SRI International

Technical Note 539 • February 1994

Toward Multimodal Support of Interpreted Telephone Dialogues

Prepared by:

Sharon L. Oviatt
Sr. Cognitive Scientist
Artificial Intelligence Center
SRI International

This research was supported by a contract from ATR International to SRI International.

Toward Multimodal Support of Interpreted Telephone Dialogues*

Sharon L. Oviatt
Artificial Intelligence Center
SRI International

1 Introduction

Real-time interpretation of telephone dialogues presents a difficult array of long-term empirical and computational research problems. Certainly, we are a long way from fully understanding the unique discourse and performance characteristics that will require accommodation during such dialogues, much less are we prepared to automatically process and interpret them for foreign speakers actively engaged in real tasks. On the other hand, research has intensified toward the development of both spoken language systems and translation of text – two prerequisites for developing the more sophisticated systems needed to handle interpretation of telephone dialogues. Furthermore, the specific long-term goal at organizations like ATR's Interpreting Telephony Research Laboratories in Japan has been real-time interpretation of spoken dialogues, and an interest in similar research pursuits is currently being generated by the Verbmobil research program in Germany (Kay, Gawron & Norvig, forthcoming).

As a parallel development, there is emerging commercial interest among telecommunications companies in providing worldwide telephone interpretation services, in this case with the aid of skilled human interpreters. For example, AT&T's "Language Line" in the United States has recently begun advertising rapid access 24 hours a day to professional telephone interpreters representing more than 140 languages – ranging from common languages like French to the more obscure Hausa. Similar services, although less ambitious with respect to the scope of languages represented, are offered by companies like KDD in Japan. Irrespective of whether our goal is the eventual automation of interpreted telephone calls or simply the support of human interpreters during such calls, our ultimate success would be enhanced by a clear understanding of how human telephone interpretation actually is conducted.

From a more theoretical perspective, we currently have very limited knowledge about the form and dynamics of spoken language, and about the important underlying factors that cause it to change during qualitatively different types of interaction. Such information is basic to our ability to model spoken language aptly for different computational purposes. For example, in

*This research was supported by a contract from ATR International to SRI International.

cases in which the basic modality of communication is spoken, we still need to identify the gross structural factors capable of influencing the communication in major ways, such as: 1) channel characteristics – e.g., whether spoken language is produced face-to-face, by telephone, or by computer; whether the communication is transmitted via a single modality or multimodally, and 2) the communication partner – e.g., whether none (i.e.- as in noninteractive monologue), one, or multiple partners are present; whether the partner is a computer system or fellow human; whether the partner speaks the same language or not. As we are just beginning to learn, distinctions at this relatively gross level can have substantial consequences for the form of the dialogue, as well as for task performance (Oviatt & Cohen, 1991a; Oviatt & Cohen, 1991b, Oviatt & Cohen, 1992). From an empirical standpoint, we can leverage considerable predictive information by identifying the most important among these gross structural factors that drive dialogue, and by examining how and why dialogues vary as a function of them. Given our lack of systematic knowledge about dialogue, it is at this level that we are most likely to account for the largest proportion of variability evident in human dialogue, which ultimately will reap research findings with the greatest predictive strength.

In terms of both discourse modeling and human performance, interpreted telephone dialogues present an example of spoken language during communications that are complex in the sense of being multiparty, multilingual, and mediated. In addition, of course, bandwidth limitations imposed by the telephone further influence these dialogues. As a result, this type of spoken interaction is quite demanding for the participants involved. The resulting dialogues also are structured radically differently than written text, which historically has provided our models of discourse structure. Even within the realm of speech, and as this chapter shall outline in some detail, three-person interpreted telephone dialogues differ substantially from the more common two-person noninterpreted ones (Oviatt et al., 1990; Oviatt & Cohen, 1991a; Oviatt & Cohen, 1992). As these data on interpreted telephone dialogues will illustrate, dialogue in the real world is variable, multidimensional, and complex, and our theoretical models of discourse must reflect this if they are to have any predictive strength.

With respect to the theoretical modeling of discourse, the idea that a single generic form exists is overly simplistic, and will prove insufficient in the long-range. If there are qualitatively distinct types of dialogue associated with different gross structural factors, as suggested by recent data, then a more comprehensive theoretical framework will be needed to encapsulate this diversity and to represent the most influential features in the communication setting. For the purpose of the present research, dialogue is viewed as it relates to a theoretical framework incorporating three interdependent levels of information. At the first level, distinctions are made about the individual or combined gross structural factors that influence dialogue, including the kinds of channel and partner characteristics discussed above. At the second level, task analytical information is incorporated into the dialogue model. This is followed, at the third level, by a more granular treatment of the interactants' specific goals, intentions, and plans within the given task. It is this third local level of analysis that has received the most research attention to date (see chapters in Cohen, Morgan & Pollack, 1990 for a sample of this work). In comparison with most current theories, the inclusion of these three nested tiers of information fulfills what some researchers have argued is a need for relatively broader and better integrated discourse models (Sparck Jones, 1991). Such a theoretical model, represented in Figure 1, bears potential relevance to a wide variety of empirical and computational research projects. As a long-term goal, this type of theoretical framework could be developed and used to make predictions about newly encountered dialogues that require modeling. Of course, the development of this kind of theory requires a firm empirical foundation, based on bottom-up research[1] defining the major factors at each level, how the three levels of information operate in an integrated manner, and what the net influence is on discourse.

INSERT FIGURE 1 ABOUT HERE

This paper compares three-person interpreted telephone dialogues with the more common two-person noninterpreted ones. To date, very little research has considered the topic of human dialogues during telephone interpretation, with the exception of recent research by Iida and colleagues (1987a, 1987b) and by Oviatt and colleagues (1988, 1992). Consequently, little is known of the language and behavior that speakers naturally engage in to support this demanding form of communication. The exploratory research presented in this chapter is based both on interviews with professional telephone interpreters, and on a detailed experimental study of Japanese–English telephone interpretation during service-oriented exchanges.

This chapter begins by summarizing research on the unique discourse and performance properties of interpreted telephone dialogues, with special emphasis placed on their structural and referential features, length and miscommunication patterns, confirmatory language, and linguistic indirection. A brief description also is presented of the pivotal role played by professional telephone interpreters as they manage the turn shifting, content, and sequencing of information passed among speakers. In the latter part of the chapter, several of the more problematic performance features of interpreted telephone dialogues are presented and analyzed. In particular, the experimental results indicated strains associated with the length, error-proneness, confirmatory requirements, and third-party waiting time during this type of communication. As an alternative to voice-only telephone interpretation, a multimodal system design is proposed that potentially could alleviate many of the identified difficulties. This multimodal system either could be used to support human interpreters, or eventually could be incorporated in the design of future automatic systems.

2 Overview of Experimental Study

An empirical study was designed to examine the predominant dialogue and performance characteristics of three-person interpreted telephone speech during service-oriented dialogues, in comparison with those of two-person noninterpreted dialogues.

