Jones et al.'s (1998) concept of "Gaze Awareness", would suggest this is because being able to see an opponent permits a negotiator to better judge their emotional stance and motivations and potentially preempt opponents' attempts to introduce newly prompted arguments.

5.4.2.3 PROCESS

Three hypotheses were made about the effect of the medium of communication on the PROCESS of the communication between the agent and customer.

HC3 Fewer words will be produced in the one-way video consultations than in the audio-only, and this will be lead by the customer.

HC4 Fewer turns will be produced in the one-way video consultations than in the audio-only.

HC5 There will be more interruptions when the one-way video is used, and this will be lead by the customer.

Three-way ANOVAs with medium of communication (2 levels: audio-only and one-way video) and style of prompt (3 levels: dialogue box, integral window and spoken prompt) as between subjects factors and role of speaker (2 levels: agent and customer) were used to analyse the amount of words produced. Although, as was predicted, there were on average fewer words when one-way video was used than when audio-only was used there was no main effect of the medium of communication (Three-way ANOVA, $F[1,84]$ 1.554, $\underline{P}$: 0.216) (see Table 5/12).

Measure	Audio-Only	One-way Video	ANOVA	<u>P</u>
Words	1074	1050	F(1,84) 1.554	0.216

Table 5/12 Summary table of the effect of medium of communication on the number of words produced.

The hypothesis about the number of turns was supported however, with a Two-way ANOVA with medium of communication (2 levels: audio-only and one-way video) and style of prompt (3 levels: dialogue box, integral window and spoken prompt) as between subjects factors finding there to be a main effect of the medium of communication (Two-way ANOVA, F[2,84] 4.534, P: 0.036) and there being 27 (14%) more turns in the audio-only consultations than the one-way video (see Table 5/13).

Measure	Audio-Only	One-way Video	ANOVA	P
Turns	219	192	F(1,84) 4.534	0.036

Table 5/13 Effect of medium of communication on the number of turns.

Finally, the counts of overlapping and interrupting speech were analysed. A Four-way ANOVA with medium of communication (2 levels: audio-only and one-way video) and style of prompt (3 levels: dialogue box, integral window and spoken prompt) as between subjects factors and role of speaker (2 levels: agent and customer) and type of overlapping speech (2 levels: overlaps and interruptions) as within subjects factors.

No effect of the medium of communication was found on the overall amount of overlapping speech (Four-way ANOVA, F[1,84] 0.491, P: 0.485), but there were effects of role (Four-way ANOVA, F[1,84] 5.939, P: 0.017), the type of overlapping speech (Four-way ANOVA, F[1,84] 63.744, P: <0.001), a significant interaction between role and the type of overlapping speech (Four-way ANOVA, F[2,84] 8.511, P: 0.005) and a trend in the interaction between all factors (Four-way ANOVA, F[2,84] 2,561, P: 0.083).

Medium of Communication	Agent Overlaps	Customer Overlaps	Agent Interrupt-ions	Customer Interrupt-ions
Audio-only	5.0 (2.3%)	7.1 (3.2%)	11.2 (5.1%)	10.6 (4.8%)
One-way	4.4 (2.3%)	7.0 (3.6%)	9.2 (4.8%)	10.1 (5.2%)

Table 5/14 The number and percentage of overlaps by each speaker. The percentages given are of the totalled agent and customer turns.

Although the customer produced more overlaps and interruptions as a proportion of the total number of turns, the difference is small (see Table 5/14), and since there was no interaction between the medium of communication, type of overlapping speech and role of speaker (Four-way ANOVA, $F[2,84]$ 0.231, $\underline{P}$: 0.794), the medium of communication and role (Four-way ANOVA, $F[2,84]$ 0.850, $\underline{P}$: 0359) or the medium of communication and type of overlapping speech (Four-way ANOVA, $F[2,84]$ 0.683, $\underline{P}$: 0.411), this was not investigated further.

No other main effects or interactions were significant at greater than the 10% level, and the hypothesis that the the medium of communication would lead to more interruptions, particularly by the customer, does not seem to be supported.

Because the pattern of results is slightly unusual, with no major indication of an effect on turn-taking, but a considerable (14%) effect on turns when other studies (O'Conaill et al., 1993; Anderson et al., 1994) had not found such, a certain amount of consideration of these results is necessary.

It is quite possible that the results are related to an interaction between the methodology and the nature of tripartite communication. Greatbatch et al. (1993) noted that in tripartite communication in doctor's surgeries, patients timed their dialogue turns to coincide with turns in the doctor-computer interaction. This implies that the added visibility of the agent afforded by the one-way video may lead to the customer coordinating their turns with the agent-computer interaction as well as just with the agent.

Both the counts of turns and the measures of turn-taking only accounted for turns within the agent-customer dialogue, and so would have missed such an interaction. Although this is unlikely to have influenced the counts of turns, since one of the predictions was that viewing the agent's interaction with their computer could lead to the pre-emption of agent turns, it could affect the counts of overlapping speech more seriously. In retrospect, it could even have been predicted that the one-way video could lead to fewer customer interruptions because the increased fluency will be with inaudible agent-computer turns.

For greater sensitivity to the nature of tripartite communication, it might be beneficial to include these human-computer turns within the measure.

5.4.2.4 Summary

The introduction of the one-way video link through which the customer can see the agent led to no significant main effects on the customer's PERCEPTION of their consultations, nor on the amount of sales made, the time taken to complete consultations or the number of words produced. There was, however, a significant effect on the amount of airports included in the itineraries the participants built and decrease in the number of turns they took.

The finding of an effect of medium on the number of airports included in itineraries suggests that the presence of the sales task leads to negotiation for time to complete each participant's individually goal oriented tasks, and because it is the PERFORMANCE of the customer's task which benefits from the one-way video that would suggest that being able to see one's opponent in negotiation tasks is more empowering than being seen; an important basic finding.

The reduction in the number of turns tentatively suggests that the customer may also be better able to coordinate with the agent's use of their interface.

5.4.3 Analysis of interaction between medium of communication and style of prompt

In Chapter 3 and the earlier sections of this chapter (see Sections 5.2 and 5.3), the effects of the styles of prompt were independently reported for what have become the audio-only and one-way video conditions in this section's comparative statistical analyses. The earlier reports clearly showed that the patterns of effects were different when the human-human communication was via audio-only or one-way video, so the effects of the style of prompt in the combined analyses will not be reported. Instead the concentration will be on the interaction between the medium of communication and the style of interface.

The aim is essentially to find evidence for there being an interaction between the medium and style of prompt. This is intended to support the argument that the medium of communication has influenced the customer's awareness of the agent's interaction with their system and their ability to counter the disruptive style of prompts' damaging effects.

In Section 5.4.1 the following hypotheses were made about these interactions.

HC7 The one-way video will lead to lower ratings of the dialogue box and spoken prompt consultations, but higher ratings of the integral window, than audio-only.

HC8 There will be an interaction between the medium of communication and the style of prompt on the time taken to complete the consultation.

HC9 The medium of communication will affect the number of words produced in each style of prompt's consultations to different degrees.

5.4.3.1 PERCEPTION

As stated above, during the comparative analysis, the customer's responses to the Likert-style questions in the questionnaires were analysed with Two-way ANOVAs taking medium of communication (2 level: audio-only and one-way video) and style of prompt (3 levels: dialogue box, integral window and spoken prompt) as between subjects factors. The sole interest in this section is the interaction between the medium of communication and style of prompt.

The only significant interaction between the medium of communication and style of prompt was found on the customers' responses to the Osgood "Good-Bad" scale, which had found an effect of style in the one-way video study, ($F[2,84]$ 4.789, $\underline{P}$: 0.011).

Simple main effects of the medium of communication on each style of prompts' consultations taken individually showed that the effect of the medium was concentrated in the integral window consultations, with the one-way video integral window consultations being rated more favourably than the audio-only ($F[1,28]$ 8.805, $\underline{P}$: 0.006). Although the one-way video dialogue box and spoken prompt consultations were in general rated less favourably than the audio-only, neither of these simple main effects were statistically significant (dialogue box: $F[1,28]$ 0.376, $\underline{P}$: 0.545) (spoken prompt: $F[1,28]$ 2.406, $\underline{P}$: 0.132).

The only other scale which showed an interaction between the style of prompt and medium which was significant at more than the 10% level was a trend in the customers' ratings of the consultation on the Osgood "Weak-Strong" scale ($F[2,84]$ 2.762, $\underline{P}$: 0.069), apparently because the customers rated the integral window consultations more "strong" and the spoken prompt consultations more "weak" in the one-way video study. This was not statistically analysed further because it was only a trend.

Specifically worth noting is that there were no further statistically significant effects on ratings of those scales which had shown an effect of the style of prompt in the one-way video study. There was no interaction between the medium of communication and style of prompt on customers' ratings of the consultation on the CIQM "Unfriendly-friendly" scale (F[2,84] 1.322, P: 0.272) nor on the CIQM "Inefficient-efficient" scale (F[2,84] 0.542, P: 0.584).

Although it is only statistically supported by the customers' ratings on one of the questionnaires scales, because of the blunt nature of questionnaire data, this still provides persuasive support for **HC7**.

It can be speculated that the customers prefer being able to see the agents, but that this is being counteracted by detrimental effects of the more imposing prompts' interruption of the agents' performance of core travel game task.

There is evidence that the customers are more aware of the imposition of the spoken and dialogue box prompts when they can see the agent, in their responses to the CIQM "Unfriendly-friendly" and "Inefficient-efficient" and LOGOS "Good-bad" scales (Section 5.2.1) in the one-way study, and the non-significant but on average higher ratings given to the dialogue box and spoken prompt consultations on the LOGOS "Good-bad" scale in the audio-only than in the one-way video study (see above).

The difference in the customer's rating of the Integral Window and Dialogue Box and Spoken Prompt consultations may be accounting for there being no main effects of the medium of communication; leading to there only being a weak hint – in the trend towards customers rating the one-way video consultations higher than the audio-only consultations on the "Satisfaction" scale (Section 5.4.2.1) – that, overall, the customers prefer the one-way video to the audio-only.

5.4.3.2 PERFORMANCE

The Two-way ANOVAs with medium of communication (2 level: audio-only and one-way video) and style of prompt (3 levels: dialogue box, integral window and spoken prompt) used to analyse the number of sales, the number of airports included in itineraries and the time taken to complete the consultations found no interactions between the medium of communication and the style of prompt which were significant at greater than the 10% level (time taken: F[2,84] 1.769, P: 0.177) (airports: F[2,84] 0.250, P: 0.779) (sale: F[2,84] 0.246, P: 0.782).

To look for specific effects of the medium of communication on each style of prompts' consultations, the simple main effects of the medium on each style of prompts' consultations were calculated. The only statistically significant simple main effect of the medium was on the number of airports in the dialogue box consultations (F[1,28] 4.523, P: 0.042), in which the one-way video lead to 1.3 (13.8%) more airports than audio-only (see Table 5/15).

Measure	Audio-only	One-way video	ANOVA	P
Airports	9.4	10.7	F[1,28] 4.532	0.042

Table 5/15 Simple main effect of medium of communication on number of airports included in itineraries during dialogue box consultations.

None of the other simple main effects were significant at more than the 10% level; the eighth comparative hypothesis (**HC8**) was therefore not supported.

5.4.3.3 PROCESS

The Two-way ANOVA with medium of communication (2 levels: audio-only and one-way video) and the style of prompt (3 levels: dialogue box, integral window and spoken prompt) as between subjects factors performed on the counts of turns found no interaction between the medium of communication and the style of prompt (F[2,84] 0.565, P: 0.571).

There was in fact a decrease in the number of turns in all styles' consultations (see Table 5/16), but simple main effects of medium of communication on the turns in each style of prompt's dialogues were calculated to determine whether one style of prompt's dialogues was more affected by the change in medium than others.

	Dialogue Box	Integral Window	Spoken Prompt
Audio-only	202	213	242
One-way Video	176	202	198

Table 5/16 Summary table of mean turns by medium of communication and style of prompt.

In fact, the only simple main effect of the medium of communication on the number of turns during a single styles' consultations which reached a level of significance greater than 10% was a trend in the effect on the spoken prompted consultations (F[1,28] 3.787, P: 0.062). The one-way video leads to 44 (22%) fewer turns in spoken prompt consultations than audio-only.

The Three-way ANOVA with medium of communication (2 levels: audio-only and one-way video) and the style of prompt (3 levels: dialogue box, integral window and spoken prompt) as between subjects factors and role of speaker (2 levels: agent and customer) as a within subjects factor performed on the counts of words also found no interaction between the medium of communication and the style of prompt (F[2,84] 2.329, P: 0.104).

The only simple main effect of the medium of communication on each style of prompt's consultation which was significant at more than the 10% level was a strong trend in the simple main effects of medium in the spoken prompt consultations (F[1,28] 4.113, P: 0.052) (see Table 5/17).

	Dialogue Box	Integral Window	Spoken Prompt
Audio-only	1719	1613	2124
One-way Video	1487	1803	1751

Table 5/17 Summary table of mean words by medium of communication and style of prompt.

A similar Three-way ANOVA performed on the counts of words per turn, taking medium of communication (2 levels: audio-only and one-way video) and style of prompt (3 levels: dialogue box, integral window and spoken prompt) as between subjects factors and role of speaker (2 levels: agent and customer) as a within subjects factor, found neither the main effect of the style of prompt nor any interactions to be significant at more than the 10% level.

On the other hand, simple main effects of the effect of the medium of communication on the customers' words in each style's consultations showed a significant simple main effect in dialogue box consultations (F[1,28] 5.214, P: 0.030) and a trend in the simple main effect in spoken prompt dialogues (F[1,28] 3.208, P: 0.084). The customer produced 231 (43%) more words in the audio-only dialogue box consultations than in the one-way video dialogue box consultations and 200 (30%) more words in the audio-only spoken prompt consultations than in the one-way video spoken prompt dialogues (see Table 5/18).

	Dialogue Box	Integral Window	Spoken Prompt
Audio-only	763	614	860
One-way Video	532	699	660

Table 5/18 Summary table of mean customer words by medium of communication and style of prompt.

