Understanding Acknowledgments

by

Karen Ward

B. Sci., University of Oregon, 1978

M.S., Oregon Graduate Institute of Science and Technology, 1992

A dissertation presented to the faculty of the

Oregon Graduate Institute of Science and Technology

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science and Engineering

June, 2001

The dissertation "Understanding Acknowledgments" by Karen Ward has been examined and approved by the following Examination Committee:

David G. Novick
Professor
Thesis Advisor

Peter A. Heeman
Assistant Professor

Sharon Oviatt
Professor

David Traum
Assistant Professor

Acknowledgments
(of another sort)

Several years ago, when I was struggling to understand what a Master's thesis might be, I spent a long afternoon in the library leafing through the theses and dissertations of students who had preceded me in this endeavor. I recall being struck by the fact that each began with several pages of flowery and sentimental acknowledgments that insisted, in one way or another, that the author could not have accomplished this work without the help, support, and guidance of what sometimes seemed to be a Cast of Thousands. I wondered whether this was part of the form of a thesis. Surely they exaggerate, I thought.

Now I look back on the 17 years that have passed since I first began taking classes at OGC as a non-matriculated student, and I realize that there was no exaggeration. If anything, words are inadequate to convey my indebtedness to my own Cast of Thousands who supported, who encouraged, who assisted or who simply didn't give up hope. And despite the certainty that I will inadvertently leave out many folks who should have been mentioned, I nonetheless attempt to thank them here.

To Debbie Gribskov, Becky Lakey, Bart Schaefer, Doug Pase, Fred Loney and the other students who befriended and encouraged me during those years I was a part-time student taking courses at OGC purely for my own amusement. To Dave Maier for making every course he taught seem like the most interesting and accessible topic imaginable. Yes, even Automata. You folks made OGC a special place to be.

To Buko, for selflessly encouraging and supporting me in the not-easy decision to quit my job and become a full-time Ph.D. student.

To Jenny Orr, Philipp Schmid, Nanda Kambhatla, Jeff Lewis, David Hansen, Brian Hansen, Mike Noel, Jon Inouye. What an amazing summer we had together, preparing for quals! We were some of the last students to have the opportunity to take charge of our own

learning in such a focused and cooperative manner. It was an empowering experience that I will never forget.

To the students, staff and faculty of CSLU, especially (in no particular order) Dan Burnett, Zhihong Hu, Yeshwandt Muthusamy, Mike Noel, Saarel Van Vuuren, Paul Hosom, Neena Jain, Jacques de Villiers, Johann Schalkwyck, Kal Shobaki, Johan Wouters, Terry Lander, Terri Durham, Charlene Edayan, David House, Ed Kaiser, Andrew Cronk, David Cole, Mike Macon, Todd Leen, Ron Cole, Misha Pavel, Hynek Hermansky. You created a supportive and intellectually stimulating place to grow as a researcher. Each of you at one time or another went out of your way to help me when I needed it and did so with generosity and graciousness.

To the intrepid and always-helpful library staff, especially Maureen, Mary Vatne Hultine and Kristine Roley. Your can-do attitude more than made up for the lack of library space.

To the administrative staff, particularly Shirley Kapsch, Barb Mosher, Julie Wilson and Anne Herman. You always find a way to make it work.

To Nick Horton, Marion Hakanson, Mark Morrissey and Bruce for being the best sysadmins one could ask for. You bailed me out more times that I can count and did so with patience and good humor. Bonus points to Nick for teaching me to juggle. Thanks also to Curt and his staff for last-minute assistance in getting the slides and signature page for my defense printed; few people knew that 30 minutes before show time I had most of the systems support staff engaged in trying various methods of getting those slides to paper.

To Jenny Orr, Philipp Schmid, Kay Berkling, Scott Daniels, Judy Cushing, Ira Smith, Tanya Widen, Walid Taha, David McGee, Takayuki Arai, Akiko Kusumoto for your friendship. It has meant a lot to me.

To the CSE faculty for not giving up on me. I especially thank Jim Hook; I've always been able to turn to you for sensible advice and encouragement.