[1] I quite agree with Webber (1991) that local problems, or "life at the bottom," have been a productive focus for discourse research. At the same time, a more global and expansive theoretical framework, including research findings relevant to the upper levels, contributes better perspective.

2.1 Method

Interpreted and noninterpreted dialogues were compared based on the same pool of subjects and tasks, nsing a within-subjects design. The study involved 12 native English speakers, who each made one telephone call through an experienced telephone interpreter to a Japanese confederate who did not speak English, and a second call to a Japanese confederate fluent in English. In total, 24 dialogues were collected, each one containing two successfully completed service tasks, or 48 tasks total.

For the interpreted calls, professional Japanese–English interpreters from AT&T's Language Line were patched in as part of their normal work routine. These interpreters were selected for their high ACTFL ratings and maximum experience conducting telephone interpretation. They represented a unique population with experience specifically in telephone interpretation. Their participation was designed to enhance both the naturalism and external validity of the study's results, since only interpreters who habitually operate in this modality can be expected to have established a stable interpretation style with respect to performance and discourse patterns.

During all calls, a native English-speaking researcher contacted a native Japanese registrar to complete a conference registration and a travel task. Registrars had familiarized themselves with reference materials needed for these tasks, and had practiced fielding calls in advance of the study so that they were able to act as skilled agents. Both types of task involved service-oriented exchanges, which included collecting information, solving simple problems, and making arrangements to obtain forms, all of which entailed the frequent exchange of proper names and numbers. During each task, both the researcher and the registrar wrote down information needed to follow through in actually accomplishing it. This information was collected both to make the tasks more realistic, and to provide an objective index of task performance.

The interpreted calls were conducted as conference calls among the three parties, so the English researcher and Japanese registrar each could hear incomprehensible speech in the other's language. During both the interpreted and noninterpreted calls, speakers were separated and visually inaccessible to one another. Interviews with Language Line interpreters had indicated that most of the Japanese–English calls they handle are service-oriented business ones, with three-person conference calls in which the participants are physically separated being the most common structural arrangement. That is, these features of the present study were designed to reflect high-frequency interpreted calls, which enhances the external validity of the results, as well as generating data in support of emerging commercial needs.

All calls were tape-recorded and transcribed, with the Japanese segments translated into English. Special attention was paid to accurately transcribing and translating all spoken language phenomena and their precise sequencing, including interjected backchannel confirmations, false starts, self-corrections, two- and three-person simultaneous speech, and so forth, rather than "cleaning it up" in any way. The content of the dialogues was coded and analyzed for linguistic and performance features of interest. Methodological details, including coding categories, reliabilities, statistical significance levels, and examples of task materials are available elsewhere (Oviatt & Cohen, 1992). All comparisons between the interpreted and noninterpreted calls reported in the results section of this chapter were statistically significant. The appendix provides brief samples from two subjects' interpreted telephone calls and one subject's noninterpreted call during the conference registration task.

2.2 Results and Discussion

The interpreted calls in this research were organized into a series of extended subdialogues between the interpreter and each of the two primary speakers in their native language. The tasks averaged seven to eight subdialogues apiece, half English and half Japanese. The average length of each subdialogue was 39 seconds, with no difference between the Japanese and English subdialogues. However, the English speakers who were interviewed often reported that they believed the Japanese turns took longer. That is, there appears to have been subjective distortion of the length of subdialogue turns when a person was listening to incomprehensible language. This may be a related phenomenon to telephone speakers' distorted perception that silent intervals are exceedingly long (Butterworth et al, 1977). Nevertheless, speakers reportedly liked hearing the foreign conversation in progress, perhaps partly because a sense of progress was communicated by prosodic cues.

With respect to basic performance issues, the interpreted calls averaged about 2.5 times longer than noninterpreted calls made by the same subjects accomplishing the same pool of tasks. The frequency of all speakers' task-relevant miscommunications, including both those that remained uncorrected and those that were corrected by another speaker, was judged from the transcripts and performance sheets. This total rate of miscommunications averaged 1.93 per task set for the interpreted calls, in comparison with 0.45 for the noninterpreted ones – a 4-fold difference. Of these, the average rate of errors that remained uncorrected was 0.93 for the interpreted calls, compared to 0.13 for noninterpreted calls – a 7-fold difference. The elevated miscommunication rate during interpretation included errors ranging from the transmission of spelled letters, digits, and proper names to terminology errors and occasional misunderstanding of task requests. Most errors were minor (e.g., "Harry" mispelled as "Hary") in the sense that they did not lead to consequential performance problems. Among the critical uncorrected errors, 71% were attributable to mistransmission of one or more spelled letters within a proper name, and 86% were due to the mistransmission of either spelled letters or digit strings.[2]

During interpreted calls, subjects also engaged in a higher rate of confirmation language geared toward double- and triple-checking the accuracy of information obtained from the interpreter. In fact, an examination of subjects' requests for confirmation and confirmations of information revealed that their rate of confirmation language per total words spoken increased from 23.5% during noninterpreted calls to 31.5% during interpreted ones. During the interpreted calls, then, nearly one-third of their language was exclusively concerned with the verification of information – a remarkably high rate, even for task-oriented dialogues. Speakers clearly were substantially more concerned about being understood during interpreted calls, which gener-

[2] Substitution of spelled letters was one common source of errors, with N/M being the most prevalent for Japanese–English speakers in this study. A second common source of spelling errors involved either erroneously doubling a letter (e.g., "Janett" for "Janet") or not doubling one (e.g., "Hary" for "Harry"). Since the overall rate of double- and triple-confirming information was extremely high in these interpreted dialogues, this latter type of error may have represented listeners' confusion over whether they were hearing a new letter or a confirmation of the original.

ated a more conservative communication style. However, this heavy reliance on confirmation exchanges still was insufficient to control the markedly higher rate of miscommunications that occurred during interpreted calls.

When first attempting to satisfy their task goals during interpreted calls, subjects used more indirect linguistic forms to request information or actions. Rather than using imperatives or direct "Wh" questions, they used indirect requests 92% of the time during interpreted calls, whereas these same subjects completing the same set of tasks reduced their rate of indirect requests to 60% during noninterpreted calls. Subjects may have used a direct style more often during noninterpreted calls simply because they expected the registrar to know the information requested, whereas they did not have such expectations of the interpreter. During interpreted calls, subjects instead transmitted their goals and needs to the interpreter, since they expected the interpreter to collaborate with the registrar in helping them to solve their problem.