Similar analyses of the simple main effects of medium of communication on the agent's word production showed no simple main effects significant at more than the 10% level (see Table 5/19).

	Dialogue Box	Integral Window	Spoken Prompt
Audio-only	956	999	1265
One-way Video	954	1104	1091

Table 5/19 Summary table of mean agent words by medium of communication and style of prompt.

In both the ANOVA of the counts of words and of the words per turn there were main effects of role (words, $F[2,84]$ 76.302, $P < 0.001$) (words per turn, $F[2,84]$ 82.547, $P < 0.001$) with the agent producing more in both cases, but because no interactions with role were significant at more than the 10% level this was presumed to be because of the different communicative requirements of each role.

Although no direct support is found for **HC9**, the finding that the medium of communication does affect the customer's production of words in the dialogue box consultations, and trend towards such an effect in the spoken prompt consultation, does suggest that the medium of communication does affect the PROCESS of the communication, but that because the one-way video only affords the customer increased vision of the agent, that effect is stronger in the customer's production.

5.4.4 Summary of Comparative Analysis

The most important finding from the comparative analysis is that the one-way video affects customer's verbosity during the consultations which used the more disruptive styles of prompt; dialogue box and spoken prompt. With the introduction of the one-way video, there was a significant decrease in the amount of words the customers produced during the dialogue box consultations, and a trend towards such an effect during the spoken prompt consultations.

Coupled with the finding that there was an interaction between the medium of communication and the customer's PERCEPTION of each style's consultations, this provides good evidence that the presence of the one-way video is permitting the customer to overcome the detrimental effects of the disruptive styles of prompt because they can see their effects on the agent.

5.5 Discussion

The comparison of the dialogue box, integral window and spoken sales prompting styles when the customer could see the agent through a one-way video link demonstrated that the effects of the style of prompt could vary with the visibility between the two human participants.

In this study, the customers perceived those consultations in which the spoken prompts were used, which was expected to be more disruptive to the agent-customer dialogue than the other styles, as producing less favourable consultations. They were rated less well on scales of friendliness, efficiency and general pleasure. At the same time, the only major effect on the PERFORMANCE and PROCESS of the dialogues was an improvement in the itineraries the customer built.

It was contended that the increased awareness the customer gained of the agent's site through the one-way video both made them more aware of disruption of the agent's cooperation with them, and better able to counter that disruption.

Comparative analysis of the audio-only and one-way video studies ensued to test this contention and proved highly fruitful.

Simple consideration of the effects of the medium of communication on the PERCEPTION, PERFORMANCE and PROCESS of the overall consultations provides insight into the potential effects of video communication on communication. The addition of the sales task to the otherwise cooperative travel game task makes the whole consultation negotiative, because the participants can negotiate for the amount of time to be spent on each subtask.

Observation of the interactions between the media of communication and the styles of prompt did indeed find evidence that the one-way video made the customer more aware of the disruption caused by the more demanding styles of prompt, dialogue box and spoken prompt, and because they were found to be leading the decrease in the verbosity of the dialogues again suggested that the benefit of the one-way video was to the customer.

A clear design implication would be that less invasive prompting styles are preferable in interfaces to be used in tripartite customer service interactions when the human participants are covisible. The less invasive integral window in this study lead to improved customer PERCEPTION of their consultations and no significant effects on the task's PERFORMANCE.

Overall, because in the comparative analysis mere variation in the visibility between the agent and customer lead to paradigm changes in the effects of the style of prompt, it can be concluded that tripartite relationship between agent, customer and computer is indeed triadic with each interaction dependent on each other one.

Chapter 6

—

Thesis Conclusions

—

The Experimental Investigation of Tripartite Communication in Simulated Customer Service Interactions

6.0 Introduction

This thesis has involved the experimental investigation of tripartite communication between two humans and a computer as it occurs in customer service situations. The aim has been to inform HCI about the nature of such tripartite communication through the application of suitable psychologically derived techniques.

This type of tripartite communication involves one "agent", who has direct access to the computer, communicating with a second "customer" who needs to find out information from the computer. It is essentially the situation which occurs remotely during a call to telephone call centres or face-to-face in travel agencies or doctor's offices.

Although there have been suggestive field studies of similar tripartite interactions, the studies in this thesis represent the first attempt to study the internal relationships of such interactions in the laboratory. The aim was to test whether changes in the design of the agent's interface can permeate to their dialogues with the customer, and their customers' attitude towards the consultations; thereby providing evidence of a causal relationship between the design of the interface and the agent-customer dialogues, and indirectly of the treatment of the computer in these interactions as a third-party rather than an inanimate object.

In Chapter 2, the current literature on such tripartite communication was reviewed. It was found that, although early studies had conceived the role of the computer user to be that of a "surrogate user" who just retrieved information from their system at the beck and call of the third party, this had developed into a recognition that the role of the computer user was more active, and even to the extent that there was a triadic relationship which involved each participant not only being influenced by the other partners, but also by the interactions between them.

The conclusion from the review was that those studies of tripartite communication which had noted the triadic relationship had studied the dialogues between the two human but had been informal, typically ethnographic, and those which had not recognised the triadic relationships had been controlled experiments or formal studies which had not paid attention to the dialogues between the two humans.

Further review was made of models of current HCI methods in controlled studies, human-human dialogue and the known effects of technology on human-human verbal dialogue, with the intention of building a suitable set of evaluative tools with which effects on the dialogues could be measured. It was clear that multiple measures would be necessary which individually might miss effects but together would build an accurate picture of the effects of the interfaces' variation.

The available measures were categorised according to the three Ps of PERCEPTION, PERFORMANCE and PROCESS. Measures of PERCEPTION were intended to capture the subjective effects of the variation in the interfaces through questionnaires. Measures of PERFORMANCE were intended to capture the effects of the interface variation on the two humans' performance of their tasks. Measures of PROCESS were intended to capture the effects of interface variation on the ways in which the human-human dialogues were structured.

6.1 Key Results from the Studies

These measures were applied in the evaluation of three variations on a simulated customer service task. In all cases, the task involved a customer contacting a remote company to arrange a trip around the United States, "The Travel Game". The customer knew what airports were available to them, but only the agent's computer held information about the flights which linked what links there were between the airports. The customer's aim was to build an itinerary which included as many airports as possible. Such a task was chosen because of its ecological validity; containing similar elements to those tasks observed in real customer service interaction (Section 2.0.1).

The first study compared two broadly different human-computer interaction styles and found their influences could be detected in the human-human dialogue. The second study investigated whether more subtle variation in the interface would permeate to the human-human dialogue, and was again successful. This was followed by two strands of research. The first looked at the process of the human-human dialogues in greater depth, to find evidence that analysis of the dialogues could inform the design of the interfaces. The second introduced a one-way video system into the tripartite communication to assess the effects of visibility on these tripartite relationships and the evaluative tools.

6.1.1 Comparison of scripted and open interfaces

The first study compared two broadly different human computer interaction styles which are common in customer service, scripted and open, and found effects on all three categories of measure employed.

The customers were found to perceive their agents as being easier to understand and more helpful when the consultations were open than when scripted; and their responses were noted to be more polarised when judging the agent's performance than when judging other aspects of the consultation and communication.

The analysis of the performance hinted at what the underlying effects of these perceptual differences were. Although there was no difference in the number of airports the customers fitted into their itineraries, the scripted consultations were over a minute and a half longer. The scripted interface seemed to be leading to problems in the agent's sharing of information.

Analysis of the process of the dialogues, looking at the number of turns and words and the amount of overlapping speech during the consultations, found that the most important effect of the different styles was on the amount of words the agent produced, though there was a suggestion that this was influencing the customer's production. The agent produced 292 (51%) more words and 3.2 (35%) more words per turn in the scripted interface than in the open, accounting for more than the total difference in both speakers' production. The effect was attributed to the agent leading the scripted dialogues, asking the customer to provide simple answers to extended questions, but the customer leading the open dialogues, asking the agent where was available to them.

The study had successfully shown in an experimental setting that broad differences in interaction style between an agent and their computer could permeate to the agent-customer dialogue and confirmed that the evaluative tools derived from the experimental study of mediated communication could capture these effects. The next step was to determine whether more subtle differences in the interface could be captured.

6.1.2 Comparison of prompting styles – audio-only

The second study applied the tools which had been validated in the first study to the analysis of smaller variations in the interface. The travel game was developed to simulate the developing practice of having service agents who take calls also be prepared to sell to the customer. As well as helping the customer to build an itinerary of flights around the United States, the agent was asked to sell a travel pass when prompted by their system. Four prompts were presented through either a dialogue box, an integral window or a spoken prompt, and the intention was to discover whether predicted effects of these presentation styles would be detectable in the human-human dialogue.

Analysis of questionnaires showed that the style of prompt had had little or no effect on the customer's perception of the consultation. Similarly, there were no effects on the participants' success in completing their tasks, with no style leading to larger or smaller itineraries or fewer or more sales. Detailed analysis of the sales process following each prompt detected none of the effects which would have been predicted from previous ergonomic work.

There were, on the other hand, sizable effects on the process of the whole consultations.

The spoken prompts lead to consultations taking 154 seconds (21%) longer to complete than the integral window, and this was due to an increase in the amount of words both speakers produced, with spoken prompted consultations containing on average 511 (33%) more words than the integral window and each turn containing on average 1.44 (19%) more words.

This provided good evidence that even small differences in the interface could have large impacts on the human-human dialogues, and was in line with suggestions about the different affective impacts of written and spoken language which would alter the relationship between the agent and their computer. It was contended that the spoken prompt could be leading to the computer being considered a commander, but the integral window to it being considered an adviser.

The observations that the design of a computer interface could impact on the human-human dialogue in tripartite communication had again been experimentally confirmed. The next steps were to see to what extent the human-human dialogues could be used to inform that design and what influence variation in the physical relationship between the three parties would have.

6.1.3 Dialogue Analysis

To determine whether analysis of the human-human dialogues could inform the design of interfaces to be used in tripartite interactions it was decided to subject the process of those dialogues to closer analysis. The method used was a functional coding scheme based closely on Conversational Games Analysis, but influenced also by previous findings about technology's potential influences on the content of dialogues.

This analysis lead to the discovery that the increase in words was strongly lead by the participants' use of two types of function; the description of new information and the asking of yes-no questions. There were 203.8 (45%) more words spent on EXPLAINs in spoken prompt consultations than in integral window consultations. There were 116.7 words (23%) more words spent on QUERY-YNs in spoken prompt consultations than in integral window consultations.

Although there was no interaction between the style of prompt and the amount of words each participant produced over the whole dialogues, the effect was found to be particularly concentrated in the agent's queries whether the customer wanted to buy the travel pass. These queries accounted for 38 (78%) of the 49 word difference between the agents' and customers' increase in use of QUERY-YN words over the whole dialogues.

The spoken prompt seemed to be compelling the agent to volunteer more information about their products, and this was affecting the whole nature of information sharing within the dialogues.

The use of in depth functional analysis of the process of the human-human dialogues had indicated what the influence of the spoken prompt was that lead to the longer dialogues, suggesting that such analysis could be used to inform the design of the interfaces.

6.1.4 Comparison of prompting styles – one-way video

In the final experimental chapter, the physical relationship between the three parties was altered by introducing a one-way video link through which the customer could see the agent, and which simulated future customer service over interactive television.

In this case, the influence of the prompts did not appear in the performance and process, but in the customer's perception of their consultations.

There was no effect of the style of prompt on the performance measures of the number of airports planned, the number of sales made or the time taken to complete the consultations. Neither was there any effect of the style of prompt on the number of turns or words used during the consultations.

On the other hand, the customers rated the consultations with the integral window more efficient, more friendly and generally better than those with the spoken, and generally better than those with the dialogue box.

These results were very different from those in the audio-only study, and it was speculated that the one-way video made the customer more aware of the influence of the computer on the agent because they could perceive the agent's being distracted from the building of the itinerary, but also better able to compensate for this influence by pre-empting the effects of the distraction on their dialogues.

To investigate these speculations further, statistical comparison was made between the results from the audio-only prompting study and from the one-way video prompting study; a comparison which was considered appropriate since the only alteration between the two was the inclusion of the one-way video link.

These analyses found that the inclusion of the one-way video had had overall effects on the perception and performance of the tasks, which seemed to show it benefiting the customer. The customers rated the one-way video consultations more satisfactory than the audio-only, and there was an average increase in the number of airports they fitted into their itineraries (1, 10%).

More importantly still, the video link had afforded the customer the ability to overcome the impairing effects of the spoken prompt. The spoken prompt had lead to an average 246 (40%) more words than the integral window in the customer's speech during the audio-only consultations, but an average 39 (6%) less words than the integral window during the one-way video consultations.

A foil for this was provided by the dialogue box. The customer's lower perception of its consultations compared with the integral window when they have the one-way video link, was accompanied by a decrease in the amount of words the customer produced in its consultations between the audio-only and one-way video studies (231 words, 43%).

The speculation that the customer was able to see the influence of the computer on their agent, and compensate for any detrimental effects, was supported and with it so was the contention that visibility could have a serious effect on the relationships within a tripartite communication.

6.2 Contributions to the literature

The research during this thesis contributes to the literature in a number of ways. Firstly, it supports earlier studies' observations about the nature of tripartite communication with findings from controlled laboratory experiments; increasing what is known about the influence of the agent-computer interaction within tripartite communication. Secondly, it has implications for the methodology of studies not only of tripartite communication but also of any human-human dialogue which interacts with technology. Finally, it has both substantive and methodological implications for Human Computer Interaction.

6.2.1 Tripartite communication

Tripartite communication involving two humans and a computer is becoming increasingly common, but until now controlled studies have tended to focus on the human-computer interaction and not considered its relationship with the third-party human. Those studies which have considered this relationship have been observational and for all the rigour in their methodologies, have provided no more than intriguing indications that there is an interaction between all three parties.

The thesis took a situation in which these sorts of tripartite communication is almost ubiquitous, that of remote customer service, simulated it as closely as possible in a controlled laboratory setting, and included structured analysis of the human-human dialogues to evaluate the interaction between variation in the human-computer interaction and the human-human dialogue.