To my committee for taking the time to read and comment upon this dissertation. I appreciate it greatly. The award for catching the most typos, sloppy wording and sloppy thinking goes to David Traum; I very much appreciate your constructive and positive manner of holding my feet to the fire.

To Peter Heeman for taking me in as an "orphaned" student. Your advice and encouragement, especially for the Switchboard study and the Wizard study, have been appreciated more than I probably managed to say.

To David Novick. For you, I do not have adequate words of thanks. You have been mentor, advocate, friend and life-saver. In all of these things, you have given me more than I deserve and more than can be repaid. I would not have begun this journey, and I most certainly could not have finished it, without your guidance and encouragement. I thank you for that great gift.

And, finally, to my family, for whom I've never had to be anything more than I am. Your love is the net that keeps me safe and secure no matter what happens. Mom, Robin and Ron, Marlin and Donna, Warren and Mary; Sami, Holly, Justine, Sean, Julia and Heather, who were not born when this began; Dad, Gram, Grandma and Jessica, who did not live to see it finished. I'm so far from home, yet I feel your love and support constantly.

And so I realize that since that long-ago afternoon in the library I've come to understand acknowledgments in more ways than I had expected. I thank you all.

Table of Contents

List of Tables

List of Figures

Abstract

Understanding Acknowledgments

Karen Ward

Ph. D., Oregon Graduate Institute of Science and Technology

June, 2001

Thesis Advisor: Dr. David G. Novick

As our ability to build robust and flexible spoken-language human-computer interfaces increases, we must consider whether and how we should incorporate various human-human discourse mechanisms into our dialogue models. In this dissertation I examine the use and probe the potential usefulness of one of these, acknowledgment. Acknowledgments signal understanding but not necessarily agreement; they serve to assure the conversants that information has been conveyed successfully. They also play a role in managing turn-taking.

Before we can incorporate acknowledgments in human-computer interfaces in an effective manner, there are several basic questions that should be answered. In this dissertation I report on a three-part research program in which I examine the use of acknowledgments in human-computer interaction from several perspectives:

- *Recognizing acknowledgments*: How can acknowledgments be recognized using low-level prosodic features and contextual cues? In two studies, I analyze corpora of human conversation for prosodic and contextual cues that might be useful for recognizing that an acknowledgment has occurred.
- *Predicting acknowledgments*: Can subjects predict where acknowledgments might occur in human-human dialogue? Two studies

probed subjects' ability to determine whether acknowledgments might occur after a turn.

- *Eliciting acknowledgments*: Are subjects are willing to use acknowledgments in human-computer interaction? I present and discuss a Wizard-of-Oz study in which subjects could control the presentation of information using either acknowledgments or commands.

By combining the three approaches, I was able to probe various aspects of the larger issue of understanding how and whether we should incorporate acknowledgments in spoken-language interfaces. Both the corpus studies and the perceptual studies suggest that dialogue-level context will be more important than local cues both for recognizing and for predicting (or generating) acknowledgment behavior in human-computer interfaces. The Wizard-of-Oz study shows that some subjects are willing to use acknowledgment as a turn-taking mechanism even in a fairly limited interface, although other subjects report resistance to the idea; more study is needed to understand the strength and implications of that resistance.

Chapter 1

Introduction

Over the past fifteen years, spoken-language human-computer interfaces have progressed from research systems that recognized a handful of disconnected words (e.g., Lee et al., 1990) to commercial dictation systems capable of recognizing over one hundred thousand words in fluent speech (Dragon, 2001).The quality of the interaction between human and computer differs markedly from that seen in human task-oriented conversation, however. In human conversation, contributions are offered and accepted routinely, often with little explicit verification that content has been conveyed accurately. Speakers exchange turns rapidly and smoothly with remarkably little overlap (Sacks et al., 1974). In human-computer spoken dialogue, however, the interaction is much less sure. Uncertain speech recognition may lead to frequent explicit requests for confirmation (McGee et al., 1998). Furthermore, human-computer interfaces are built around a one-sided, or single-initiative, interaction style in which the course of the interaction is controlled primarily by one party, be it system or user, with the other party left to respond in a relatively passive manner. Such systems typically support either an interaction based on verbal menus or forms, in which the system prompts a user through a series of choices, or a command-style interface in which the system responds to a series of user commands. Turn-taking may be controlled explicitly through mechanisms such as beeps ("please speak after the tone") or push-to-talk interfaces. Compared to talking with a human conversant, using a spoken-language interface can seem awkward, slow, and repetitive.