The basic interpretive approach adopted by professional telephone interpreters in this study involved assuming the role of an independent agent who actively managed the information needed to complete a subject's task. That is, analyses indicated that interpreters did not conform to a strictly literal style, since they: 1) provided speakers with extra task-critical information that had not been requested (e.g., dialing information for foreign country codes), 2) directly answered questions when they already knew the answers, thereby shortcutting the usual interpretive process of passing the question to the other speaker (e.g., spelling the name of a Japanese company), 3) selectively omitted passing information that wasn't central to task solution (e.g., not transmitting the student fee when only the general fee had been asked), 4) offered to provide task-relevant information or actions that had not been raised by the speaker (e.g., offering to find out a hotel cost), and 5) prompted a speaker for task-relevant information that had not been raised (e.g., asking whether a quoted fare is one-way or round-trip). All interpreted calls contained examples of these initiating behaviors, which may have occurred partly in response to realistic requirements for successful and efficient task completion, or to limitations inherent in the telephone modality. At any rate, interpreters clearly played an active role in directing the content of these service-oriented calls, and in organizing the flow and sequencing of task information. However, they rarely altered the content of task-critical new information presented by either of the main speakers (e.g., name, FAX number). This information generally was passed immediately and literally to the other main speaker.

Telephone interpreters also took considerable initiative in the management of turn regulation. Of course, they effectively signaled turn shifts by initiating switches between languages, a forceful sign to both speakers of the opening and closing of the communication channel. In addition, interpreters provided clear linguistic marking of the start and end of most subdialogues, with the start of subdialogues indicated over 91% of the time, and both the start and end indicated over 57% of the time. Consequently, the data revealed that three-person interpreted calls were no more disorganized than two-person noninterpreted ones with respect to turn shifts among subdialogue partners. For example, two-person simultaneous speech was no more frequent during interpreted than noninterpreted calls, and third-party interjections and three-person simultaneous speech during an ongoing subdialogue were extremely rare. In short, turn-shifting between the English and Japanese subdialogues was remarkably clean during interpreted calls, and it appears that interpreters were quite skillful at managing this regulatory function.

In connection with interpreters' adoption of an active managerial role, they also explicitly referred to themselves an an independent agent through the use of first-person pronouns and noun phrases, rather than adopting a transparent interpretive style in which only the main speakers are represented. In addition, interpreters used third-person pronouns and noun phrases to refer to the waiting third person. In fact, all interpreters habitually adopted this style of explicitly referring to each of the three parties in the interpreted conversation, independent of whether the two main speakers did so. Given the lack of visual access in the telephone modality, this explicit referential style may have helped to prevent confusion about which of the three speakers was being represented.

3 Performance Problems Unique to Interpreted Telephone Dialogues

Participants engaged in interpreted telephone conversations reported advance concern about whether they would be understood, and whether their conversation could succeed. Afterwards, they also reported that completing the call had required patience and persistence. These subjective impressions are not surprising, given the length, structure, and uncertainty of interpreted calls from the main speakers' point of view. That is, these conversations required them to remain engaged for a much lengthier time than a typical two-person dialogue involving a simple service exchange. In addition, half of this time was spent waiting without feedback while the interpreter spoke to the other party. To compound this problem, participants' waiting time was distorted to seem lengthier than it actually was. Furthermore, no meaningful backchannel feedback from the other main speaker was forthcoming throughout the entire exchange, leaving speakers without this usual stream of reassurance, and totally dependent on the intermediating interpreter. Finally, while talking to the interpreter, speakers' advance concern about the success of the conversation was heightened further by experiencing a higher rate of errors that required correction. Under these circumstances, speakers evidently felt more vulnerable and more concerned about the fragility of the conversation's success. This may have led to their attempts to compensate by requesting and providing more confirmations.

One general performance problem for interpreted telephone dialogues clearly is the miscommunication rate. As described earlier, this rate is 4-fold higher than noninterpreted calls for total miscommunications, and 7-fold higher for those critical errors that remain uncorrected. The unique structure of interpreted calls results in the highest differential being for errors that remain uncorrected, since all three participants are never able to simultaneously confirm information in a final way. That is, the structure of the exchange precludes the source speaker from ever directly witnessing that his or her contributed information has been passed and received correctly. Furthermore, the process of explicitly reconfirming information back with the original source is, at best, imperfect. In order to reduce these uncorrected errors, task-critical information could be made confirmable automatically and simultaneously among all three speakers, rather than just the two engaged in the current subdialogue. This would permit the initiator to check for and correct errors in information that he or she had supplied. Such a confirmation capability might best be achieved through the visual channel, with supplementary visual feedback available to all three parties.

Since 86% of all task-critical errors in these tasks involved either spelled letters in proper names or else digit strings, these classes of information should be the primary focus of any coordinated confirmation system. In fact, it may be advantageous simply to send this type of information visually, with all three parties able to view it, and the interpreter providing supplementary commentary as needed. This arrangement could prevent the vast majority of errors, or about 80 to 90%, that occurred during service-oriented interpreted calls. Such a substantial reduction in errors would lead to a proportional decrease in the subsequent clarification subdialogues needed to correct them, at least in the 50% of cases in which such errors were noticed. This in turn would lead to a reduction in the overall length of calls.

A second general set of performance problems is associated with the lengthiness of interpreted telephone dialogues, including: 1) a 2.5-fold excess duration beyond that of comparable noninterpreted calls, 2) the need for speakers to spend half of this time waiting between subdialogues, and 3) speakers' subjective distortion of the length of this idle time.[3] Apart from speakers' reports that interpreted calls require considerable patience, let's take a moment to consider why lengthiness per se should be viewed as a performance "problem." First, interviews with interpreters and the main speakers confirmed that efficiency is a primary client expectation of a telephone interpretation service. To date, people tend to be seeking telephone interpretation in situations where efficiency is wanted (e.g., routine service-oriented transactions) or needed (e.g., medical or police emergencies). Of course, long-distance and per-minute interpreter charges also encourage people to feel that time matters. Furthermore, interpreters report that they feel internal pressure to complete each call quickly, and this is reinforced by the commercial telephone interpretation services that employ them.[4] In short, a mutual desire for efficiency appears to be a primary feature of this communication modality, and clients' expectations for efficiency may require that it receive priority in future systems that are developed.

Given this clash between speakers' expectations for brevity and the actual lengthiness of interpreted telephone conversations, both a reduction in the overall length of interpreted calls and in speakers' perceived waiting time would be advantageous, assuming that these changes could be accomplished without increasing the miscommunication rate. Perceived waiting time could be reduced by engaging the waiting person with meaningful feedback. If, as suggested earlier, the waiting person received visual feedback about task-critical information as they simultaneously heard the foreign subdialogue between the other two parties, then this would

[3] Although a simultaneous interpretation system for handling telephone dialogues theoretically could overcome this set of performance problems, it also would exacerbate the already high error rate in this type of call, precluding consistently high quality interpretation. At least for the moment, technology in support of consecutive human interpretation is believed to be the only tenable alternative, since it is at least *relatively* error-free. It is an empirical question whether fully automatic simultaneous systems can be developed for handling interpreted telephone dialogues in a manner that is accurate and successful.