The three studies and their in depth analyses not only supported the informal observations but also made basic findings about the nature of the interaction between the three parties.

6.2.1.1 The nature of tripartite communication

Two important findings were made about the nature of tripartite communication; that not only broad differences, but also smaller variations in the human-computer interaction can permeate to the human-human dialogue, and also that visibility plays as important a role in these interactions as it does in those involving only humans.

The first study showed that broad differences in the interactional style of the human-computer interface could be impact on the human-human dialogue, providing essential support to previous researchers' work, but it was in the second study that it became clear that the triadic nature of the tripartite communications could not be ignored. A variation in the system's presentation of four 5 lexical item messages lead to a 2 1/2 minute and 511 word difference in the customer service consultations it was supporting.

The dialogue analysis of this study suggested that the small variation had changed the whole dynamic of the interaction between the agent and their computer, and this in turn had altered the way in which information was passed in the agent-customer dialogue. The change in the style of prompt had lead to a change in how active or passive a role the computer was assigned – it being given the role of a commander when the prompts were spoken and that of an adviser when they were presented through the integral window.

The third study introduced one-way visibility into the tripartite interaction, to simulate an innovative customer service situation. This was intended to investigate the observation that reports of the greatest coordination between the three parties had occurred in those studies which also involved the greatest amount of co-visibility and co-presence. It was found that the added visibility permitted the customer increased awareness of the agent's interaction with their computer and the agent's visual feedback signals so that they were both more aware of the distracting prompts' presentation and able to overcome effects on their dialogues.

Overall, the strong contention that the interactions are triadic, with each participant influenced not only by the other participants but also by their interactions, was supported.

6.2.2 Methodology

Several contributions to psychological methodology were also made, the most basic of which was a confirmation that multiple measures are necessary when studying the interaction of technology with human-human communication. In each of the three studies, the influence of the variation of the interface was most evident in a different categories of measure. In the the comparison of open and scripted interfaces, it appeared in all three categories, in the audio-only comparison of styles of prompt, it appeared in the PERFORMANCE and PROCESS, in the one-way video comparison, it appeared in the PERCEPTION, and when the audio-only and one-way video studies were compared, it appeared in all three measures.

The major methodological development was the application of structured dialogue analysis in the study of the tripartite communication, and this proved capable, not only of capturing the effects of interface variation on the human-human dialogues, but also of indicating who the agent-computer interaction had lead to that variation. The measures of words and turns hinted at effects, which were given extra credence when combined with the dialogue coding scheme; the increase in the verbosity of agent's dialogues after the presentation of the spoken prompts directly indicated those prompts' effects.

On a more technical note, the development of Conversational Games Analysis, to account for the task orientation of the utterances whose function it captures, proved useful in indicating particular uses and sections of dialogue in which the influence of the interfaces was most evident. It formalised previous unstructured observations of the way in which technology could affect dialogue by such researchers as Wright and Monk (1990) and Newlands (1998).

6.2.3 Human-computer interaction

The contributions to the HCI literature will be discussed within the two frameworks for its research which are most prominent, the user modelling school which emphasizes the development of basic knowledge – typically in laboratory studies – about human-computer interaction on which user models can be built, and the contextual school, which emphasizes that human-computer interaction must be studies in context, and which is typically applied in the solution of practical design issues.

6.2.3.1 The User Modelling school

The contribution to the user modelling school of HCI, which was represented in the literature review by the formal and experimental work on interface consistency, is that the concept of user models may have to be adjusted to take account of situations where the model the agent-user has of how to use an interface has to accommodate itself to the model their customer has.

The thesis has shown how simple measures of user interaction, such as the time taken to press a key, though they can be built into complex models of individual instances of human-computer interaction (see Gray et al., 1990, 1992, 1993), may be insufficient to appreciate the effects interfaces may have on the agent user's dialogues with their customer. For instance, predictions which could be made from evaluative studies of technology in the human factors tradition proved insufficient to the purpose of predicting the effects of the different styles of prompt used in the second and third studies. Instead reference had to be made to work in social psychology and qualitative studies in HCI, CMC and CSCW which indicated different emotive influences the styles of prompt could have on the agent.

6.2.3.2 The Contextual school

There are two major contributions to the contextual school; one substantive and one methodological. The substantive contribution is the laboratory testing of observations made during naturalistic studies of tripartite communication, and the methodological contribution is the development of tools which could be applied in field studies. A further contribution may be an indication that Activity Theory's general imutable conception of computers as tools may have to be revised, in the face of some software applications.

Within social science the perfect progression of research is from contextual field studies to empirical testing of the observations in laboratory settings, to ensure causality has been correctly assigned, then back to the field to interpret the empirical results and focus future research (Richardson, 1996). The field work reviewed in Section 1.1 provides the initial contextual studies which suggested that there might be inter-relationships amongst the two humans and their computer partner, and personal observation suggested aspects of the inter-relationships which suggested suitable areas in which to study tripartite communication, but this thesis has closed the first loop of the research by providing the empirical support for those observations.

The methodological contribution is in the development of the evaluative techniques which have been used in the analysis of field studies, CSCW and CMC. The use of the widespread measures of the process of communication in the study of tripartite communication inherently indicates the relationship between the two, and the development of the Conversational Games Analysis scheme provides a new tool specifically intended to capture the impact of technology on dialogue in a more structured way than such analysis has typically been performed until now. Although those who study dialogue per se doubt that features of dialogue can be studied or even recognised in a foolproof fashion without a complete model of the dialogue around them, such as is provided by a coding scheme (Carletta, pers comm 1998), it is rare for such schemes to be applied in CSCW schemes. Instead researchers often try to pick out the features intuitively (e.g. O'Connaill, Whittaker and Wilbur, 1993)[1].

Although these coding schemes are not cheap to apply, in the face of the cost of developing, emplacing and supporting new software's use with hundreds of customer service agents, it may appear more feasible.

A further methodological development may come if the tripartite relationships are considered within the Activity Theory's action model. Typically, machines may only be considered tools which mediate activity, and this is suitable for the study of computers in standard desktop work and even in computer-mediated communication (Bødker, 1991; Kaptelinin, 1996). At first glance it appears to be suitable for the study of tripartite communication as well; the customer subject can be considered to use the agent tool to effect changes in the computer object, the agent subject can be considered to use their computer to effect changes in the company's databases, and the company might even be seen to use the computer to effect changes in the dialogue object between the agent and customer. It becomes less clear, however, how the computer should be considered if it can be shown to have direct effects on the customer which are not intended by the company or agent. In this situation, but also in others where computers may be given the roles

---

[1] The search for breakdowns under the aegis of Activity Theory (e.g. Bødker, 1996) contradicts this, somewhat, but Activity Theory works on a considerably different level from dialogue coding schemes.

of advisers[2], and hard as it may be to assign them intentions, it may be necessary to reconsider them as subjects.

6.3 Future Work

The directions this thesis leads in are as varied as the contributions it has made. The relationships between the three parties and their interactions could be investigated further. The methods for evaluating tripartite communication could be developed to take even more account of the inter-relationship between the parties, and this development could feed back into their use in the evaluation of other technologies.

6.3.1 Future Studies

One unspoken aim of the thesis has been to seed interest in the future study of tripartite communication between two humans and a computer, and future opportunities for investigating the relationships between the three parties are almost unlimited.

Just as has been found the study of mediated communication, it is likely that there will be interactions between co-presence, visibility, sharing of the computer display, the tasks the participants are involved in, synchronicity and quality of both the human-human dialogue and human-computer interaction and the experience of users. These are all multiplied two-fold by the additional feature of asymmetry. Although a conservative programme intended to further the building of a taxonomy of tripartite communication would see the investigation of two-way video, the sharing of the computer display and co-presence, even more unusual relationships may have real applications. It is not too far-fetched to imagine a blind person using a braille interface and a non-haptic display, depending on a third person to transfer information back to them.

If a programme were to begin, the next step would be to alter the relationship between the customer and the computer. Reflecting Bright's (1991) variation of the relationship between patients and computers in doctors' surgeries, the

---

[2] For instance, in the use of automated medical diagnosis software (e.g. Cicourel, 1990).

customer could be given different amounts of access to the output from the agent's software.

The hypotheses for such studies could be grounded in the findings from such studies as Olson, Olson and Meader's (1997) which found clear benefits to the sharing of whiteboards in design tasks, Raeithel and Velichkovsky (1996) who found benefits to being aware of collaborators' visual focus of attention in the use of CSCW software, Gaver, Sellen, Heath and Luff's (1993) Multiple Target Video study or Daly-Jones, Monk and Watts (1998) study of gaze awareness in video-conferencing. A theoretical background can be provided by Vygotsky (1962), Bruner (1981) and Clark (e.g. Clark and Brennan, 1991) who have all stressed the importance of being able to infer what a conversational partner is refering to.

An initial study, based on the prompting studies, could permit the customer just to see the building of their itinerary, probably allowing them to see the list of airports which are available to them as the agent accesses it, but not giving them visual access to the agent, nor making them aware of the prompts which the agent receives. This would permit the study of the customers' reaction to the agent receiving "secret" messages even though there appears to be open access to their computers' output, and would model situations such as simultaneous access to web pages in customer service such as has been suggested by Zhang et al. (1998). Particular reference might be made to Gaver, Sellen, Heath and Luff's (1993) observation of subjects who hid objects outside of their collaborators' view in the MTV study; to investigate further the effects of purposefully misleading anisotropic views.

A follow-up study could investigate the potential cancelling of this effect by giving the customer full access to all relevant information presented by the agents' software. The customers themselves could receive an automated message letting them know that purchasing opportunities have become available to them, and that they can enquire about more details from the agent.

In both the above cases, the visibility of the agent could be varied in the same way that it was in the studies reported in this thesis; giving the customer either audio-only or one-way video access to the agent. Two alternatives for adjusting this visibility would make natural extensions to the work which has already been performed or suggested.

The first is to introduce two-way video between the agent and customer. This option would parallel the introduction of symmetric access to the output from the agent's computer.

The combination of two-way vision and symmetric access to the agents' computers' output has already been studied in customer service situations (Anderson et al., 1997), but the isolated effects of two-way video and whiteboards, and their relationship to the effects of one-way video, has not been studied in customer service situations[3]. It is only with each variation in the complete taxonomy studied in isolation that causal relationships can be abstracted and an accurate model built.

A second is more unusual, and more in keeping with the design of previous studies which have emphasized the fact that the technology to make them practical is on our doorstep. Currently, low-cost two-way video such as might be applied in customer service suffers badly from the restrictions placed on it by the available technology (Watson and Sasse, 1996). This thesis presented one-way video as one more immediately available option, but one further option which would make even less demand on bandwidth would be the display of still photographs.

These are already used in television news broadcasts when correspondents are unable to provide video footage, and have been applied in trials of distance learning support software at British Telecom (Ellis, 1997). However, the effects of these on the process of collaborative tasks, whether cooperative or competitive has not been systematically studied.

Even if they were not found to have short-term effects during the performance of single collaborative tasks – and it is difficult to see how they would improve the coordination of any of the mechanisms in Whittaker and O'Conaill's (1997) taxonomy of conversational functions (Figure 5/1) – they might have less direct benefits. The mere inclusion of the still photograph, giving a greater sense of social presence with collocutors, may produce similar effects to those related to greater visibility in negotiative tasks (see Section 1.3.2.1.0), and may be

[3] They have been studied in other business applications (see Section 1.3.2; particularly Olson, Olson and Meader, 1997) but, as was discussed in Chapter 2, customer service has several unique features, not least its inherent asymmetry, and findings from other situations may not be directly applicable.

beneficial to longer term socialisation, just as Tang and Isaacs (1993) suggested suitably designed video-conferencing equipment might be[4].

The value of any such programme is clear from the different interactions between the styles of sales prompt and the human participants in the audio-only and one-way video comparisons of sales prompting styles.

6.3.2 Future Analysis

The most obvious improvement to the analysis would be to develop schemes for the analysis of participants' gaze and gesture which could be used to measure the distribution of their attention and how they use visibility to coordinate with the two other participants. The analysis of controlled studies using such a scheme could be used to test Greatbatch et al.'s (1993) observation that doctors used specific gestures to indicate the progress of their interaction with their computers and patients were able to use this to inform their initiation of turns. The findings from this thesis would indicate that the customers also use visibility to detect when there are unexpected events in the agent-computer interaction, and the scheme might also be used to test this through analysis of gaze around apparently disruptive events.

Such a scheme might also be developed to take account of the human-computer interactional turns. This could be useful in tripartite communication because it might explain the discrepancy between Greatbatch et al.'s (1993) observation of synchronisation between customer's turns and the doctor's turns with their computer, and the insensitivity of the measure of overlapping speech in this thesis' studies.

Such a scheme could also have benefit in the wider world of computer supported collaborative work and computer-mediated communication, where it could be used to capture the effects of collaborators or collocutors interactions with their systems. It is rare, in the real world, for any CSCW or CMC session not to

---

[4] Although long-term benefits may seem to be of less interest in the remote customer service interactions which were simulated in this study – where the participants may never encounter each other again – the long term benefits may accrue to the collaboration between the customer and the company.
It is, of course, predicted that these long-term effects may be observed in the short-term PROCESS of individual dialogues (Isaacs and Tang, 1994).

involve a considerable amount of time during which systems are configured to be mutually compatible or to overcome changes in the networks which connect them.

Such a development might be complex and time-consuming, however – gaze is notoriously difficult to analyse in any but the simplest laboratory setting – and the further application of the coding scheme developed in Chapter 4 may be more feasible.

The most obvious further application of the coding scheme would be in the analysis of the first and third studies in the thesis, but since it has been proved to be capable of capturing effects which have only been observed informally during previous applications of the scheme – for instance, Newlands' (1998) observation that much of the increase in EXPLAINs during her studies was not related to the task the speakers were engaged in – it might also be tested in other situations[5].

Its application might also provide opportunities to investigate further ways in which the coded dialogues can be analysed.