As our ability to build robust and flexible spoken-language human-computer interfaces increases, we must consider whether and how we should incorporate various human-human discourse mechanisms into our dialogue models. For straightforward tasks such as looking up a telephone number, for example, a single-initiative interaction may be both adequate and appropriate. For more complex tasks such as planning or scheduling, however, a more flexible style of interaction may offer benefits. People engaging in such tasks are seen to exhibit a flexible style of interaction, one in which either participant may take or decline a turn, suggest alternate courses of action, or even change the task itself (e.g., Gross et al., 1993). Conversants usually are able to negotiate meanings (Clark & Wilkes-Gibbs, 1986) and to determine that their contributions have been adequately understood (Clark & Schaefer, 1989). Human-computer interaction, by contrast, is typified by cumbersome confirmations, both explicit ("Did you say New York?") and implicit ("When do you want to go to New York?"). The more fluid interaction of human conversation is marked by the use of a rich suite of discourse mechanisms not currently in use in typical spoken-language interfaces. In this dissertation I examine the use and probe the potential usefulness of one of these, acknowledgment.

The term "acknowledgment" is used according to the definition put forth by Clark and Schaefer (1989). Acknowledgments signal understanding but not necessarily agreement; they often appear in American English conversation as "okay" or "uh-huh." Although acknowledgments contribute no new domain information to the conversation, they serve to assure the speaker that information has been conveyed successfully. They also play a role in managing turn-taking; although an acknowledgment may preface a new contribution by the same speaker (Novick & Sutton, 1994), often they occur alone as a single-phrase turn that appears to serve the purpose of explicitly declining an opportunity to take a turn (Sacks et al., 1974). Acknowledgments are common in many types of human-human conversation; in a corpus of problem-solving spoken dialogues, for example, Traum and Heeman (1996) found that 51 percent of turns began with or consisted of an explicit acknowledgment.

Despite their ubiquity in human conversation, acknowledgments are rarely incorporated into human-computer dialogue models. One possible explanation for this lack may be seen in the design of spoken-language interfaces. Current-generation interfaces are still relatively fragile, and so designers go to some effort to structure dialogues and create prompts that guide the user toward short, high-content, in-vocabulary responses (e.g., Basson et al., 1996; Cole, et al., 1997; Oviatt et al., 1994; Dybkjaer et al., 1996). Acknowledgments by definition contribute no new domain content, so dialogue models usually are designed to discourage their use.

For example, the dialogue models are often designed to support a style of interaction that might be characterized as master-slave (Grosz & Sidner, 1990): either the user is limited to responding to a series of system prompts, or the system is limited to responding to a series of user commands. In either case, such single-initiative interaction tends to discourage acknowledgment behavior for several reasons:

- An acknowledgement usually is not a cooperative response to a non-rhetorical question, at least not as the sole response; one does not respond to the request "please say your age" with a bald "okay," for example.
- Because the systems do not model acknowledgment behavior and so do not expect acknowledgments to be offered, an acknowledgment is likely to be either misunderstood as an in-vocabulary word or thrown away as an out-of-vocabulary word.
- The rigid command/response dialogue structure leaves no question at any point as to whose turn it is to speak, thus eliminating the need for acknowledgments as a turn-taking mechanism.
- The use of barge-in defeats the common interpretation of an acknowledgment. When the user speaks, the system contribution is cut off before the user utterance is interpreted. If the utterance was intended to signal that the contribution was understood and that the system should continue, the effect is exactly the opposite of the one desired.

Thus, current design and implementation practices both discourage and render meaningless the standard uses of acknowledgments. If these impediments were removed, what role might acknowledgments play in a human-computer interface?