[4] AT&T Language Line advertisements emphasize that an appropriate interpreter will be patched into an incoming call with an average delay of just 14.5 seconds. Training of telephone interpreters also encourages a focus on expediting the specific task at hand in a timely way.

provide them with specific progress information that could reduce their subjective sense of elapsed time. As mentioned earlier, it also would reduce the subsequent need for follow-up confirmation and clarification subdialogues, thereby reducing the overall length of the call.

A third performance problem for interpreted telephone dialogues involves the very high proportion of speakers' time that is spent seeking or supplying confirmation on central aspects of the task, which may occur in part because the structure of these dialogues prevents speakers from receiving reassuring backchannel feedback. By increasing and streamlining speakers' receipt of confirmation feedback, it should be possible to relax their sense of uncertainty, thereby reducing the amount of time they feel compelled to invest in confirmatory exchanges. If important proper names, their spelling, and digits could be transmitted and confirmed visually during the call then, according to calculations based on the original transcripts, the overall proportion of confirmation language out of all language transmitted could be reduced by 44%, or from 31.5% to 19% of the total dialogue. This latter percentage is more consistent with the proportion of confirmatory language typically found in noninterpreted task-oriented calls, which has ranged from approximately 18 to 23.5% in this and other studies (Oviatt & Cohen, 1991a & 1992). Based on computations from the present transcripts, this single system feature thereby could reduce the overall length of this type of service-oriented dialogue by approximately 12 to 13%. In addition, of course, speakers would be free of the anxiety caused by not receiving direct and timely feedback, and of the tedium involved in constantly having to check everything, so that they instead could focus their attention more productively on the task at hand.

4 Multimodal Support of Interpreted Telephone Dialogues

A multimodal system potentially offers solutions to this constellation of problems for voice-only telephone transmission. To begin, consider a multimodal telephone that simply transmits communication signals in support of a human interpreter, with no speech recognition, natural language, or interpretation capabilities. Further, consider only interpretation during the general class of service-oriented exchanges, including such transactions as arranging for car rentals, airline reservations, conference registration, and bank transfers. What critical features would be needed in a multimodal system to diminish the outlined performance problems for interpreted telephone dialogues?

One key feature is a coordinated confirmation system. Such an arrangement would need to be capable of confirming task-critical information simultaneously among all three participants, rather than just the two engaged in the current subdialogue. Of course, it would be easiest for such simultaneity to be achieved visually. Since spelled proper names and digits accounted for most of the task-critical errors in the present research, it would be strategic for the confirmation system to focus on this content. If the letters in spelled names were transmitted visually, this could avoid the typical cross-linguistic problems associated with foreign speakers not hearing native sound contrasts in one another's languages. The inability to distinguish certain foreign contrasts, such as Japanese speakers' difficulty hearing the English R and L, was one source of spelling errors during voice-only telephone transmission.

To provide coordinated visual confirmations, consider a multimodal telephone in which an electronic tablet and stylus are used to transmit handwritten and drawn marks.[5] At least for the class of service-oriented transactions, all three participants could receive simultaneous confirmations of certain task-critical information by viewing the completion of a "receipt" on their multimodal phone. As the conversation progressed through its series of subdialogues, information on this receipt could be filled out by each of the two main participants, with completed information automatically echoed to all parties. The content of the receipt would be tailored to the type of transaction involved, and could be supplied by the service agent. For example, in the case of conference registration, the receipt contents displayed in Figure 2 might be presented. The conversation could begin with a blank receipt presented to all three parties. Since a standard set of prompts could be developed for many service transactions, such as car rentals or bank transfers, the prompts for these receipts could be translated in advance and made available to the two main speakers in their native language. Then the receipt's contents could be filled in by the two main speakers at appropriate times during the conversation. In the case of the conference registration receipt presented in Figure 2, for example, the registrant could directly transmit his or her handwritten name, business affiliation, address, phone, credit card information and, of course, signature. The registrar subsequently could fill in the registration fee, inclusive items, and the confirmation number.

INSERT FIGURE 2 ABOUT HERE

These handwritten items of information often may not require translation by the interpreter during the conversation, although she may add explanatory comments. That is, the need for translation depends on the intended uses of the receipt's contents, and on the languages involved. For example, conference registration information collected simply as an official record, or for the purpose of addressing envelopes or name badges, may or may not require translation. Even in cases in which translation of certain items is needed, the possibility exists for the interpreter to postpone this work until after the call has been completed. Essentially, then, this system could free the interpreter from the time, effort, and risk of mediating these error-prone items. The interpreter's function instead could focus more on managing other aspects of the exchange. In addition, of course, each foreign speaker would have specific advance expectations for the kind of information that would be completed after each prompt (e.g., the person's name, a credit card company, etc.), since they would be viewing the prompts in their own native language. As a result, untranslated material following these prompts may not be disturbing. This may likewise be true when the two main speakers have different alphabetic codes, as with Japanese and English speakers, although translation is likely to be required more often in such cases for practical reasons (e.g., an address written in Kanji could not be left untranslated if it was needed to address an envelope that had to pass successfully through the U.S. postal system).

[5] The underlying hardware for this type of multimodal telephone is available already, although these phones have yet to be specifically developed to support any particular application such as telephone interpretation. For example, AT&T has developed the "Smartphone," and Wacom the "Viewphone," both of which transmit the usual telephone speech signals as well as handwritten or drawn markings. If such phones became available to support interpreted dialogues for service transactions, then both the telephone companies that provide interpretation and the companies selling commercial services would be more likely candidates than private individuals for acquiring and maintaining the necessary new hardware.

When foreign speakers do not share the same numerical and alphabetic codes, it can be particularly difficult for them to engage in voice-only interpretations that require the frequent exchange of spelled proper names and numbers. In the present study, the transmission of many names and spelled letters during Japanese–English interpretations tended to be laborious and error-prone.[6] Visual transmission of such information would be particularly valuable in these cases. In this study, the Japanese native speaker and interpreter spontaneously used Japanese to express the sounds of foreign English names, then spelled English letters within those names, provided other English words beginning with the same letter ("T as in Tokyo"), preceding letters in the English alphabet ("B as in A B"), and so forth. That is, they sometimes needed to present a time-consuming collection of sound and concept clues in order to clarify English words and their spellings, since Japanese does not use the modified Latin alphabet, and no perfect correspondence exists between sounds in the two languages. If it were not for the fact that most educated Japanese are familiar with the English alphabet, such an approach would not have been possible at all. In short, when alphabetic codes differ as they do for Japanese and English, a multimodal transmission system of the type proposed has considerable potential for improving the accuracy and efficiency of interpretation.

In addition to the receipt, it would be advantageous to include free space on the tablet for drawing maps, sketching discussed objects, or recording personal notes. Interviews with professional interpreters have indicated that drawing simple maps while directions are given in parallel would be particularly helpful during many conversations with foreign speakers. This is the case because foreigners' service transactions often are travel oriented, and locations in the host country tend to be unfamiliar. Interpreters also report that speakers frequently would benefit from a view of objects under discussion. Object diagrams could be of assistance in clarifying the meaning of unfamiliar or technical foreign concepts that are being interpreted. In addition, as they interpret, many professional interpreters report that they take notes as a mnemonic aid, so they can recall and transfer information accurately to the other speaker. Free space on the proposed multimodal phone could be used for personal notetaking, either by designating an area for this purpose, or by having the interpreter select "private" rather than "public" transmission.