For instance, one possible explanation for the effects seen in the first study might be differences in the way the floor is controlled, with the agent controlling it more when their interfaces are scripted and the customer gaining more control when their agent's interface is open. If the dialogues were to be coded, one way to investigate the sharing of control would be to look at the proportion of dialogues which each participant spends in initiating and response moves. A higher proportion of initiating moves by one participant would indicate that they are taking control of the dialogues[6]. In the case of the first experiment it could be predicted that the open interface would lead to a higher proportion of initiating moves by the customer, and the scripted interface to a higher proportion of initiating moves by the agent.

The structure of the use of the conversational moves might also be investigated further. A relatively simple measure such as the proportion of moves which are embedded in other moves might give an indication of how difficult it was to achieve grounding (Clark and Brennan, 1991). A preliminary investigation of this

---

[5] Bearing in mind Bakeman and Gottman's (1997, p.15; Section 4.1.0) warnings about the application of coding schemes outside their original domain.

[6] As was mentioned in Chapter 4, initiating moves are not entirely independent of response moves, so it is impossible to draw conclusions from response moves.

following the coding of the second experiment found that it was insensitive to the effects of the style of prompt, but also suggested that the reliability of coding would have to be particularly high because a difference in the coding of a single move could have wider reaching effects on the embeddedness of following moves than it has on their categorisation.

Another option would be the conflation of categories. If the QUERY-W, QUERY-YN, CHECK and ALIGN moves are conflated into a single "QUERY" move, not only could it be predicted that coding would be simpler, but the three categories of initiating move would also represent the three core uses of language; making statements (EXPLAINs), making questions (QUERYs) and making commands (INSTRUCTs).

This would remove the potential criticism that QUERY-Ws and QUERY-YNs, and CHECKs and ALIGNs are taxonomised along different criteria. The two QUERYs are distinguished by the expected structure of the response, CHECKs refer back to information which has been given, and ALIGNs determine that the listeners' focus of attention is current. Although coders apparently find the scheme reliable, there is room for the categories to overlap, particularly since all CHECKs and ALIGNs are inherently questions expecting open or yes-no answers.

On the other hand, the sensitivity of the conflated category could be weaker than that of the split categories; and it is worth remembering that some comparisons have only found effects in ALIGNs (Doherty-Sneddon et al., 1997).

Another potential clarification to the coding scheme would be a change in the definition of QUERY-W and QUERY-YN moves. Currently, these are distinguished according to the response they require, but in a very restrictive way since only an expectation of "yes" or "no" warrants a QUERY-YN classification. Occasionally during the coding, it was felt that a better option would be to distinguish between questions which allowed open responses (which could be called QUERY-OPENs) and those which include a closed selection of responses (which could be called QUERY-CLOSED). QUERY-OPENs would be questions such as "Where would you like to go?" whereas QUERY-CLOSEDs would be questions such as "Would you like to go to Denver or Houston?" As with the measure of proportions of initiating moves which was mentioned above, this could help to capture differences in the distribution of control, since QUERY-OPENs essentially cede the floor without limitations, whereas QUERY-CLOSEDs expect the floor to be returned.

This would amount to a considerable change to the coding scheme, and its effects are impossible to predict. Certainly, in most dialogues, many moves which are currently classified as QUERY-Ws would become QUERY-CLOSEDs, and there would be fewer QUERY-OPENs than QUERY-Ws.

Another potential change would be to subcategorise the type of ON-TASK behaviour which is taking place in a conversational move, just as OFF-TASK categories have been subcategorised. A potential difficulty is that it may prove difficult to specify categories which are general enough to be as widely applicable as the current scheme of OFF-TASK subcategories was intended to be.

It might be that such sub-categories of ON-TASK moves would be better left task specific, in the same way Longabaugh's (1963) Resource Process Analysis *Resource* dimension is left unspecified. For instance, in coding the Travel Game, ON-TASK subcategories might refer to the part of the on-going flight which is being planned – destination or flight time – or in coding the map task, ON-TASK subcategories might refer to whether the participants are planning the beginning of the traversal around an object or the completion of a traversal.

A search for more generally applicable ON-TASK subcategories might begin with Morley and Stephenson's (1977) Resource dimension, or the categories related to the conduct and planning of meetings in the Olsons' scheme (Olson, Olson, Carter and Storrøsten, 1992; Olson, Olson and Meader, 1995). Such categories might capture whether a move is concerned with the definition of the next topic, the transfer of information related to topics, or the making of decisions.

The final development of the coding scheme does not concern the definition of its categories, but the way in which coded dialogues are analysed. Historically, once dialogues have been coded with Conversational Games Analysis, the analysis of the codings has been fairly quick and dirty – the majority of studies have merely compared the number of conversational games which occurred. This has been proved fairly successful, but as was shown in Chapter 4, further analysis of the coded dialogues can prove even more enlightening.

In that chapter, not only the counts of turns, but also counts of the words spent on turns were analysed. This proved to give a better indication of how much attention the participants were donating to each category of move.

More detailed analysis may also be valuable, and the most obvious development would be to look at the transitions between moves. The Olsons (Olson, Olson, Carter and Storrøsten, 1992; Olson, Olson and Meader, 1995) have produced simple measures of how much each turn is followed by others, but time series analysis as suggested by Bakeman and Gottmann (1997) might be even more beneficial. A further alternative would be the definition of the structure of dialogues to produce phrase structure trees such as might be produced during syntactic analysis of sentences.

It is difficult to predict how valuable such analyses might be. During the analysis of the CIQM dialogues (see Section 4.1.2.3) they were coded using the basic Conversational Games Analysis scheme, and preliminary attempts were made to produce parse trees of their dialogue structure, considering each coded utterance a leaf node, to compare the good and bad dialogues.

In this case, it was found that comparisons which could be used to produce similar conclusions about the dialogues as were generated during the CIQM focus groups required that multiple null nodes had to be proposed within the parse trees to account for the failure of the bad dialogues. For instance, one bad dialogue could be specifically criticised because the agent never told the customer what would be happening in the next phase of the call – for instance, rather than saying "For identification purposes, I'm going to need your name, please?", a poor agent might simply say "Your surname with initials?". This only appeared in the parse trees as gaps in the structure of the dialogues.

The prediction of where gaps should occur is undesirable, if not impossible, but may cause problems in drawing conclusions from methods of analysis, such as the Olsons' and time series analysis, which rely only on the surface behaviours seen in the dialogues.

6.3.3 Contextual Studies

What this last point implies, of course, is that the context of the consultations needs to be considered, and now that there is experimental evidence that the design of the computer can have an effect on the relationships between agent and customers, the context could be returned to, to discover what can be done to implement these findings.

The focus groups used during the CIQM project have given an abstract idea of what is important to customer during their consultations, observations of call centre staff and discussions with management may have been able to produce broad-brushed summaries of what is important to them, but a more focussed contextual inquiry might serve better to show how well these abstractions and summaries fit the actual practice of BT's customer service.

Uniting these with an analysis of the relationships between the call centre personnel, their management, computers, customers and even the designers of their software within an Activity Theory framework may indicate just what needs to be done to ensure the findings from the thesis are implemented, and implemented well[7].

It is fairly certain that management will have to be brought on board for any of recommendations to be implemented and for them to succeed, based on Bellamy's (1996) finding that a similar need arose in the implementation of educational software. It can also be predicted that the call centre personnel will have to included within the design and implementation, both so that improvements are relevant to everyday practice, and so that they are accepted once implemented. For the latter, it is worth reflecting on one of the few implementations of video-conferencing which has been successful; that at Boeing, where all the workers recognised that due to the time constraints in the design and construction of a new airplane they would not be able to waste time in travelling between multiple sites (Egido, 1990).

Such contextual studies would also complete the perfect circle of social science research, in which observation studies are used to discover issues which need testing, these are tested in the laboratory, and then the field is returned to to interpret the laboratory findings in context and decide what aspect of the issue would best be tested next (Richardson, 1996). As Galegher, Kraut and Egido (1990, p. XIV) said in the preface to their book:

Nothing is as generative of good theory as good practical problems

6.4 Application

It has already been argued by human factors experts at British Telecom (Millard, 1995) that the quality of the interaction between customer service agents and their customer can be improved by suitable implementation of their support systems.

---

7 It may, of course, also remind us that the requirements made of such technologies are often extreme complex, or changing, and that the implementation process will be formidable indeed (Galegher and Kraut, 1990).

Although the primary concern of this thesis was to improve basic knowledge about tripartite communication, rather than to answer specific practical questions, it is also possible to draw conclusions which can be applied in the design of real customer handling systems and influence customer service practice.

Broadly speaking, the findings have implications in three areas; decisions about what is important in customer service, the training of customer service agents, and the design of future customer handling systems.

6.4.1 Defining what is important in customer service

The implications for the definition of what is important in customer service are important for two reasons, firstly, for the inherent value of improving customer service, but secondly because this is the level at which management will appreciate changes in the design of software, and as was intimated in the penultimate paragraph of the last section (Section 6.3.3), it is unlikely that any of these results will be applied unless the management which sponsors the design perceives their benefit.

The strong implication from all three studies is that customers like to feel in control of their dealings with companies. In the comparison of open and scripted interface styles, the customers perceive the agent as less helpful and easy to understand when the scripts lead to their controlling the dialogues. Similarly, a consideration of the two prompting studies shows the customer preferring the one-way video to the audio-only; apparently because it allows them to overcome the effects of the prompts which threaten to take the control of the dialogues away from them.

The implication is that customer service dialogues must be supported in such a way that the customer feels in control of them.

6.4.2 Training of customer service agents

Clearly, if the implication is that customers like to feel in control of their consultations with companies, it is necessary that customer service agents are trained to support this feeling.

This can be fostered in simple ways, such as those uncovered in BT's CIQM study (Marle and Coyle, 1995); which amount to a large extent to being polite. More generally, it can be inferred from the studies in this thesis, that the agents should be encouraged to give the customer equal access to the floor in the consultations' dialogues.

Obviously, there is information which the customer will need to give, and therefore the agent will have to request it from the customer, but this should be done in such a way that the customer neither feels imposed upon, nor that the agent is trying to take control of the dialogues as they make the request. This means both that the agent should not ask questions bluntly, but also that they should not be too verbose.

This verbosity can either be seen as unhelpful in itself – as in the comparison of open and scripted interfaces – or can influence the customer to be more verbose themselves – as in the audio-only comparison of prompting styles – which can be inferred to be seen as perceived as unsatisfactory from the way the customer controls their verbosity better in the one-way video comparison of prompting styles.

It is worth considering, however, that the extensive training which has been used in the past is being supplanted by increased support, both in the way customer service teams are structured and in the design of agent's software (Millard, pers comm 1998).

6.4.3 Design of customer handling systems

The implications for the design of future customer handling systems are clearest. In the absence of other influences, open interface design should be perferred to scripted, and non-invasive means should be perferred when giving alerts about new information in the agent-computer interface.

Both of these changes seem to improve the customer's feeling of control within the dialogues; the open interface because the agent does not monopolise the dialogue, and the less invasive prompting methods because they are less likely to be perceived as distracting the agent from their interaction with the customer.

The qualification, "in the absence of other influences", is important, however, because there are situations in which practical demands may over-ride these guide-lines; such as requirements that the agents' dialogues contain exact wordings for legal reasons in the remote sale of insurance.

A second recommendation concerns the inclusion of one-way video. Which appears to increase the customers' feeling of control during the dialogues, and could also be considered for future implementation.

The caveats here are that the design of the agent-computer interface becomes even more important because the customer appears to be more readily able to perceive its distractions to the agent, and some consideration will have to be given to the context in which the agent works. If the agents are to viewed as well as heard, then both their appearance and the appearance of their offices may influence the customers' perception of the consultation, and the company which employs them. A slovenly agent will rarely be perceived well, nor, most likely, would a view of a cluttered, unprofessional office.

6.5 Students' Reflections

At the end of the thesis it is worth reflecting on the process of the research; its achievements, mistakes, and what has been learned.

6.5.1 Achievements

Looking back over the research which underlies this thesis, the achievements which seem most important to the author are the practical ones.

Building software which modelled customer service interactions as closely as possible, based on observations and discussions with British Telecom experts, but which also contained tasks with definable results so that performance could be reliably compared, was the first of these.

An even bigger personal achievement was the development of the task orientation dimension for Conversational Games Analysis, which drew from considerable experience in coding gained during the first year of research training, in depth review of previously developed coding schemes, observations of real customer service calls and discussion of what was important within them, and acknowledgement of aspects which had been considered important in CMC and CSCW studies. This is the achievement which the author most wants to share with the academic community.

In terms of research findings, the insight during the analysis of the prompting studies that the effects which were being observed could best be explained by anthropomorphising the computer's role was a defining moment. It lead the research away from fairly systematic and technical consideration of the immediate effect of prompts, and towards a more holistic consideration of the relationships between the parties.

6.5.2 Mistakes

The mistakes which have been made are far more widespread than those which the author is willing to admit to.

Perhaps, the most important mistake was progressing without a programme clearly laid out, and without a clear enough idea of how results might fit in to the literature. This lead to the first study, which the author considered for some time to have been a false start because the interfaces were so different from each other that the results would be irrelevant. It was only with further consideration of the results, interpreting measures of PERCEPTION, PERFORMANCE, and PROCESS in concert that the value of the study became apparent.

Not entirely a mistake, but certainly something the author would like to return to, and reconsider, is the design of the questionnaires. Sadly, it is likely that they could only have been improved with hindsight. The questionnaire used in the first study had inconsistently focused questions, and the responses to them were difficult to interpret. Although the questionnaires used in the second and third studies resolved some of its failings, having been developed while the results from the first study were still being analysed, it was impossible to draw conclusions about how all the questions could have been better focussed. This lead to a fairly heavy questionnaire which had a number of pointless items.

The final mistake, and one it would be difficult to have avoided, was that of intending to do to much; and the consequent disappointment when these intentions were not realised. An initial aim, as has been indicated above, had been the development of the means to formalise the structure of the functional process of dialogues. In the face of the demands of gathering the data from three fairly large studies and their already extensive analysis, the analysis of the structure evaporated. The author suspects that the setting of reasonable goals is one of the most important learning experiences during his, and any, doctoral research programme.