In the long term, I expect that spoken-language and multi-modal interfaces will improve to the point that we will be able to participate in a conversation with a computer as smoothly and effortlessly as we talk to a person. Human conversations are rarely built around a master-slave model, unless there is an extreme power difference between the conversants. Instead, turn-taking and conversational initiative shift fluidly between conversants based on both verbal and non-verbal cues (e.g., Novick et al., 1996; Chu-Carroll & Brown, 1997). Once taking a turn becomes optional, once we move past a rigid model of a user simply responding to prompts or the system responding to commands, we must confront the problem of how best to communicate to the other party that one declines the opportunity to take a turn and that one understands the contribution well enough for the moment. The use and acceptance of acknowledgments may provide an intuitively obvious and convenient way to do this.

But I would argue that there are also more immediate reasons to apply a better understanding of acknowledgment use to spoken language interface design. With improved voice synthesis and spoken-language recognition, we are expanding the range of applications for which a spoken-language interface is useful. Some of these domains will require the presentation or acceptance of lengthy or complex information through spoken language, as for hands-busy applications or interfaces for the vision-impaired. Furthermore, system intelligibility may be less than ideal; spoken-language interfaces are being used in cars, for example, where street and engine noise may interfere with hearing. And voice synthesis, although improving rapidly, is still not as intelligible as human speech.

In situations such as these, people are especially likely to offer explicit verbal confirmation. While acknowledgments in such circumstances can certainly be explained in terms of signalling understanding, they also can be viewed in a more mechanistic sense as

controlling the pace at which information is presented. Acknowledgment may offer an intuitive means of improving the efficiency and usability of human-computer interaction by allowing users to control the flow of complex or lengthy information. By recognizing and responding to acknowledgment, for example, a system may be able to reduce the delay between contributions, thus speeding and smoothing the interaction while preserving a small granularity of information presentation. In situations where understanding may be task-critical, acknowledgments can be used to convey and verify understanding as information is presented without having to interrupt the presentation to request explicit confirmations. A system which accepts user acknowledgment may be able to reduce the occurrence of misunderstanding while still preserving an efficient transfer of information.

There may be additional applications of acknowledgment expertise, such as language training. Acknowledgment use differs across languages and cultures; Japanese, for example, has a rate of acknowledgment feedback that is approximately twice that of American English (Ward, 1996). At the same time, people learning a new language often have difficulty in understanding and hence need to be able to signal understanding or lack of understanding appropriately in the target language. If we better understood the rules for use of acknowledgment in various languages, we could incorporate that knowledge in conversational practice systems designed to give feedback to language learners about their use of acknowledgment and other dialogue management strategies.

Before we can incorporate acknowledgments in human-computer interfaces in an effective manner, there are several basic questions that should be answered. In this dissertation I report on a three-part research program in which I examine the use of acknowledgments in human-computer interaction from several perspectives:

- *Recognizing acknowledgments*: How can acknowledgments be recognized using low-level prosodic features and contextual cues? Acknowledgments in American English are expressed using words like "right," "uh-huh," and "yeah" that can be ambiguous with other domain or discourse usages, so we would like to be able to use additional sources of information to differentiate acknowledgments from other intended

meanings. In two studies, described in Chapter 4, I analyze corpora of human conversation for prosodic and contextual cues that might be useful for recognizing that an acknowledgment has occurred.

- *Predicting acknowledgments*: Can subjects predict where acknowledgments might occur in human-human dialogue? A corpus study can show us only where acknowledgments happened in a particular conversation, but not where they might have happened. We would like to know where acknowledgments might occur, however, both for generating acknowledgments appropriately and for recognizing acknowledgments. In Chapter 5, I report on two experiments in which I probe subjects' ability to determine whether acknowledgments might occur after a turn.
- *Eliciting acknowledgments*: Are subjects willing to use acknowledgments in human-computer interaction? Corpora of human-computer conversation show few if any uses of explicit acknowledgment. Is this because of system design and user expectation, or does it reflect a preference in interaction style? In Chapter 6, I present and discuss a Wizard-of-Oz study in which subjects had a choice in controlling the presentation of information using acknowledgments or commands.