The proposed multimodal system would transform the interpreted telephone conversations in at least three major ways. First, it would introduce mixed communication, with both speech and writing. Secondly, it would reduce the interpreter's load and control during the conversation.[7] With respect to the issue of control, the "receipt" itself now would guide more of the informational content and ordering of information discussed, and task-critical information covered by the receipt would be supplied directly by the main speakers without routine

[6] See the second interpreted dialogue sample in the appendix, involving subject #1 and interpreter #1, for an example of the time and effort typically required by two Japanese speakers to spell an English name. In spite of the effort expended during this spelling exercise, an error was detected on the performance sheets, where the registrar failed to record the entire second syllable of the registrant's surname, "der." This syllable also was the only one that the registrar did not confirm verbatim during the exchange.

[7] This shifting of some control from the interpreter to main speakers may be desirable from all parties' point of view, since the interpreter often is overloaded, and the main speakers dislike feeling dependent on the interpreter and uncertain about whether they are being understood.

mediation. Thirdly, the proposed system would introduce segments of simultaneity into an otherwise consecutive interpretation, in the sense that task-critical information appearing on the receipt could be transmitted directly, with no interpretation lag. This latter feature could assist greatly in speeding up interpreted conversations, but without the considerable loss of accuracy that would be associated with voice-only simultaneous interpretation.

To summarize anticipated improvements in the performance problems outlined earlier, the proposed multimodal system potentially could reduce miscommunications in these kinds of service exchanges by as much as 80 to 90%. This estimate is based on the finding that errors in spelled names and digits accounted for 86% of all task-critical errors in the present study. A reduction of this magnitude could be accomplished in part by presenting information in a visual form that would be much easier for foreign speakers to perceive and use. In particular, it would eliminate speakers' inability to distinguish difficult sound contrasts, including nonnative ones. Using the proposed system, errors could be avoided by permitting each main speaker to send and receive task-critical information directly, which essentially would eliminate the need for interpreter mediation of this error-prone material, as well as reducing the interpreter's load. Finally, in cases where errors did occur, the proposed system would enable easier error detection and correction, since the initiator of information could check for and correct any errors in information that he or she had supplied. Essentially, then, the proposed system has the potential to bring the error rate into line with that typically experienced during two-person noninterpreted calls covering the same content.

It also is estimated that the proposed system could reduce the high proportion of total language consumed by confirmatory exchanges by as much as 44%. This estimate, based on calculations from the present set of transcripts, derives solely from the expectation of reduced confirmations and requests for confirmation focused on spelling and digits. This change would result in a drop in the proportion of confirmatory language from its present 31.5% level down to approximately 19% of the overall dialogue. Once again, this level would be more in line with that typically found in task-oriented noninterpreted calls between two speakers. Of course, such a reduction in confirmatory language also would be expected to reduce the total length of the conversation and, in turn, the total task completion time. For example, given that the proposed system could reduce confirmation language by approximately 44%, then estimates from our current transcripts indicate that this single change could shorten the overall length of this type of dialogue by 12 to 13%.

Due to other economies associated with the proposed system, total task completion time actually would be expected to speed up considerably more than 12 to 13%. As mentioned earlier, the reduction in errors discussed above also would cause related clarification subdialogues to drop out, which in turn would shorten the dialogue. Further reductions in task completion time would be roughly proportional to the degree in which simultaneous transmission of handwritten information replaced auditory transmission, since the former would result in by-passing the consecutive interpretation process. In the present transcripts, about 25 to 30% of subjects' time was spent exchanging spelled names and digits that could be transmitted in a more direct visual manner using the system proposed. As a result, this particular alteration would most likely provide a further substantial reduction in completion time, as well as constituting a major avenue for reducing each speaker's waiting time. Since task completion time is driven by a particularly complex set of factors, it would be difficult to estimate precisely how much it would be reduced by their combined influence, beyond simply predicting a sizable overall reduction. Further research is needed on interpreted dialogues receiving the proposed system support, in order to confirm the performance improvements predicted above, to establish more accurate estimates of them, and to rule out the possibility that negative trade-offs may result in the process.

In contrast with a system designed to transmit information in support of human interpreters, the development of a partially or fully automatic interpretation system would be considerably more complex and long-term, requiring more work in areas like speech processing, natural language processing, machine translation, dialogue modeling, and human factors. Indeed, the basic design elements required to build an optimal automatic system are as yet very ill-defined. Given an automatic arrangement, however, the use of a coordinated confirmation system would be expected to reduce errors, shorten total completion time, and reduce the high proportion of confirmatory language. Another advantage of the proposed system is that some of the written receipt information could be transmitted without translation. That is, to the extent that direct transmission occurred, only partial automation would be required, thereby enhancing the likelihood of successful processing.

A multimodal pen/voice system would be preferable to speech-only dialogue interpretation in several respects. For example, service-oriented tasks require recognizing the names of individuals and commercial establishments, although recognition of proper names is a weak point for current speech systems. The option of writing task-critical names, simultaneously conveying their spelling, would be an advantageous use of the pen for this reason. Secondly, the replacement of a human with automatic speech recognition would not eliminate the problem of distinguishing difficult-to-hear or nonnative sounds with no clear designation in a foreign syllabary or phonetic system. Automatic speech recognizers and dialogue translation systems still would have difficulty processing such content, which generally could be handled more accurately and efficiently in written form. These issues simply highlight the fact that speech and handwriting recognition have the potential to be combined in ways that are complementary, and that provide compensation for one another's weaknesses.

An additional advantage of a multimodal pen/voice system would be the efficiency made possible through parallel performance. For example, an agent could speak instructions about how to get to a car rental pickup location, while simultaneously drawing a simple line map. Furthermore, by adding the multifunctional capabilities of a pen system to speech recognition, one not only could provide and confirm task-critical information in writing, but also create graphics, edit gesturally, and verify signatures dynamically. The latter of these would be particularly useful for on-line approval of financial transactions.[8] In short, in the context of automated telephone interpretation, there are reasons to believe that the multimodal system proposed in this chapter would yield strategic advantages over telephone speech alone.

[8] These multifunctional capabilities indicate why a pen-supplemented telephone system would be preferable to a keyboard one, in addition to the fact that pen/voice telephones potentially could be portable.