6.5.3 Learned

Practically, reflecting on what has been learnt has allowed the author to realise how their knowledge of the fields of social science in general, and HCI, its sub-branches CMC and CSCW, and the field of applied linguistics in particular has increased. Entering the doctoral research, the author had a general, but fading, background in psychology, and a current, but supremely unpractical, knowledge of theoretical linguistics – having just completed a master's thesis on the internal functional structure of the copula "be".

The learning curve, to acquire a feel for the nature of applied research was steep, and often frustrating, but once learnt has been rewarding. It has allowed the author to see the value in techniques and models which had previously seemed vague and over-generative, and given him an appreciation for the complexity of researching real world problems.

Probably the most important thing which has been learnt about the process of research, is the value of what for a long time the author termed serendipity, but which he has discovered CSCW researchers may have more accurately named proximity (Kraut, Egido and Galegher, 1990). Kraut, Egido and Galegher's term encapsulates the interpretation of research as a social enterprise, in which knowledge and plans develop as researchers bounce ideas off each other both in formal and informal settings. The serendipity the author referred to included the occasions on which just the right piece of research was found to be sitting in a conference proceedings which had been lying on his desk unopened for months, or in the article preceding the one he had sought out in the library.

References

Akmajian, A., Demers, R.D. and Harnish, R.M., 1984, Linguistics: An Introduction to Language and Communication, Second Edition, MIT Press: Cambridge, MA.

Anderson, A.H., 1992, Negotiating coherence in dialogue, In Gernsbacher, M.A. and Givon, T. (Eds.), 1995, Coherence in Spontaneous Text, John Benjamin: Amsterdam, pp. 1-40.

Anderson, A.H., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C. and Thompson, H., 1991, The HCRC Map Task Corpus, Language and Speech, 34, 351-360.

Anderson, A.H., Mullin, J., Newlands, A., Doherty-Sneddon, G. and Fleming, A.M., 1994, Video-mediated communication in CSCW: Effects on communication and collaboration. Position paper for Workshop on Video-mediated communication: Testing, Evaluation and Design Implications, Organised by Finn, K., Sellen, A. and Wilbur, S., CSCW '94, Chapel Hill: NC, October 1994.

Anderson, A.H., Newlands, A., Mullin, J., Fleming, A.M., Doherty-Sneddon, G., and Van der Velden, J., 1996, Impact of video-mediated communication on simulated service encounters. Interacting with Computers, 8, 2, 193-206.

Anderson, A.H., O'Malley, C., Doherty-Sneddon, G., Langton, C., Newlands, A., Mullin, J., Fleming, A.M., Van der Velden, J., 1997, The Impact of VMC on Collaborative Problem Solving: An analysis of Task Performance, Communicative Process and User Satisfaction, In Finn, K., Sellen, A., and Wilbur, S. (Eds), Video Mediated Communication, Lawrence Erlbaum Associates: Hillsdale, NJ.

Anderson, C.W. and McMaster, G.E., 1982a, Computer assisted modeling of affective tone in written documents, Computers and the humanities, 16, 1-9.

Anderson, C.W. and McMaster, G.E., 1982b, Objective analysis of emotional tone in stories and poems, ALLC Journal, 3, 2, 45-51.

Anderson, C.W. and McMaster, G.E., 1986, Modelling emotional tone in stories using tension levels and categorical states, Computers and the Humanities, 20, 3-9.

Anderson, C.W. and McMaster, G.E., 1987, Quantification of writing by the Brothers Grimm: A comparison of successive versions of three tales, Paper presented at the International Conference on Computers and the Humanities, April 1987, Columbia, SC.

Anderson, C.W. and McMaster, 1992, Pantextual indices of emotional tone, In Nardocchio, E.F. (Ed.) Reader response to literature: The empirical dimension, Mouton de Gruyter: New York.

Anderson, C.W. and McMaster, G.E., 1993, Emotional tone in Peter Rabbit before and after simplification, Empirical Studies of the Arts, 11, 2, 177-185.

Argyle, M. and Dean, J., 1965, Eye-contact, distance and affiliation, Sociometry, 28, 289-304.

Argyle, M. and Ingham, R., 1972, Gaze, Mutual Gaze and Proximity, Semiotica, 6, 1, 32-49.

Argyle, M., Ingham, R., Alkema, F. and McCallin, M., 1973, The Different Functions of Gaze, Semiotica, 7, 19-32.

Argyle, M., Lalljee, M. and Cook, M., 1968, The effects of visibility on interaction in a dyad, Human Relations, 21, 3-17.

Austin, J., 1962, How to do Things with Words, Urmson, J.O. (Ed.), Clarendon Press: Oxford.

Baber, C., Stanton, N.A. and Stockley, A., 1992, Can speech be used for alarm displays in "process" control type tasks? Behaviour and Information Technology, 11, 4, 216-226.

Bach, K. and Harnish, R., 1979, Linguistic Communication and Speech Acts, MIT Press: Cambridge, MA.

Baddeley, A.D. and Hitch, G.J., 1974, Working Memory, In Bower, G. (Ed.) Recent advances in learning and motivation, Vol. VIII, Academic Press: New York.

Baecker, R.M., Grudin, J., Buxton, W.A.S. and Greenberg, S., 1995, Reading in Human Computer Interaction: Toward the Year 2000. 2nd Ed., Morgan Kaufman: San Francisco, California.

Bakeman, R. and Gottman, J.M., 1997, Observing Interaction: An Introduction to Seequential Analysis, 2nd Ed., Cambridge University Press: Cambridge, UK.

Bales, R.F., 1950, Interaction Process Analysis: a method for the study of small groups, Addison-Wesley: Cambridge, MA.

Bellamy, R., 1996, Designing educational technology: Computer-mediated change, In Nardi, B.A. (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction, Cambridge, MA, MIT Press.

Besson, M., Faita, F., Czternasty, C., and Kutas, M., 1997, What's in a pause: event-related potential analysis of temporal disruptions in written and spoken sentence, Biological Psychology, 46, 3-23.

Beyer, H. and Holtzblatt, K., 1998, Contextual Design: Defining customer-centred systems, Morgan Kaufman: San Francisco, CA.

Blattner, M.M., Sumikawa, D.A. and Greenberg, R.M., 1989, Earcons and icons: their structure and common design principles, Human-Computer Interaction, 4, 11-44.

Blomberg, J.L. and Henderson, A., 1990, Reflections on Participatory Design: Lessons from the Trillium experience. Proceedings of Computer and Human Interaction (CHI) '90, ACM Press, pp. 353-359.

Bødker, S., 1991, Through the Interface: A Human Activity Approach to User Interface Design, Hillsdale, NJ: Lawrence Erlbaum.

Bødker, S., 1996, Applying Activity Theory to Video Analysis: How to Make Sense of Video Data in HCI, In Nardi, B. (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction, Cambridge, MA, MIT Press.

Boyle, E., Anderson, A.H. and Newlands, A., 1994, The effects of visibility on dialogue performance in a cooperative problem solving task, Language and Speech, 37, 1, 1-20.

Brennan, S.E., 1991, Conversations with and through computers, User Modelling and User-Adapted Interaction, 1, 67-86.

Brewster, S.A., Wright, P.C. and Edwards, A.N., 1992, A detailed investigation into the effectiveness of earcons, In Kramer, G. (Ed.), Auditory display, sonification, audification and auditory interfaces, The Proceedings of the First International Conference on Auditory Display, Santa Fe Institute, Addison-Wesley: Santa Fe: NM, pp. 471-498.

Brewster, S.A., 1997, Using non-speech sound to overcome information overload, Displays, Special Issue on multimedia displays, 17, 179-189.

Bright, S., 1991, Nearest and Dearest, British Journal of Healthcare Computing, 8, 58-59.

Bruner, J., 1981, The pragmatics of acquisition, In Deutsch, K., The Child's Construction of Language, Academic Press: New York.

Bull, M. and Aylett, M., 1998, An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue, Proceedings of ICSLP '98, Sydney, Australia.

Byblow, W.D. and Corbett, J.T., 1989, Effects of linguistic redundancy and coded voice warnings on system response time, Applied Ergonomics, 20, 105-108.

Card, S.K., 1982, User perceptual mechanisms in the search of computer command menus, Proc. Human Factors in Computer Systems, Washington, D.C., Chapter of ACM (March 1982), pp. 190-196.

Card, S.K., Moran, T.P., and Newell, A., 1983, The Psychology of Human-Computer Interaction. Erlbaum: Hillsdale, NJ.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., 1996, HCRC dialogue structure coding manual, Human Communication

Research Centre (HCRC) Publication, http://www.hcrc.ed.ac.uk/CARLJ962.html.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., Anderson, A.H., 1997, The reliability of a dialogue structure coding scheme, Computational Linguistics, 23, 1, 13-32.

Carroll, J.M., 1982, Learning, using and designing command paradigms, Human Learning, 1, 31-62.

Carroll, J.M., 1997, Human-comptuter interaction: psychology as a science of design, International Journal of Human-Computer Studies, 46, pp. 501-502.

Chapanis, A., 1975, Interactive Human Communication, Scientific American, 232, 36-42.

Chomsky, N., 1965, Apects of the Theory of Syntax, MIT Press: Cambridge, MA.

Cicourel, A.V., 1990, The integration of distributed knowledge in collaborative medical diagnosis, In Galegher, J.R., Kraut, R.E., and Egido, C. (Eds.), Intellectual Teamwork: Social and technological foundations of cooperative work, Lawrence Erlbaum Associates: Hillsdale, NJ.

Clark, H.H. and Brennan, S.E., 1991, Grounding in Communication. In J. Levine, L.B. Resnick and S.D. Behrend (Eds.), Shared Cognition: Thinking as Social Practice. APA Books: Washington DC.

Clark, H.H. and Marshall, C., 1981, Definite Reference and Mutual Knowledge, In Joshi, A., Webber, B. and Sag, I. (Eds.), Elements of Discourse Understanding, Cambridge University Press: Cambridge, UK, pp. 10-63.

Clark, H.H. and Schaefer, E.F., 1987, Collaborating on contributions to conversations, Language and Cognitive Processes, 2, 19-41.

Clark, H.H. and Schaefer, E.F., 1989, Contributing to Discourse, Cognitive Science, 13, 259-94.

Clark, H.H. and Wilkes-Gibbs, D., 1986, Referring as a collaborative Process, Cognition, 22, 1-39.

Cohen, P.R., 1984, The pragmatics of referring and the modality of communication, Computational Linguistics, 10, 2, 97-146.

Cohen, B.H., 1997, Explaining Psychological Statistics, Duxbury.

Cook, M. and Lalljee, M., 1972, Verbal substitutes for visual signals in interaction, Semiotica, 6, 212-221.

Cotton, J.C. and McCawley, M.E., 1983, Voice technology design guides for Navy training systems: Final report for the period 23 April 1980-2 January 1982 (Report No. NAVTRAEQUIPCEN 80-C-0057-1), Naval Training Equipment Center: Orlando: FL.

Cruikshank, P.J., 1984, Computers in medicine: patients' attitudes, J. Royal College of General Practitioners, 34, 77-80.

Daft, R.L. and Lengel, R.H., 1986, Organizational information requirement: Media richness and structural design, Management Science, 32, 5, 554-571.

Daly-Jones, O., Monk, A. and Watts, L., 1998, Some advantages of video-conferencing over high-quality audio conferencing: fluency and awareness of attentional focus, Int. J. Human-Computer Studies, 49, 21-58.

Davies, M., 1971a, Cooperative Problem Solving, Communications Studies Group paper no. E/71159/DV.

Davies, M., 1971b, Cooperative Problem Solving: A follow-up study, Communications Studies Group paper no. E/71252/DV.

Deatherage, B.H., 1972, Auditory and Other Sensory Forms of Information Presentation, In Van Cott, H.P. and Kinkade, R.G. (Eds.) Human Engineering Guide to Equipment Design (Revised Edition), McGraw-Hill.

Doherty-Sneddon, G. Anderson, A.H., O'Malley, C., Langton, S., Garrod, S. and Bruce, V., 1997, Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance, Journal of Experimental Psychology: Applied, 3, 2, 105-123.

Duncan, S. and Fiske, D.W., 1977, Face-to-face interaction: Research, methods and theory, Lawrence Erlbaum: Hillsdale, NJ.

Egido, C., 1990, Teleconferencing as a technology to support cooperative work: Its possibilities and limitations, In Galegher, J.R., Kraut, R.E., and Egido, C. (Eds.), Intellectual Teamwork: Social and technological foundations of cooperative work, Lawrence Erlbaum Associates: Hillsdale, NJ.

Ellis, N., 1997, Merlin Gamma Trial: Educational Evaluation Report, Draft C, BT Internal Document.

Engestrøm, Y., 1987, Learning by Expanding, Helsinki: Orienta-konsultit.

Ericsson, K. and Simon, H.A. , 1980, Verbal Reports as Data, Psychological Review, 87, 3, 215-251.

Fitter, M.J. and Cruikshank, P.J., 1982, The computer in the consulting room: a psychological framework, Behaviour and Information Technology, 1, 81-92.

Furnham, A., 1982, The message, the context and the medium, Language and Communication, 2, 1, 33-47.

Gaines, B.R. and Shaw, M.L.G., Foundation of dialog engineering: the development of human-computer interaction. Part II, International Journal of Man-Machine Studies, 24, 101-123.

Gale, S., 1990, Human aspects of interactive multimedia communication, Interacting with Computers, 2, 175-189.

Galegher, J.R. and Kraut, R.E., 1990, Technology for Intellectual Teamwork: Perspectives on research and design, In Galegher, J.R., Kraut, R.E., and Egido, C. (Eds.), Intellectual Teamwork: Social and technological foundations of cooperative work, Lawrence Erlbaum Associates: Hillsdale, NJ.

Gaver, W., 1986, Auditory icons: Using sound in computer interfaces, Human-Computer Interaction, 2, 167-177.

Gaver, W., 1989, The Sonicfinder: An interface that uses auditory icons, Human-Computer Interaction, 4, 67-94.