In Chapter 2, I survey previous work in this area. Chapter 3 discusses the methodological issues that arose in designing these studies and explains the basis for the decisions made. Chapters 4 through 6 follow the outline of the questions above. In Chapter 7, I discuss the conclusions to be drawn from the work completed to date and describe the next steps to be taken in this research program.

Chapter 2

Related Work

In the previous chapter, I posed three questions about the role of acknowledgments in human-computer interaction: how can acknowledgments be recognized using low-level prosodic features and contextual cues, can subjects predict where acknowledgments might occur, and are subjects willing to use acknowledgments in human-computer interaction? Because acknowledgment behavior is explicitly discouraged in most spoken-language interfaces, there are few examples of acknowledgments in human-computer interaction from which we might form hypotheses. I turn, then, to research into the role of acknowledgment in human-human understanding. In this chapter I first describe the role of acknowledgments in terms of the collaborative view of conversation. I then discuss the cues that I investigate during the course of this study.

2.1 Acknowledgments and the Collaborative View of Conversation

The term "acknowledgment" is drawn from Clark and Schaefer's (1989) description of methods used in conversation to signal understanding. Clark and his colleagues explained conversation as collaborative process in which conversants work together to construct a mutual model of jointly held beliefs (Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989; Clark & Schaefer, 1989). In this view, conversants build upon a basis of shared knowledge drawn from the information considered to be commonly known, knowledge from their prior interaction, and information observed from the physical world around them. They add to this mutual model though collaboration; each

conversant makes contributions to the model of mutually-held beliefs, and both conversants are responsible for ensuring that a contribution has been understood "to a criterion sufficient for current purposes" (Clark & Schaefer, 1989, pg. 163).

In this collaborative view of conversation, dialogue is seen as a set of presentation-acceptance pairs in which a presentation by one conversant is accepted by another conversant through presentation of evidence of understanding. This evidence might consist of one of the following (taken from Clark & Schaefer, 1989, pg. 267):

- Continued Attention. By continuing to listen, B indicates that A's presentation has been understood to B's satisfaction.
- Initiation of the Relevant Next Contribution. B shows that A's contribution has been understood by starting in on the next relevant contribution.
- Acknowledgment. B nods, says "uh huh[1]", or makes some other overt indication that A has been understood.
- Demonstration. B demonstrates understanding, e.g., B performs the action that A has requested.
- Display. B repeats verbatim all or part of A's presentation, e.g., B repeats back the address that A has dictated.

Notice that an acknowledgment is itself a contribution to the conversation, thus requiring acknowledgment. How do conversants keep from generating an infinite loop of acceptances of acceptances? Clark and Schaefer proposed that the types of evidence are ordered from weakest (continued attention) to strongest (display); we accept a presentation at one level by offering evidence of understanding at a weaker level. Thus, one would expect display (repetition of the presentation) to be used when the need for confirming mutual understanding is very high, as in air traffic control communications (Ward, 1992),

1. Although Clark and Schaefer did not make this point explicitly, it should be noted that the words commonly used in American English to convey acknowledgment—"okay," "uh-huh," "yeah," and "right"—are also used for other discourse purposes, such as expressing agreement or committing to some future action. Not all instances of "uh-huh" are acknowledgments.

and that a moderately strong form of acceptance such as acknowledgment might be accepted by a weaker form such as next contribution.

There are difficulties with this model, particularly as a basis for designing a computational model of collaborative interaction. Although the definitions of presentation and acceptance seem straightforward enough, my own experiences in analyzing the presentation and acceptances seen in transcripts of real conversations suggest that they may be difficult to track reliably. This may be in part because the presentation/acceptance analyses are constructed from the perspective of an idealized overhearer and so may not reflect the perspective of any individual conversant (Novick et al., 1996b). Also, as Traum (1999b) noted, the analysis is constructed with knowledge of the subsequent course of the conversation, knowledge that would not be available either to a conversant or to a computational system attempting to represent the state of the conversation as it unfolds.