5 An Empirical Perspective on Dialogue Research

The research described in this chapter is representative of an experimental approach, which aims to: 1) identify undiscovered dialogue phenomena, 2) quantify the actual prevalence of various dialogue phenomena, 3) isolate and manipulate influential factors that drive dialogue patterns, 4) interpret dialogue patterns as they relate to other phenomena (e.g., task performance) in order to leverage explanatory and predictive power, and 5) enhance objective analysis through the adoption of methodological tools and scientific conventions. In order to permit ourselves the opportunity to discover the breadth of dialogue phenomena that exist, which is essential to the aims of empirical research, it is necessary to study actual human communication within its full range of natural contexts. That is, if empirical work is in the business of accounting for variability, then we must first establish what the full range of variability is. To do so permits us to begin distinguishing types of discourse that are qualitatively very different, so that we then can begin identifying the important underlying factors that drive discourse and performance to change in major ways. It also permits us better opportunities for uncovering new discourse phenomena, and for identifying the boundary conditions that operate on them. Considerably more information will be needed about the variable, multidimensional nature of discourse phenomena that exist, including those reported here for interpreted telephone dialogues, before models can be developed that are broad, well integrated, and have predictive strength.

Given the present research findings, which exemplify the real-world complexity of data, it is clear that there exist qualitatively different types of dialogue. These dialogue types are associated with different sets of gross structural factors, including the overriding channel and partner characteristics discussed in this chapter's introduction. As a result, the pursuit of a single generic dialogue model is an overly simplistic goal. From a pragmatic view, more divergent and accurate models will be needed to build communication systems serving an array of different applications. From a theoretical view, a broader framework will be needed that integrates the major factors influencing dialogue, in order to account for the complexity of real data sets. The relatively local theories of discourse that are currently available will be insufficient for these purposes. This chapter presents an example of an alternative theoretical framework, one that includes three fundamental levels of information within an interdependent hierarchy. A more comprehensive framework of the type proposed is needed, and should be specified further based on related empirical findings, so that we can begin generating useful predictions about newly encountered dialogues.

In comparing interpreted telephone dialogues with noninterpreted ones, both of which represent complex and naturally occurring communication modalities, this research clearly opted to focus on dialogue differences as a function of naturally co-occurring clusters of factors within the context of real communication modalities. That is, the present investigation does not represent factorial experimentation aimed at isolating individual factors. Experimental manipulation and analysis at the more global level found in this study can provide a clearer orientation during the early stages of exploring new topics, as well as enhancing the external validity and generalizability of research findings. Once again, since the present research agenda was an exploratory probe, an eclectic research strategy was selected that included both interviewing and experimentation. In short, given the lack of literature on interpreted dialogues, the present research was conducted in the spirit of scouting new territory. Therefore, the methodology was designed to sweep broadly across the landscape for signs of new modality landmarks and research routes.

6 Summary

This chapter outlined the unique discourse and performance properties of interpreted telephone dialogues during service-oriented exchanges, with special emphasis placed on their structural and referential features, length and miscommunication patterns, confirmatory language, and linguistic indirection. In addition, a description was presented of the pivotal role performed by professional telephone interpreters as they manage turn shifting, and the content and sequencing of information passed among speakers. The constellation of performance problems that characterize interpreted telephone dialogues was analyzed, including excessive length, third-party waiting time, error-proneness, blockage of backchannel confirmations between the main speakers, and elevation in the overall proportion of confirmatory language. As an alternative to voice-only telephone interpretation, a multimodal system was proposed that potentially could reduce or eliminate these outlined difficulties. The most salient features of this system include multimodal pen/voice transmission with multifunctional capabilities, and a coordinated confirmation system based on the completion of "receipts." The general strategic advantages of the proposed system were outlined, and estimates were provided of anticipated performance improvements for dialogues receiving this support. The long-term goals of the exploratory research described in this chapter include the description, theoretical account, and modeling of qualitatively different types of dialogue, the specification of preliminary target requirements for future automatic systems, and the optimization of human performance within those systems.

7 Acknowledgments

This research was supported by ATR Interpreting Telephony Laboratory, ATR International. Thanks to Drs. Phil Cohen, Megumi Kameyama, and Ray Perrault of SRI International, Dr. Jurgen Sottung of AT&T's Language Line, and Dr. Wilhelm Weber, Dean of Translation and Interpretation at the Monterey Institute of International Studies, all of whom provided valuable contributions on various aspects of this research. Of course, any errors or controversial views expressed herein are entirely those of the author.

References

[1] B. Butterworth, R. R. Hine, and H. D. Brady. Speech and interaction in sound-only communication channels. *Semiotica*, 20(1-2):81–99, 1977.

[2] P. R. Cohen, J. Morgan, and M. E. Pollack, editors. *Intentions in Communication.* MIT Press, Cambridge, Massachusetts, 1990.

[3] H. Iida, K. Kogure, I. Nogaito, and T. Aizawa. Analysis of telephone conversations through an interpreter. Technical Report TR-1-002, ATR Interpreting Telephony Laboratories, Osaka, Japan, May 1987.

[4] H. Iida, M. Kume, I. Nogaito, and T. Aizawa. Collection of interpreted telephone conversation data. Technical Report TR-1-000X, ATR Interpreting Telephony Laboratories, Osaka, Japan, October 1987.

[5] M. Kay, J. M. Gawron, and P. Norvig. Verbmobil: A translation system for face-to-face dialog. Technical Report, Center for the Study of Language and Information, Stanford University, forthcoming.

[6] S. L. Oviatt. Management of miscommunications: Toward a system for automatic telephone interpretation of Japanese-English dialogues. Technical Report 438, Artificial Intelligence Center, SRI International, Menlo Park, California, May 1988.

[7] S. L. Oviatt and P. R. Cohen. Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. *Computer Speech and Language*, 5(4):297–326, 1991a.

[8] S. L. Oviatt and P. R. Cohen. The contributing influence of speech and interaction on human discourse patterns. In J. W. Sullivan and S. W. Tyler, editors, *Intelligent User Interfaces*, ACM Press Frontier Series, Addison-Wesley Publishing Co., New York, New York, ch. 3, 69–83, 1991b.

[9] S. L. Oviatt and P. R. Cohen. Spoken language in interpreted telephone dialogues. *Computer Speech and Language*, 6, 1992, in press.

[10] S. L. Oviatt, P. R. Cohen, and A. M. Podlozny. Spoken language in interpreted telephone dialogues. In *Proceedings of the 1990 International Conference on Spoken Language Processing*, the Acoustical Society of Japan, Kobe, Japan, 1305–1308, 1990.

[11] K. Sparck Jones. Discourse modelling: Where are we now, and where should we be going? In *Fall Symposium Series on Discourse Structure in Natural Language Understanding and Generation*, American Association for Artificial Intelligence, Asilomar, California, 142–145, November 1991, working notes from panel discussion.

[12] B. Webber. Discourse modelling: Life at the bottom. In *Fall Symposium Series on Discourse Structure in Natural Language Understanding and Generation*, American Association for Artificial Intelligence, Asilomar, California, 146–151, November 1991, working notes from panel discussion.

Appendix: Transcription Samples of Interpreted and Noninterpreted Telephone Calls

This appendix includes segments from part of task 1 for subject #11's interpreted and noninterpreted calls, and for subject #1's interpreted call. The heading on top of each page identifies the type of call and subject. To protect anonymity, subjects' names have been altered throughout the transcriptions.