Gaver, W., 1992, The affordances of media spaces for collaboration, In Proceedings of CSCW '92, Toronto, ACM Press.

Gaver, W., 1996, Affordances for interaction: the social is material for design, Ecological Psychology, 8, 2, 111-129.

Gaver, W., Sellen, A., Heath, C.C., and Luff, P., 1993, One is not enough: multiple views in a media space, In Proceedings of the ACM INTERCHI '93 Conference on Human Factors in Computing Systems, ACM Press: New York, pp. 335-341.

Gaygen, D.E. and Luce, P.A., 1998, Effects of modality on subjective frequency estimates of spoken and printed words, Perception and Psychophysics, 60, 3, 465-483.

Gernsbacher, M.A., 1984, Resolving 20 years of inconsistent interaction between lexical familiarity and orthography, concreteness and polysemy, Journal of Experimental Psychology: General, 113, 256-281.

Gibson, J.J., 1979, The ecological approach to visual perception, Houghton Mifflin: New York.

Goldberg, G.N., and Mettee, D.R., 1969, Liking and perceived comunication potential as determinants of looking at another, Psychonometric Science, 16, 277-78.

Good, M., 1989, Seven experiences with contextual field research, SIG-CHI Bulletin, 20(4), 25-33.

Good, M., 1992, Participatory Design of a portable torque-feedback device, Proceedings of Computer and Human Interaction (CHI) '92, ACM Press, pp. 225-232.

Gray W.D., Johns B.E., Atwood, M.E., 1992, A Precis of Project Ernestine or An Overview of a Validation of GOMS. Proceedings of Computer and Human Interaction (CHI) '92, ACM, pp. 307-312.

Gray W.D., Johns B.E., Atwood, M.E., 1993, Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. Human-Computer Interaction, 8, 3, 237-304.

Gray W.D., Johns B.E., Stuart, R., Lawrence, D. and Atwood, M.E., 1990, GOMS meets the Phone Company: Analytical Modelling Applied to Real-World Problems, In Diaper, D., Gilmore, G., Cockton, G. and Shackel, B. (Eds.), Human-Computer Interaction – INTERACT '90, North-Holland: Elsevier Science Publishers.

Greatbatch, D., Luff, P., Heath C., and Campion, P., 1993, Interpersonal communciation and human-computer interaction: and examination of computers in medical consultations, Interacting with Computers, 5, 2, 193-216.

Grice, H.P., 1975, Logic and Conversation, In Cole, P. and Morgan, J. (Eds.) Speech Acts (Syntax and Semantics, Vol. 3), Academic Press: New York.

Grice, M. and Savino, M., 1995, Intonation and communicative function in a regional variety of Italian, In Phonus 1, pp. 19-32, Institute of Phonetics, University of Saarland.

Grudin, J., 1989, The case against user consistency, Communications of the ACM, 32, 1164-1173.

Halliday M.A.K., 1989, Spoken and Written Language, 2nd Ed., Oxford University Press.

Harman, G., 1977, Review of "Linguistic Behaviour" by Jonathan Bennet. Language, 53, 417-424.

Heath, C.C., and Luff, P., 1992a, Collaboration and Control: Crisis management and multimedia technology in London Underground line control rooms, Computer Supported Cooperative Work, 1, (1-2), 69-94.

Heath, C.C., and Luff, P., 1992b, Media space and communicative assymetries: Preliminary observations of video mediated interaction, Human-Computer Interaction, 7, 99-103.

Heinicke, C. and Bales, R.F., 1953, Developmental trends in the structure of small groups, Sociometry, 16, 7-38.

Heise, D.R., 1965, Semantic differential profiles for 1000 most frequent English words, Psychological Monographs: General and Applied, 79, 8, 1-31.

Henderson, R., Podd, J., Smith, M. and Varela Alvarez, H., 1995, An examination of four prominent user-based software evaluation methods, Interacting with Computers, 7, 4, 412-432.

Hodgdon, 1990, ShowDialog Version 1.6, Hypercard Shareware.

Hole, L., Crowle, S. and Millard, N., 1998, The Motivational User Interface, Proceedings of HCI '98.

Holtzblatt, K. and Jones, S., 1993, Contextual Inquiry: a participatory technique for system design, In Schuler, D. and Namioka, A. (Eds.), Participatory Design: Principles and Practice, Lawrence Erlbaum Associates: New York.

Hone, K.S., Graham, R., Maguire, M.C., Baber, C. and Johnson, G.I., 1998, Speech Technology for automated teller machines: an investigation of user attitude and performance, Ergonomics, 41, 7, 962-981.

Houghton, A., Gardner, M.J. and Gould, P., 1991, ESCFE: Expert Systems for Customer Facing Environments, IEEE Journal on Selected Areas in Communcations, 9, 4.

Houghton, G and Isard, S.D., 1987, Why to speak, what to say, and how to say it. in P. Morris (Ed.) Modeling Cognition, Wiley.

Howell, D.C., 1997, Statistical Methods for Psychology, Wadsworth.

Isaacs, E.A. and Tang, J.C., 1994, What video can and cannot do for collaboration: a case study, Multimedia Systems, 2, 63-73.

Kaptelinin, V., 1996a, Computer-Mediated Activity: Functional Organs in Social and Developmental Contexts, In Nardi, B.A. (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction, Cambridge, MA, MIT Press.

Kaptelinin, V., 1996b, Activity Theory: Implications for Human Computer Interaction, In Nardi, B.A. (Ed.), Context and Consciousness:

Activity Theory and Human-Computer Interaction, Cambridge, MA, MIT Press.

Kellogg, 1987, Conceptual Consistency in the User Interface: Effects on User Performance, In Bullinger, H.J. and Shackel, B. (Eds.), Human-Computer Interaction – INTERACT '87, Elsevier Science Publishers: North-Holland.

Kendon, A., 1967, Some functions of gaze direction in two-person conversation, Acta Psychologica, 26, 22-63, Reprinted in Kendon, A., Conducting Interaction: Patterns of behaviour in focused encounters, Cambridge University Press: Cambridge, UK.

Kiger, J.I., 1984, The depth/breadth trade-off in the design of menu-driven user interfaces, International Journal of Man-Machine Studies, 20, 201-213.

Kobayashi, M., Shinozaki, M., Sakairi, T., Touma, M., Daijavad, S., Wolf, C., 1998, Collaborative customer services using synchronous web browser sharing, Proceedings of CSCW '98, pp. 99-108

Kowtko, J.C., 1997, The function of intonation in task-oriented dialogue, Doctoral thesis available as Research Paper HCRC-RP84, Human Communications Research Centre: University of Edinburgh.

Kowtko, J.C., Isard, S.D. and Doherty-Sneddon, G.M., 1992, Conversational Games within Dialogue, Proceeding of the Workshop on Discourse Coherence, Human Communications Research Centre, University of Edinburgh.

Kraut, R.E., Egido, C. and Galegher, J.R., 1990, Patterns of contact and communication in scientific research collaboration, In Galegher, J.R., Kraut, R.E., and Egido, C. (Eds.), Intellectual Teamwork: Social and technological foundations of cooperative work, Lawrence Erlbaum Associates: Hillsdale, NJ.

Kraut, R.E., Miller, M.D. and Siegel, J., 1996, Collaboration in Performance of Physical Tasks: Effects on Outcomes and Communication, Proceedings of ACM CSCW '96 Conference on Computer-supported cooperative work, pp. 57-66.

Krippendorff, K., 1980, Content Analysis: An introduction to its methodology, Sage Publications.

Kuuti, K., 1996, Activity Theory as a Potential Framework for Human Computer Interaction Research, In Nardi, B.A. (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction, Cambridge, MA, MIT Press.

Labov, W. (1966), Hypercorrection by the lower middle class as a factor in linguistic change, In Bright W. (Ed.), Sociolinguistics, Mouton: The Hague.

Labov, W. and Fanshel, D., 1977, Therapeutic Discourse, Academic Press: New York.

Landis, J.R. and Koch, G.G., 1977, The measurement of observer agreement for categorical data, Biometrics, 33, 159-174.

Lawrence, D., Atwood M.E. and Dews, S., 1994, Surrogate Users: Mediating Between Social and Technical Interaction, Proceedings of Computer and Human Interaction (CHI) '94, ACM Press: Boston, MA, pp. 399-404.

Lawrence D., Atwood M.E., Dews S. and Turner T., 1990, Social Interaction in the use and design of a workstation: Two contexts of interaction, in Thomas, P.J. (Ed.), The Social and Interactional Dimensions of Human-Computer Interfaces, Cambridge University Press: Cambridge, UK.

Leont'ev, A.N., 1974, The problem of activity in psychology, Soviet Psychology, 13 (2), pp. 4-33.

Leont'ev, A.N., 1978, Activity, Consciousness, and Personality, Prentice-Hall: Englewood Cliffs, NJ.

Liederman, R., 1990, The Telephone Book, McGraw-Hill.

Longabaugh, R., 1963, A category system for coding interpersonal behaviour as a social exchange, Sociometry, 26, 319-344.

Luff, P., 1990, Introduction, in Luff, P., Gilbert, G.N. and Frohlich, D.M. (Eds.), 1990, Computers and Conversation, Academic Press.

Luff, P., Gilbert, G.N. and Frohlich, D.M., 1990, Computers and Conversation, Academic Press.

Marle, J. and Coyle G., 1995, Validation of Customer Interface Quality Measurement Tool, BT Human Factors Unit Internal Publication.

Martin, D., Bowers, J., and Wastell, D., 1997, The Interactional Affordances of Technology: An Ethnography of Human-Computer Interaction in an Ambulance Control Centre, in Thimbleby, H., O'Conaill, B., and Thomas, P., (Eds.) People and Computers XII: Proceedings of Human-Computer Interaction '97, pp. 263-281.

McCarthy J.C. and Monk A., 1994, Measuring the quality of computer-mediated communication. Behaviour and Information Technology, 1994, 13, 5, 311-319.

Millard, N., 1995, Best Practice: Final Report, Future Customer Handling Systems Report. BT Internal Document.

Monk, A., McCarthy, J., Watts, L., and Daly-Jones, O., 1994, Measures of Process. in M. MacLeod and D. Murray (Eds) Evaluation for CSCW, Springer-Verlag: Berlin.

Morley, I.E. and Stephenson, G.M., 1969, Interpersonal and interparty exchange: a laboratory simulation of an industrial negotiation at the point level. British Journal of Psychology, 60, 543-545.

Morley, I.E. and Stephenson, G.M., 1970, Formality in experimental negotiations: a validation study, British Journal of Psychology, 61, 383.

Morley, I.E. and Stephenson, G.M., 1977, The Social Psychology of Bargaining, George Allen Unwin: London.

Mountford, S.J. and Gaver, W., 1990, Talking and listening to computers, In Laurel, B. (Ed.), The Art of Human-Computer Interface Design, Addison-Wesley: Reading, MA, pp. 319-334.

Muller, M.J., Carr, R., Ashworth, C., Diekmann, B., Wharton, C., Eickstaedt, C., and Clonts, J., 1995, Telephone Operators as Knowledge Workers: Consultants Who Meet Customer Needs. Proceedings of Computer and Human Interaction (CHI) '95, ACM Press.

Nardi, B.A., Ed., 1996a, Context and Consciousness: Activity Theory and Human-Computer Interaction, Cambridge, MA: MIT Press.

Nardi, B.A., 1996b, Activity Theory and Human Computer Interaction, In Nardi, B.A. (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction, Cambridge, MA, MIT Press.

Nardi, B.A., 1996c, Some Reflections on the Application of Activity Theory, In Nardi, B.A. (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction, Cambridge, MA, MIT Press.

Nielsen, J., Ed., 1989, Coordinating User Interfaces for Consistency, SIG-CHI Bulletin, 20, 3, 63-65.

Newell, A. and Card, S., 1985, The prospects for psychological science in Human-Computer Interaction. Human Computer Interaction, 1, 3, 209-242.

Newlands, A., 1998, The Effects of Computer Mediated Communication on the Processes of Communication and Collaboration, Unpublished Phd. Thesis, Glasgow University: Department of Psychology.

Norman, M. and Thomas, P., 1990, The Very Idea: Informing Design from Conversation Analysis, in Luff, P., Gilbert, N. and Frohlich, D. (Eds.), Computers and Conversation, Academic Press.

O'Conaill, B and Whittaker, S. and Wilbur, S., 1993, Conversations over Video Conferences: An Evaluation of the Spoken Aspects of Video-Mediated Communication, Human-Computer Interaction, 8, 389-428.

Ogden, W.C. and Boyle, J.M., 1982, Evaluating human-computer dialog styles: Command versus form fill-in for report modification, Proc. Human Factors Society, Twenty-Seventh Annual Meeting, Santa Monica, CA, pp. 542-545.

Olson, G.M., Olson, J.S., Carter, M. and Storrøsten, M., 1992, Small group design meetings: An analysis of collaboration, Human Computer Interaction, 7, 347-374.

Olson, J.S., Olson, G.M. and Meader, D.K., 1995, What mix of video and audio is useful for remote real-time work, In Proceedings of

Computer and Human Interaction (CHI) '95 (Denver, CO, May 1995) ACM Press, pp. 362-368.

Osgood, C.E., 1962, Studies in the general affective meaning systems, American Psychologist, 17, 10-28.

Paget, R., 1930, Human Speech, Harcourt: New York.

Payne, S.J. and Green, T.R.G., 1986, Task-action Grammars: a model of the mental rpresentation of task languages. Human Computer Interaction, 2, 93-133.

Payne, S.J. and Green, T.R.G., 1989, The Structure of Command Languages: an experiment on task-action grammar, International Journal of Man-machine Studies, 30, 213-234.

Pisoni, D.B., 1982, Perception of speech: the human listener as cognitive interface, Speech Technology, April, pp. 10-23.

Polson, P.G., 1988, The consequences of consistent and inconsistent user interfaces, In Guindon, R. (Ed.) Cognitive Science and its Applications for Human-Computer Interaction, Lawrence Erlbaum: Hillsdale, NJ, pp. 59-108.