Traum (1999b) also observed that the ordering of the types of evidence is not always well-motivated. In particular, demonstration would seem to require comprehension of the intended meaning of the contribution whereas display merely requires perception of the syllables: I can parrot a short phrase in Spanish, a language I do not speak, without understanding what was said. The relative strength of the types of evidence would matter, of course, if we are to construct a model defining in quantitative terms how much evidence is needed to accept a given contribution. I touch on this issue indirectly in the perceptual studies reported in Chapter 5 in which I ask subjects to judge whether out-of-context presentations should be followed by (accepted by) an acknowledgment. The larger question of whether the relative strength of the types of evidence is correct and quantifiable—or even necessary—remains an open one, however.

Traum addressed these problems—and others—by proposing a Grounding Acts model that does not rely on two phases of presentation and acceptance with ranked evidence of acceptance. Instead, each discourse unit is grounded through one of a small set of grounding acts that are not considered to be ordered. The various evidences of understanding in the Clark and Schaefer model are all considered to be a single kind of

grounding act, which Traum termed acknowledgment. Traum's model did not address the issue of explaining why certain types of evidence are more appropriate in some situations than in others, although in later work he did discuss the importance of considering the cost of performing a grounding act in relation to the importance that the particular information be grounded (Traum, 1992). For example, display (repeating the other's contribution) is unremarkable when writing down a phone number but might be unsettling if used frequently in a casual conversation, and this may be explained in terms of the presumedly higher cost of display. This model is attractive in that it places more emphasis on the costs and benefits of grounding a particular discourse unit and sidesteps the problem of how to quantify various types of evidence of understanding. Because the research reported here is limited to a single type of grounding behavior, however, that shortcoming of the Clark and Schaefer model assumes less importance for this particular study.

2.1.1 ABC's of Acknowledgments: Assessments, Back-Channels and Continuers

Clark and Schaefer's definition suggests that an acknowledgment is fundamentally a brief contribution, that it is made as a speaker's turn in preference to a more extensive presentation, such as a relevant next contribution, and that it conveys only understanding of the previous contribution. This example from Novick and Sutton (1994) illustrates this form:

(1.1) Wizard: On Evans, you need to turn left and head West for approximately three quarters of a mile to Clermont.

(1.2) User: Okay.

(1.3) Wizard And, um, on Clermont you turn left, heading South for about two blocks to Iliff.

Novick and Sutton (1994) cataloged twelve patterns of acknowledgment behavior, including several in which an acknowledgment occurs at the beginning of a longer turn such as a relevant next contribution. They also identified and analyzed the case of acknowledgments appearing in the midst of an extended turn:

(4.1) Wizard: All right, um, the first thing you need to do is go

> South on Logan Street for one and a half miles to Evans Avenue. Then turn left on Evans Avenue and go one and a quarter miles to South Josephine Street. Okay, then you'll turn left on South Josephine Street. Nineteen Forty South Josephine is within the first block.

Novick and Sutton characterized this as a self-continuer, a temporizing move in which the speakers signals an intention to continue speaking.

Schegloff (1982) drew a distinction between two types of acknowledgment, continuers (such as "uh-huh" or "I see") and assessments (such as "wow"). Continuers are inserted during the other speaker's turn and convey an additional understanding that the primary speaker is not yet through speaking; these are also termed "back-channels." Assessments combine the acknowledgment function with an assessment of the previous contribution and are offered as separate turns.

Several researchers distinguished between an acknowledgment and a back-channel. Nigel Ward (1996), for example, defined back-channel feedback as follows:

1. responds directly to the content of an utterance of the speaker,
2. is optional, and
3. does not require acknowledgment by the speaker.

With this definition Ward excluded self-acknowledgments, acknowledgments that preface longer turns, and acknowledgments that do not appear to respond to the other's contribution. The second and third characteristics were intended to exclude answers to questions and back-channel-like presentations that function as questions, such as "huh?" In my work, these last two cases would be considered relevant next contributions and not acknowledgments at all.

In this research, I will use the term "acknowledgment" to refer to the broad class of acknowledgment behaviors, including acknowledgments which preface longer turns, and I will use the terms "assessment" and "back-channel" to specify the more specialized forms as identified above.