Interpreted Call

Subject #11; Interpreter #3

English Speaker	Interpreter	Japanese Speaker
	⇐	
Hello?		
	Hello, this is Japanese interpreter six-five-four. May I help you?	
Yes, I need to, uh register for the Pacific Rim Internatioual Conference * [on] Artificial Intelligence.	Yes, all right. [?XX?]	
	All right. Can I ask your full name, please?	
Harry.		
	Could you spell, please?	
H-A-R-R-Y.		
	Uh huh.	
L.		
	L.	
Shoemaker.		
	Shoemaker. Is this, ah your last name?	
Yes, it is.		
	Okay, Shoemaker, hh', all right. Then, ah would you like to have the, ah form by, ah FAX or by airmail, or...	
Would you please send the form by FAX?		

Subject #11; Interpreter #3

English Speaker	Interpreter	Japanese Speaker
	FAX, okay.	
And, and let me give you the FAX number.		
	Yes, can I have a FAX number, please?	
Four–zero–eight–,		
	Four–zero–eight–,	
eight–zero–three–,		
	eight–zero–three–,	
four–five–four–six.		
	four–five–four–six. Okay, I will repeat. Your name is Harry L. Shoemaker. Then FAX number is four–oh–eight–, eight–oh–three–, four–five–four–six.	
Mhmm.		
	Okay, just a moment. ⇒ もしもし。 *Hello.*	
		はい。 *Yes.*
	はい、またよろしくお願いいたします。 *Yes, hello again.*	
		よろしくお願いいたします。 *Hello.*

Subject #11; Interpreter #3

English Speaker	Interpreter	Japanese Speaker
	あのういまこちらの男性の方がですね、* あのう用紙を、会議場に関する用紙を、ファックスで送っていただきたいとおっしゃっておりますから、* ここにお名前をいかが、あの、伺いましたから * お願いします。まずお名前のほうが、ラリー、H-A、	はい。 はい、わかりました。 はい。
	Um, now the man here says, um, (he) would like (you) to please send the forms pertaining to the conference place, by FAX, so, here, would (his) name, um, (I) have requested it, so would (you) please, starting with (his) first name, Harry, H-A-,	*Yes.* *Yes. (I) see.* *Yes.*
		H-A、 *H-A-,*
	R-Y、 *R-Y.*	
		R-Y、 *R-Y.*
	これ、そしてミドルのイニシャル、あの真ん中にあのイニシャルがありまして、L、*	はい。 L、
	This, and the middle initial, um there is an initial in the middle, L.	*Yes, L.*
	はい、そしてご名字が、* シューメイカスペル * いいますけども、S-H-O、	はい。 はい。
	Yes, and (his) last name, Shoemaker, (I)'ll spell it, it's S-H-O-,	*Yes.* *Yes.*
		S-H-O、 *S-H-O-,*

Subject #11; Interpreter #3

English Speaker	Interpreter	Japanese Speaker
	E、 *E-,*	
		E、 *E-,*
	M-A、 *M-A-,*	
		M-A、 *M-A-,*
	K-E-Rです。 *and K-E-R.*	
	E-R、シューメイカー これが名字ですね。 *E-R, Shoemaker. This is (his) last name, okay?*	K-E-Rですね。• あ、わかりました。はい。• *K-E-R, right? Ah, (I) see, yes.*
	はい、そしてファックスのナンバーが、よん-ぜろ-はち、 *Yes, and the FAX number is four-zero-eight-,*	
		よん-ぜろ-はち、 *Four-zero-eight-,*
	はち-ぜろ-さん、 *eight-zero-three-,*	
		はち-ぜろ-さん、 *eight-zero-three-,*
	よん-ごお-よん-ろくです。 *four-five-four-six.*	
		よん-ごお-よん-ろくですね。 *four-five-four-six, right?*

Noninterpreted Call

Subject #11

Researcher	English Speaking Registrar
Hello?	
	Hello?
Hi, I'd like to register for the Pacific Rim International Conference on n' Artificial Intelligence * in Nagoya.	
	Okay.
	Okay.
And I need to, ah have you FAX me a copy of the registration form, if you may.	
	Okay, could you, could I have your name and FAX number?
Yes, my name is Harry [L.]	[Har'], Harry.
Yeah, Harry L. Shoemaker, S-H-, O-E-,	
	S-H-, O-E-?
M-A-K-E-R.	
	M-A-K-E-R.
Okay, and my FAX number * is four-zero-eight-,	Yes,
	four-zero-eight-,
six-three-zero-,	
	six-three-zero-,
five-four-[four-four].	[five]-four-four-four?
Mhmm.	
[Okay.]	Okay, we'll send it, [eh] immediately.
Now I need to find out some information also.	
	Okay.
Can you tell me how much the registration fee is, per person, for those who are *not* attending the banquet?	

Interpreted Call

Subject #1; Interpreter #1

English Speaker	Interpreter	Japanese Speaker
	⇒ もしもし。 *Hello.*	
		はい。 *Yes.*
	はい。お名前ですが、* えーと、はじめの方のファーストネームですね。 *Yes. The name is, let'(s) see, the first, the first name, okay?*	はい。 *Yes.*
		はい。 *Yes.*
	お名前がスティーブ、S-T、 *The name is Steve, S-T-,*	
		S-T、 *S-T-.*
	はい、東京のTですね。 *Yes, it's T- as in Tokyo.*	
		はい。 *Yes.*
	E、 *E-,*	
		E、 *E-,*
	V、 *V-,*	
		えっ *What?*
	V、ヴィー、ベジタブルの?XXX?ですね。 * *V-, Vee, that's ?XXX? as in vegetable, okay?*	あ、V、はい、 *Oh, V-, yes.*

Subject #1; Interpreter #1

English Speaker	Interpreter	Japanese Speaker
		はい。 *Yes.*
	E、 *E.*	
		E、 *E.*
	はい。であのー、こちらの名字のほうなんですが、• B、エービーの B ですね。 *Yes, and, um, the last name is, it's B–, as in A, B.*	はい。 *Yes.*
		はい。 *Yes.*
	R-E、 *R–E–,*	
		R-E、 *R–E–,*
	D-E-R、 *D–E–R.*	
		はい。 *Yes.*
	はい、こちらになってます。 *Yes, this is it.*	
		あ、わかりました。 *Ah, okay.*
	はい。 *Yes.*	
		あと、他にどういったご質問は。 *And, any other, what kind of questions...*