Power, R., 1979, The Organisation of Purposeful Dialogues, Linguistics, 17, 107-152.

Pugh, 1995, Hierarchic Popup Menu Version 3.3.1, Hypercard Shareware.

Raeithel, A. and Velichovsky, B., 1996, Joint Attention and Co-construction of Tasks: New Ways to Foster User-Designer Collaboration, In Nardi, B.A. (Ed.), Context and Consciousness: Activity Theory and Human-Computer Interaction, Cambridge, MA, MIT Press.

Rapp, B. and Caramazza, A., 1997, The modality-specific organization of grammatical categories: Evidence from impaired spoken and written sentence production. Brain and Language, 56, 248-286.

Reddy, M., 1979, The conduit metaphor: A case of frame conflict in our language about language, In Ortoney, A. (Ed.) Metaphor and Thought, Cambridge: Cambridge University Press.

Reid, A.A.L., 1970, Electronic person-person communications, Communications Studies Group paper no. P/70224/RD.

Reisner, P., 1981, Formal grammar and design of an interactive system. IEEE transactions on Software Engineering, 5, 229-240.

Reisner, P., 1993, APT: A description of user interface consistency, International Journal of Man-Machine Studies, 39, 215-236.

Richardson, J.T.E. (Ed.), 1996, Handbook of qualitative research methods for psychology and the social sciences, British Psychology Society: Leicester, UK.

Roberts, L.A. and Sikora, C.A., 1997, Optimizing feedback signals for multimedia devices: earcons vs. auditory icons vs. speech. Proceedings of HCI International, Tampere, Finland, 224-226.

Robinson, H., 1990, Towards a Sociology of Human-Computer Interaction: A Software Engineer's Perspective, in Luff, P., Gilbert, N. and Frohlich, D. (Eds.), Computers and Conversation, Academic Press.

Root, R.W., 1988, Design of a multi-media system for social browsing, Proceedings of the Conference on Computer-Supported Cooperative Work, pp. 25-38, Portland, OR.

Rutter, D.R., 1984, Looking and Seeing: The role of visual communication in social interaction, Wiley: Chichester.

Rutter, D.R., 1987, Communicating by Telephone, Oxford: Pergamon Press.

Rutter, D.R. and Stephenson, G.M., 1977, The role of visual communication in synchronising communication, European Journal of Social Psychology, 7, 29-37.

Rutter, D.R., Stephenson, G.M. and Dewey, M.E., 1981, Visual communication and the content and style of conversation, British Journal of Social Psychology, 20, 41-52.

Ryder, M., 1999, What is Activity Theory?, http://www.cudenver.edu/~mryder/itc_data/act_dff.html

Sacks, H., Schegloff, E. and Jefferson, G., 1974, A simplest systematics for the organisation of turn-taking for conversation,

Saussure, F. de, 1990, Course in General Linguistics, Edited by Bally, C., Sechehaye, A. and Riedlinger, A., Translated by Harris, R., Duckworth: UK.

Schafer, R.W., 1995, Scientific Bases of Human-Machine Communication by Voice, Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, No. 22, pp. 9914-9920.

Schegloff, E.A., 1972, Notes on a conversational practice: Formulating place, In Sudnow, D., (Ed.) Studies in social interaction, Free Press: New York, pp. 75-119.

Schegloff, E.A., 1980, Preliminaries to Preliminaries, In Zimmerman, D. and West, C. (Eds.), Language and Social Interation, Special issue of Sociological Inquiry, 50, 3, 104-52.

Schegloff, E.A., Jefferson, G, and Sacks, H., 1977, The preference for self-correction in the organisation of repair in conversation, Language, 53, 361-82.

Schegloff, E.A. and Sacks, H., 1973, Opening up Closings, Semiotica, 7 (3/4), pp. 289-327.

Schiffrin, D., 1994, Approaches to Discourse, Blackwell: Oxford, UK.

Scott D., and Purves, I.N., 1996, Triadic relationships between doctor, computer and patient, Interacting with Computers, 8, 4, 247-363.

Searle, J., 1969, Speech Acts, Cambridge University Press: Cambridge, UK.

Searle, J., 1979, A Taxonomy of Illocutionary Acts, In Searle, J., Expression and Meaning,Cambridge University Press: Cambridge, UK.

Sellen, A.J., (1992), Speech patterns in video-mediated conversations, In Bauersfeld, P., Bennett, J. and Lynch, G. (Eds.), Proceedings of Computer and Human Interaction (CHI) '92, ACM Press: New York, pp. 49-59

Selley, C.J., Baker, S. and McKay, R., 1997, SMART — Improving Customer Service, BT Technology Journal, 15, 1, January.

Shackel, B., 1997, Human-Computer Interaction — Whence and Whither, J. American Society for Information Science, 48, 11, 970-986.

Shneiderman, B., 1991, A taxonomy and rule base for the selection of interaction styles. In B. Shackel and S.J. Richardson (Eds.) Human Factors for Informatics Usability, Cambridge University Press, pp. 325-342. Reprinted in Baecker, R.M., Grudin, J., Buxton, W.A.S. and Greenberg, S. (Eds.), Readings in Human Computer Interaction: Toward the year 2000.

Shneiderman, B., 1998, Designing the User Interface: Strategies for Effective Human-Computer Interaction, Third Edition, Addison Wesley: .

Short, J.A., 1974, Effect of medium of communication on experimental negotiation, Human Relations, 27, 225-34.

Short, J., Williams, E. and Christie, B., 1976, The Social Psychology of Telecommunications, Wiley: London.

Simpson, C.A. and Navarro, T.N., 1984, Intelligibility of computer generated speech as a function of multiple factors, In Proceedings of the National Aerospace and Electronics Conference (84CHI984-7 NAECON), IEEE: New York, pp. 932-40.

Simpson, C.A., McCauley, M.E., Roland, E.F., Ruth, J.C., Williges, B.H., 1987, System Design for Speech Recognition and Generation, In Baecker, R.M. and Buxton, W.A.S. (Eds.), Readings in HCI, pp. 400-413.

Sinclair, J.M. and Coulthard, R.M., 1975, Towards an analysis of Discourse: The language used by teachers and pupils, Oxford University Press: Oxford, UK.

Smith, S.L., and Mosier, J.N., 1986, Guidelines for Designing User Interface Software, Report ESD-TR-86-278, Electronic Systems Division, MITRE Corporation, Bedford: MA.

Smith D.E. and Whalen J., 1995, Texts in Action, Unpublished MS.

Sperber, D. and Wilson, D., 1986, Relevance: Communication and Cognition, Blackwell: Oxford.

Stephenson, G.M., Ayling, K. and Rutter, D.R., 1976, The role of visual communication in social exchange, British Journal of Social and Clinical Psychology, 15, 113-120.

Stern, K.R., 1984, An evaluation of written, graphics and voice messages in proceduralised instructions, In Proceedings of the Human Factors Society 28th Annual Meeting, Human Factors Society: Santa Monica, CA, Vol. 1, pp. 314-318,

Suchman, L.A., 1987, Plans and Situated Actions, Cambridge University Press: Cambridge, UK.

Tang and Isaacs, 1993, Why do users like video: Studies of multimedia-supported collaboration, Computer Supported Collaborative Work, 1, 163-196.

Tognazzini, B., 1992, Tog on interface, Apple Computer Inc.

Urkin, J., Warshawsky, S.S., Pliskin, J.S., Cohen, N., Sharon, A., Binstok, M., and Margolis, C.Z., 1993, How does a computerised medical record (CMR) affect physicians' work style? A video recorded study, In Proceedings of the American Medical Informatics Association 1993 Spring Congress 89.

US Department of Defence, 1981, Human engineering design criteria for military systems, equipment and facilities (MIL-STD-1472C), Author: Washington, DC.

Veinott, E.S., Olson, J. and Olson, G.M. and Fu, X., 1999, Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other, Proceedings of Computer and Human Interaction (CHI) '99, Pittsburgh, PA, ACM Press.

Vygotsky, L.S., 1962, *Thought and Language*, Edited and translated by Hanfmann, E. and Vakar, G., Cambridge, MA.: MIT Press.

Vygotsky, L.S., 1978, Mind and Society, Cambridge, MA: Harvard University Press.

Waterworth, J.A. and Holmes, W.J., 1986, Understanding machine speech, Current Psychological Research and Reviews, 5, 228-245.

Watson, A. and Sasse, M.A., 1996, Evaluating audio and video quality in low-cost multimedia conferencing systems, Interacting with computers, 8, 3, 255-275.

Waxler, N.E., and Mishler, E.G., 1966, Scoring and reliability problems in Interaction Process Analysis: A Methodological Note. Sociometry, 29, 28-40.

Westall, F., 1995, Speech Technology: It's good to talk, British Telecommunications Engineering, Vol. 14, Oct 1995.

Whalen, J., 1995a, A technology of order production: Computer-aided dispatch in 911 communications, Pp. 187-230 in Paul ten Have and George Psathas (Eds.), Situated Order: Studies in the Social Organisation of Talk and Embodied Actions, University Press of America: Washington, DC.

Whalen, J., 1995b, Expert systems versus systems for experts: Computer-aided dispatch as a support system in real-world environments, In Peter J. Thomas (Ed.), The social and interactional dimensions of human-computer interfaces, Cambridge University Press: Cambridge, UK.

Whalen J., Zimmerman D.H. and Whalen M.R., 1988, When words fail: A single case analysis, Social Problems, 35, 4, 335-361.

Whiteside, J., Bennett, J. and Holtzblatt, K., 1988, Usability Engineering: Our experience and evolution, In Helander, M. (Ed.), Handbook of Human-Computer Interaction, pp. 791-817, Elsevier Science: North-Holland.

Whittaker, S. and O'Conaill, B., 1997, The role of vision in face-to-face and mediated communication, In Finn, K.E., Sellen, A.J. and Wilbur, S.J. (Eds.), Video-Mediated Communication, Lawrence Erlbaum: Mahwah, NJ.

Wickens, C.D., Sandry, D.L. and Vidulich, M., 1983, Compatibility and resource competition between modalities of input, central processing and output, Human Factors, 25, 227-240.

Wiedenbeck, S. and Davis, S., 1997, The Influence of Interaction Style and Experience on User Perceptions of Software Packages, International Journal of Human-Computer Studies, 46, 5, 563-588.

Williams, E., 1975, Coalition formation over telecommunications media, European Journal of Social Psychology, 5, 503-7.

Williams, E., 1977, Experimental Comparisons of face-to-face and mediated communication. Psychological Bulletin, 84, 963-976.

Williams, D.H. and Simpson, C.A., 1976, A systematic approach to advanced cockpit warning systems for air transport operations: Line pilot preferences, In Proceedings of the Aircraft Safety and Operating Problems Conference (NASA SP-416), Norfolk, VA: NASA-Langley Research Center, pp. 617-644.

Winograd, T. and Flores, F., 1987, Understanding Computers and Cognition: A new foundtion for design, Addison-Wesley: Reading, MA.

Wright, P.C. and Monk, A., 1989, Evaluation for Design, In Sutcliffe, A. and Macaulay, L. (Eds.) People and Computers V, Cambridge University Press: Cambridge, UK, pp. 345-358.

Yngve, V.H., 1970, On getting a word in edgewise, In Papers from the sixth Regional Meeting of the Chicago Linguistic Society, Chicago Linguistic Society: Chicago.

Young, I., 1974, Telecommunicated interviews: An exploratory study, Communications Studies Group paper no. E/74165/YN.

Zhang, Q., Wolf, C.G., Daijavad, S. and Touma, M., 1998, Talking to customers on the web: a comparison of three voice alternative. Proceedings of CSCW '98, pp. 109-117.

Appendix 1

—

Hard Copy Resources from The Comparison of Open and Scripted Interface Styles in Simulated Customer Service Interactions with Audio-Only Communication

Instructions to Travel Game Participants

Instructions

Please read these instructions carefully before you say that you are ready to start the game.

You have a map of the western states of the USA in front of you, marked with the names of states and the cities which have airports used by Atlantic Airports. You should use this map to help you plan a 22 day trip around these states using an "Atlantic Air Travel Pass".

The pass allows you unlimited travel along any air routes which Atlantic Airways has with the one restriction that each airport may only be visited once. If you wish to visit an airport more than once incurs a surcharge of $50. Your budget permits you 2 such surcharges.

Furthermore, you must spend at least three days in every state you visit, for sight-seeing, but you may travel within the state during that time.

The agent will be keeping a record of your itinerary, but you may want to write a note of it yourself.

Your objective in this game is to visit as many cities as you can within the 22 day duration of the pass.

The person who visits the most airports wins a £30 prize.

Game Rules

- You must plan to visit as many airports on Atlantic Air routes as you can in 22 days.
- You must spend 3 days in every state you visit, but may travel between airports within the state during that time.
- You can afford 2 surcharges for visiting the same airport more than once.
- You may restart or change your itinerary at any point in the game without cost; it is your final itinerary which counts.
- There is a real time limit of 10 minutes on your consultation with the travel agent.

Remember you are in a competition!

Starting Point

All competitors will begin their journey in Salt Lake City.

This is an Atlantic Airways hub and has links to more airports than any other airport in the western United States.

Western US Airports by State

The following airports and states are available.

AZ - Arizona PHX - Phoenix TUS - Tucson CA - California SAN - San Diego SFO - San Francisco CO - Colorado COS - Colorado Springs DEN - Denver ID - Idaho BOI - Boise PIH - Pocatello KS - Kansas ICT - Wichita MT - Montana BIL - Billings GTF - Great Falls NE - Nebraska OMA - Omaha NM - New Mexico ABQ - Albuquerque ELP - El Paso	NV - Nevada ELY - Ely LAS - Las Vegas OK - Oklahoma OKC - Oklahoma City TUL - Tulsa OR - Oregon PDX - Portland SD - South Dakota FSD - Sioux Falls RAP - Rapid City TX - Texas FTW - Dallas/Fort Worth HOU - Houston UT - Utah CDC - Cedar City SLC - Salt Lake City WA - Washington GEG - Spokane SEA - Seattle WY - Wyoming CPR - Casper JAC - Jackson Hole

Agent Handout

Rules Summary

- You must plan to visit as many airports on Pacific Air routes as you can in 22 days.
- You must spend 3 days in every state you visit, but may travel between airports within the state during that time.
- You can afford £100 in excess surcharges.
- You may "backtrack" at any point in the game without cost; it is your final itinerary which counts.
- There is a real time limit of 10 minutes on your consultation with the travel agent.