2.1.2 Evidence of Understanding in Telephone Dialogue

Clark and Schaefer's model of evidence of understanding was developed to account for face-to-face conversations. In the telephone dialogues with which this study will concern itself, however, conversants cannot see each other. This lack of visual communication between conversants presents some difficulties in interpreting the categories of demonstration, continued attention, and initiation of the next relevant contribution.

The category of demonstration presents problems when considering telephone conversation because it relies on the physical co-presence of the conversants. Demonstration assumes that the speaker is able to perceive the demonstrator's confirming action, for example when the speaker can see that the hearer is turning left on Clermont street as directed. Such direct perception is unlikely in telephone conversation unless the action is quite noisy or unless it has some long-distance effect that is within the speaker's perception. In the studies reported here, therefore, the category of demonstration is not used.

Traum (1999b) noted that the category of initiation of the next relevant contribution is problematic in that one may not be able to distinguish the case that a conversant accepts the previous contribution from the case that a conversant failed to hear or attend to the previous contribution, especially when the next contribution would be relevant whether or not the previous contribution had occurred. I would add that, in the absence a visual communicative channel, a similar problem exists for the category of continued attention: while physically-copresent conversants might signal continued attention with gaze (Novick et al., 1996), in telephone conversation it is not possible to distinguish the silence of continued attention from the silence of boredom or distraction.

Absent visual cues, a speaker might be able to deduce continued attention by noting that the hearer had passed up an opportunity to respond. To model a negative event (the listener declined to take a turn), though, it would be necessary to hypothesize the places where the speaker felicitously could have done otherwise. Sacks, Schegloff and Jefferson

(1974) proposed that turns are constructed out of smaller units (sentences, clauses, phrases) and that transition-relevance points, where turn-taking may occur, tend to fall at the boundaries of these units. This observation is not entirely helpful for modelling telephone conversation, though, because many such transition-relevance places are seen in a typical dialogue, and in the great majority of them neither turn-taking nor explicit back-channelling (declining the turn) occur. Copresent conversants may rely on gaze cues to signal which transition-relevence points are available for turn-taking (Novick, et al., 1996), but telephone conversants manage to exchange turns smoothly without visual feedback. This suggests that other cues are available to signal the availability of the floor, a hypothesis which will be expanded and tested in the studies described in Chapters 4 and 5. Until these cues can be defined, however, the category of continued attention remains problematic and so is not used.

2.2 Cues for Recognizing and Predicting Acknowledgments

For a spoken-language system to use and understand acknowledgment behavior, it must be able both to recognize and to offer acknowledgments appropriately. Although acknowledgments often are defined informally in terms of example words or phrases, as they have been in this chapter, words alone will not be sufficient for reliably identifying acknowledgements in context. Many of the phrases commonly used to convey acknowledgment are ambiguous with expressions of agreement ("yeah," "uh-huh") and other confirming words may be ambiguous within the specific task domain, such as "right" in the context of giving directions. We would like to have additional information available during the early phases of processing of the user's presentation to aid the system in identifying the intent of the message.

In this research I examine several possible cues to understanding and predicting the occurrence of acknowledgments. I focus on factors that I expect to be useful, available, and relatively robust in current systems, with particular attention to cues that would be available early in the spoken language understanding process so that the results of considering these

cues might be available for use in syntactic and semantic processing. The cues I chose to examine are dialogue context, pitch, pause, and word choice. These factors and their representation are discussed in greater detail in the following sections.

2.2.1 Dialogue Context in Terms of Speech Acts

The dialogue context in which a contribution is presented critically affects the interpretation that the hearer places upon it. The same words are easily understood to mean very different things in different contexts. For example, the question:

Do you know what time it is?

could represent, in the appropriate circumstances, either a simple request for information or a pointed suggestion that it is time to leave — and only rarely can it be interpreted as the simple "yes-no" question that its surface form would suggest. This suggests that we understand a common intention behind these sentences, the illocutionary force, that is separate from the surface form (Austin, 1962). For modelling conversational utterances in computational terms, we would like to have an abstract representation for representing and classifying these intentions.