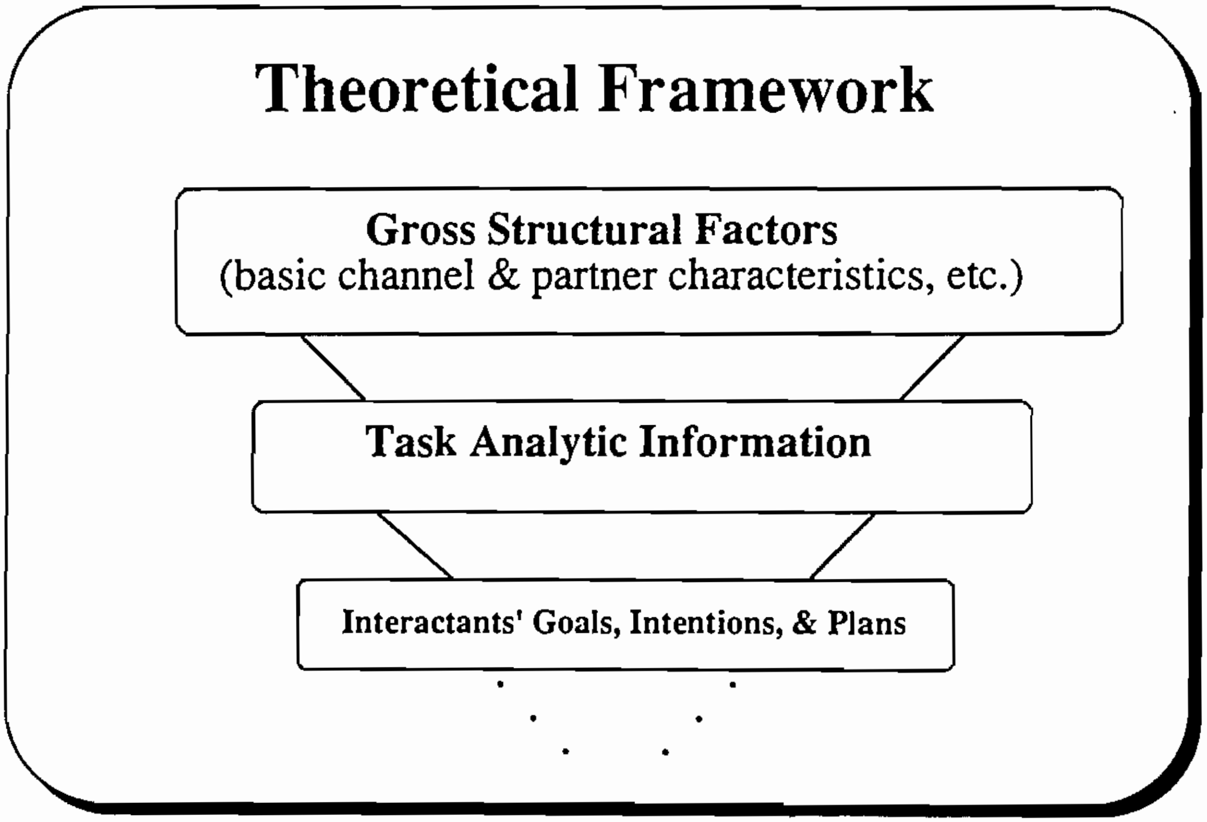
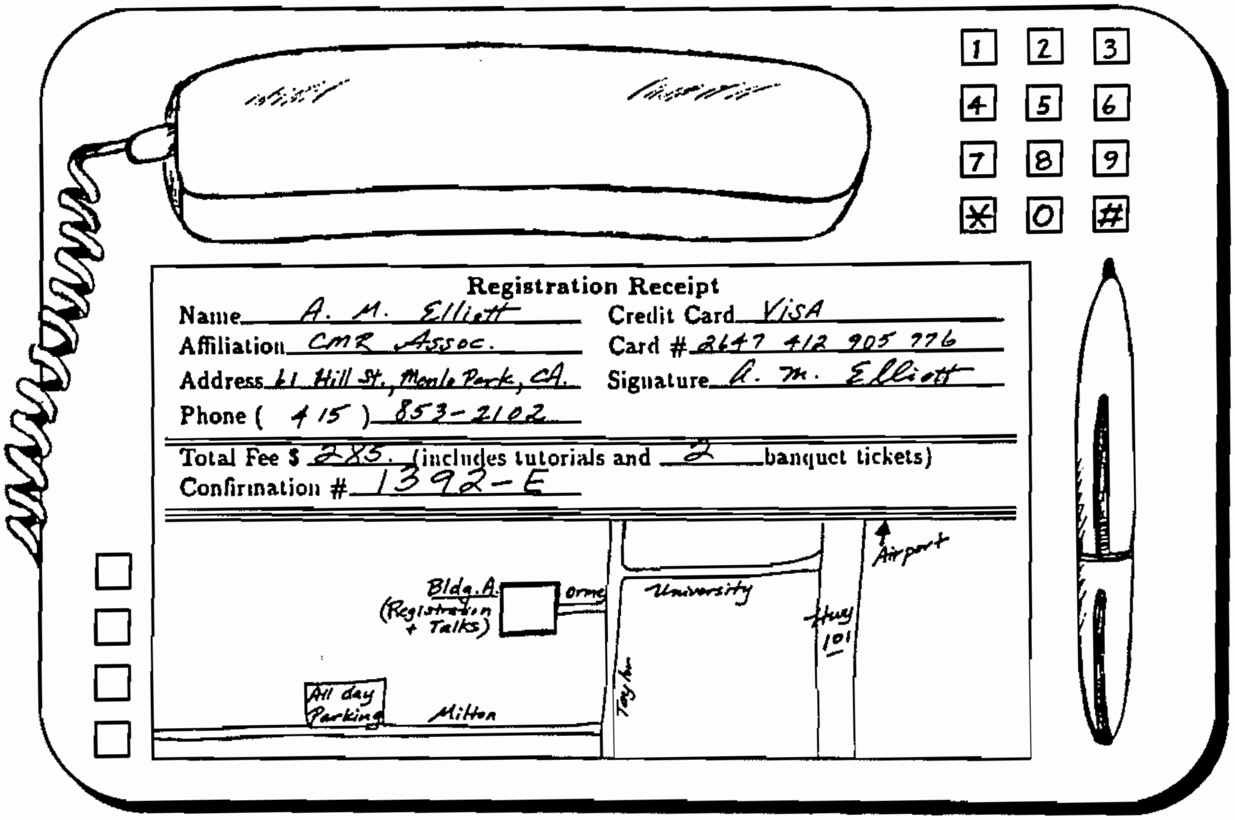

Theoretical Framework
Gross Structural Factors
(basic channel & partner characteristics, etc.)
Task Analytic Information
Interactants' Goals, Intentions, & Plans

Figure 1- Theoretical model of major factors influencing dialogue



Registration Receipt

Name A. M. Elliott Credit Card VISA
Affiliation CMR Assoc. Card # 2647 412 905 776
Address 61 Hill St., Menlo Park, CA. Signature A. M. Elliott
Phone (415) 853-2102

Total Fee $ 285. (includes tutorials and 2 banquet tickets)
Confirmation # 1392-E

Figure 2- Tablet presentation on a multimodal phone for a conference registration exchange