Starting Point

All competitors will begin their journey in Salt Lake City.

This is an Atlantic Airways hub and has links to more airports than any other airport in the western United States.

Western US states and their abbreviations

AZ - Arizona	NV - Nevada
CA - California	OK - Oklahoma
CO - Colorado	OR - Oregon
ID - Idaho	SD - South Dakota
KS - Kansas	TX - Texas
MT - Montana	UT - Utah
NE - Nebraska	WA - Washington
NM - New Mexico	WY - Wyoming

Student Travel Game: BT Simulation

Agent Questionnaire

Subject ID: ....
Experiment Date: ....

Agent Name: ....

Please circle one of the answers below to indicate your choice.

1.1 Did you find the last consultation natural?

a) very natural
b) fairly natural
c) neither natural nor awkward
d) fairly awkward
e) very awkward

1.2 How satisfied do you think the subject was with the final outcome of the consultation?

a) very satisfied
b) satisfied
c) neither satisfied nor dissatisfied
d) dissatisfied
e) very dissatisfied

1.3 How friendly did you feel that the consultation was?

a) very friendly
b) fairly friendly
c) neither friendly nor hostile
d) fairly hostile
e) very hostile

1.4 How happy did you feel with the way you dealt with the subject?

a) very happy
b) fairly happy
c) neither happy nor unhappy
d) fairly unhappy
e) very unhappy

Student Travel Game: BT Simulation

Subject Questionnaire

Subject ID:

Experiment Date:

Please answer the following questionnaire as frankly as possible.

All information is taken in strict confidence; personal details will only be used to contact you if you are the winner and as evidence that you have participated.

Personal Details

Name: ....

Date of Birth: ....

Matriculation Number: ....

Term-time address: ....

....

Term-time telephone: ....

E-mail address: ....

Signature: ....

Previous Experience

Please circle one of the answers below to indicate your choice.

1.1 How knowledgeable about the United States do you consider yourself?
   a) very knowledgeable
   b) quite knowledgeable
   c) moderately knowledgeable
   d) slightly knowledgeable
   e) not at all knowledgeable

1.2 If you have visited the United States, how long have you spent there?
   a) more than 5 weeks
   b) 3 to 5 weeks
   c) 1 to 2 weeks
   d) less than 1 week
   e) never or for stopover only

1.3 How often have you used the telephone to arrange travel plans, for instance to book flights or train tickets?
   a) very often (weekly)
   b) fairly often (monthly)
   c) occasionally (yearly)
   d) rarely (once or twice)
   e) never

1.4 How often have you used the telephone to book other services, such as concert tickets?
   a) very often (weekly)
   b) fairly often (monthly)
   c) occasionally (yearly)
   d) rarely (once or twice)
   e) never

The Experiment

Please circle one of the answers below to indicate your choice.

2.1 How easy was it to hear and understand the travel agent?

a) very easy
b) fairly easy
c) neither easy nor difficult
d) fairly difficult
e) very difficult

2.2 Did you find the conversation with the Travel Agent natural?

a) very natural
b) fairly natural
c) neither natural nor awkward
d) fairly awkward
e) very awkward

2.3 How easy was it to make changes in your itinerary?

a) very easy
b) fairly easy
c) neither easy nor difficult
d) fairly difficult
e) very difficult

2.4 How helpful did you find the travel agent?

a) very helpful
b) fairly helpful
c) neither helpful nor unhelpful
d) fairly unhelpful
e) very unhelpful

2.5 How satisfied were you with the final outcome of the consultation?

a) very satisfied
b) satisfied
c) neither satisfied nor dissatisfied
d) dissatisfied
e) very dissatisfied

2.6 Did you feel that you had lost touch with the Travel Agent at any time.

a) Yes

b) No

2.7 How friendly did you feel that the Travel Agent was?

a) very friendly

b) fairly friendly

c) neither friendly nor hostile

d) fairly hostile

e) very hostile

2.8 Please write out what you remember your final itinerary to be below.

City 1. ....

City 2. ....

City 3. ....

City 4. ....

City 5. ....

City 6. ....

City 7. ....

City 8. ....

City 9. ....

City 10. ....

City 11. ....

City 12. ....

City 13. ....

City 14. ....

City 15. ....

City 16. ....

Thankyou for your participation.

The winner will be contacted as soon as the results are available.

Appendix 2

—

Hard Copy Resources from The Comparison of Sales Prompting Styles in Simulated Customer Service Interactions with Audio Only Communication

Customer Information

I will answer any questions you have about the rules after I have trained the agent.

Instructions

One of your friends has joined a scheme which allows students to work abroad in the US over the summer holiday.

Since they payed a deposit early they were entered in a prize draw and have won a 3 week trip for 2 around the western United States. You have been invited to come along, on the condition that you organise the route you will take. All the flights have to be booked before you leave the UK, and this is the purpose of the "call" you are now making.

The Pacific Airlines Travel Pass you have been given permits unlimited travel between airports the airline uses for 22 days, with only two restrictions.

Firstly, you must spend at least 3 days in each state, but may travel to other airports within the state during that time.

Secondly, every time you visit the same airport after the first there will be a £50 surcharge. You and your friend have agreed to spend at most £100 on such travel arrangements.

Your journey will begin at Salt Lake City; a Pacific Airlines hub which has links to most other airports.

Your friend wants to visit as many cities as possible so that they can decide where they would most like to spend their summer. Therefore, your objective is to visit as many airports as possible.

If at any time, you reach a location that you find it difficult to advance from you may "backtrack" to an airport you have previously been to. This will erase your itinerary from that point onwards, including any surcharges.

The call you make may take no longer than 15 minutes – the agent will advise you when this time is up.

To demonstrate your friend's pleasure if you organise the route well, **the customer who arranges the itinerary visiting the most airports will win a £15 prize.**

Rules Summary

- You must plan to visit as many airports on Pacific Air routes as you can in 22 days.
- You must spend 3 days in every state you visit, but may travel between airports within the state during that time.
- You can afford £100 in excess surcharges.
- You may "backtrack" at any point in the game without cost; it is your final itinerary which counts.
- There is a real time limit of 15 minutes on your consultation with the travel agent.

Don't forget you are in a competition!

Good luck!

Starting Point

All competitors will begin their journey in Salt Lake City.

This is an Atlantic Airways hub and has links to more airports than any other airport in the western United States.

Western US Airports by State

The following airports and states are available.

AZ - Arizona PHX - Phoenix TUS - Tucson CA - California SAN - San Diego SFO - San Francisco CO - Colorado COS - Colorado Springs DEN - Denver ID - Idaho BOI - Boise PIH - Pocatello KS - Kansas ICT - Wichita MT - Montana BIL - Billings GTF - Great Falls NE - Nebraska OMA - Omaha NM - New Mexico ABQ - Albuquerque ELP - El Paso	NV - Nevada ELY - Ely LAS - Las Vegas OK - Oklahoma OKC - Oklahoma City TUL - Tulsa OR - Oregon PDX - Portland SD - South Dakota FSD - Sioux Falls RAP - Rapid City TX - Texas FTW - Dallas/Fort Worth HOU - Houston UT - Utah CDC - Cedar City SLC - Salt Lake City WA - Washington GEG - Spokane SEA - Seattle WY - Wyoming CPR - Casper JAC - Jackson Hole

Agent Handout

Rules Summary

- You must plan to visit as many airports on Pacific Air routes as you can in 22 days.
- You must spend 3 days in every state you visit, but may travel between airports within the state during that time.
- You can afford £100 in excess surcharges.
- You may "backtrack" at any point in the game without cost; it is your final itinerary which counts.
- There is a real time limit of 10 minutes on your consultation with the travel agent.

Starting Point

All competitors will begin their journey in Salt Lake City.

This is an Atlantic Airways hub and has links to more airports than any other airport in the western United States.

Western US states and their abbreviations

AZ - Arizona	NV - Nevada
CA - California	OK - Oklahoma
CO - Colorado	OR - Oregon
ID - Idaho	SD - South Dakota
KS - Kansas	TX - Texas
MT - Montana	UT - Utah
NE - Nebraska	WA - Washington
NM - New Mexico	WY - Wyoming

Prompts

Since the ticket the customer has purchased is for two people the surcharges indicate the amount to be paid for both travellers, not individually.

Prompt:	"sell travel pass"
Details:	Local travel pass Surcharge: £10 for 2 people This pass permits unlimited travel on local bus and commuter rail for the duration of your stay in the town. It is offered for travel around the last location the customer visited. This provides a perfect way to get around the town and its surrounding area. The pass has been arranged privately between Pacific Airlines and the local travel companies and must be bought at the same time that the itinerary is booked. If the customer accepts, mark "TP" in the relevant airport's notes section. This will only be offered if the customer spends a complete daylight period in a town, but may not be available on all cities or at all times. Hilite the "TP" button by the penultimate airport.
Prompt:	"sell flight upgrade"
Details:	Flight upgrade Surcharge: £15 for 2 people This upgrades the quality of the in-flight arrangements on the last flight the customer booked. The improved arrangements include more comfortable seats and a personal entertainment system which permits the customer to choose the video they watch. The upgrade must be bought at the same time that the itinerary is booked. This will only be offered is the customer is taking a flight of 3 hours or longer, but may not be available on all flights. Hilite the "FU" button by the current airport.

Student Travel Game: BT Simulation
Subject Questionnaire

Please answer the following questionnaire as frankly as possible.

The Experiment

Please circle one of the answers below to indicate your choice.

1.1 Please rate your consultation on the following scales
Note that you are rating the consultation as a whole, not just your partner.

a)	Unpleasant						Pleasant
	1	2	3	4	5	6	7
b)	Still						Lively
	1	2	3	4	5	6	7
c)	Tender						Tough
	1	2	3	4	5	6	7
d)	Bad						Good
	1	2	3	4	5	6	7
e)	Passive						Active
	1	2	3	4	5	6	7
f)	Weak						Strong
	1	2	3	4	5	6	7

1.2 How easy was it to hear your partner?

a) very easy
b) fairly easy
c) neither easy nor difficult
d) fairly difficult
e) very difficult

1.3 Please rate your consultation on the following scales

Note that you are rating the consultation as a whole, not just your partner.

a)	Impolite				Polite
	1	2	3	4	5
b)	Inefficient				Efficient
	1	2	3	4	5
c)	Unfriendly				Friendly
	1	2	3	4	5
d)	Unassuring				Assuring
	1	2	3	4	5
e)	Incompetent				Competent
	1	2	3	4	5
f)	Unsympathetic				Sympathetic
	1	2	3	4	5
g)	Unhelpful				Helpful
	1	2	3	4	5

1.4 Did you feel that you had lost touch with your partner at any time.

a) Yes b) No

1.5 How polite did you feel your partner was?

a) very polite
b) fairly polite
c) neither polite nor impolite
d) fairly impolite
e) very impolite

<u>1.6 How easy was it to make changes in the itinerary?</u>

a) very easy
b) fairly easy
c) neither easy nor difficult
d) fairly difficult
e) very difficult

<u>1.7 How helpful did you feel that your partner was?</u>

a) very helpful
b) fairly helpful
c) neither helpful nor unhelpful
d) fairly unhelpful
e) very unhelpful

<u>1.8 How friendly did you feel that your partner was?</u>

a) very friendly
b) fairly friendly
c) neither friendly nor hostile
d) fairly hostile
e) very hostile

<u>1.9 How satisfied were you with the final outcome of the consultation?</u>

a) very satisfied
b) satisfied
c) neither satisfied nor dissatisfied
d) dissatisfied
e) very dissatisfied

<u>1.10 How efficient did you feel your partner was?</u>

a) very efficient
b) fairly efficient
c) neither efficient nor inefficient
d) fairly inefficient
e) very inefficient

1.11 How assuring did you feel your partner was?

a) very assuring
b) fairly assuring
c) neither assuring nor unassuring
d) fairly unassuring
e) very unassuring

1.12 Did you find the conversation natural?

a) very natural
b) fairly natural
c) neither natural nor awkward
d) fairly awkward
e) very awkward

1.13 How competent did you feel your partner was?

a) very competent
b) fairly competent
c) neither competent nor incompetent
d) fairly incompetent
e) very incompetent

1.14 How sympathetic did you find that your partner was?

a) very sympathetic
b) fairly sympathetic
c) neither sympathetic nor unsympathetic
d) fairly unsympathetic
e) very unsympathetic

1.12 Please write out below what you remember the final itinerary to be.

City 1. ....

City 2. ....

City 3. ....

City 4. ....

City 5. ....

City 6. ....

City 7. ....

City 8. ....

City 9. ....

City 10. ....

City 11. ....

City 12. ....

City 13. ....

City 14. ....

City 15. ....

City 16. ....

City 17. ....

City 18. ....

Previous Experience

Please circle one of the answers below to indicate your choice.

1.1 How knowledgeable about the US do you consider yourself?

a) very knowledgeable
b) quite knowledgeable
c) moderately knowledgeable
d) slightly knowledgeable
e) not at all knowledgeable

1.2 If you have visited the United States, how long have you spent there?

a) more than 5 weeks
b) 3 to 5 weeks
c) 1 to 2 weeks
d) less than 1 week
e) never or for stopover only

1.3 How often have you used the telephone to arrange travel plans, for instance to book flights or train tickets?

a) very often (weekly)
b) fairly often (monthly)
c) occasionally (yearly)
d) rarely (once or twice)
e) never

1.4 How often have you used the telephone to book other services, such as concert tickets?

a) very often (weekly)
b) fairly often (monthly)
c) occasionally (yearly)
d) rarely (once or twice)
e) never

1.5 Have you ever been employed as a telephone based customer service representative; for instance in tele-sales, as a telephone receptionist or in telephone based product support.

a) Yes b) No

If **YES**, then for how long? ....