Speech act theory provides a conceptual framework for describing the conversational context and effect of an utterance in abstract terms (Austin, 1962). Searle proposed that speech acts could be recognized and defined by a set of rules (Searle 69). Although Searle's rules are rarely used directly in interpreting human utterances, his work makes explicit the idea that the interpretation of an utterance depends upon the context in which the utterance was made and upon the beliefs of the conversants about the state of the conversation and of the world.

There have been efforts to develop a small list of basic, irreducible speech acts or to group speech acts into a small number of related families (Wierzbicka, 1987, for example). Several researchers have suggested heuristics for recognizing verbs that can describe speech acts (for example, Austin, 1962; Stubbs 1983). Austin estimated that there roughly one thousand speech act verbs in English, and he proposed a preliminary taxonomy based

on an intuitive classification of related verbs. Searle later proposed a hierarchical taxonomy based on similarities among the speech act properties (Searle, 1985). These general formulations tend to be difficult to apply to a specific task dialogue, though, and various domains may invite finer or coarser levels of detail for domain-specific activities. For example, in the Air Traffic Control domain it may be necessary to model "contact," which is the act of formally establishing communication between pilot and controller (Ward, 1992). Thus, it is not uncommon for researchers to develop or modify a speech act taxonomy to support the particular task being studied (Traum, 1999a).

In this work, then, I will follow common practice in using a set of speech acts that were drawn from several sources, particularly Traum (1996), and will augment those with corpus-specific notations. These will be described in detail in Chapter 4.

2.2.2 Pause

Pause has been found to be a strong marker for syntactic structure in read speech (Price et al., 1991), which suggests that pause may be a helpful marker for transition-relevance points. Wang's and Hirschberg's (1992) examination of intonational phrase boundaries in read speech indicated that both syntactic cues and phrase length play a role in the placement of pauses and phrase boundaries. Models derived from read speech may be misleading if applied too strictly to spontaneous speech, however; pause seems to be less closely linked to syntactic structure in spontaneous speech (Ferreira, 1993).

Much early work on pause in spontaneous speech focused on its role in monologue as an indicator of cognitive load (for example, Goldman-Eisler, 1958, 1961). Chafe suggested that cognitive load caused by factors such as changes in context (1985) and by cognitive constraints on reference (1986) may play a more prominent role. He found that pauses and prosodic downturns, for example, suggest that speakers organize their contributions in terms of short intonation units (one to five words) containing one new idea (Chafe, 1993). While this does not necessarily violate the syntactic-structure model, it does suggest that other factors may be as important as syntax alone in predicting and interpreting

pause in spontaneous speech. If pause's role as a marker for cognitive load in monologue can be extrapolated to dialogue, then we might expect between-turn pauses to be shorter preceding an acknowledgment (which should involve little cognitive load, as the acknowl edger is not formulating a new contribution).

Pause may also function as an indication that the speaker has finished the current contribution and is willing to yield the floor to the other speaker. The Japanese Discourse Research Initiative uses a pause of 400 milliseconds or more to define utterance boundaries in their corpus (Nakazato, 2000). Project LISTEN's mixed-initiative reading tutor system incorporated a rule that suggests that a back-channel response is appropriate after a two-second silence (Aist, 1998). Interestingly, in Aist's turn-taking specification the pause threshold for signalling a change in initiative from student to tutoring system is longer for taking a full turn than for offering a back-channel, although he did not mention the exact value for the former.

In summary, then, the literature suggests that pause might be useful as an indicator of speaker willingness to yield the floor. If hearers wait until they detect a pause to take or decline a turn, then we might expect acknowledgments to be preceded by noticeable silence. Other studies suggest that pause marks cognitive load, and we would not expect acknowledgments to impose great cognitive load because they contribute no new information to the conversation. If pause is a marker of cognitive load, then we might expect the pause preceding acknowledgments to be shorter than those preceding more substantive contributions.

2.2.3 Intonational Contour

Intonational contour reflects several phenomena: word-level (syllabic) stress, phrasal tunes and pitch accents, and pause. Phrasal tunes clearly offer valuable cues to the speaker's intentions. Pierrehumbert and Hirschberg (1990) proposed that tunes signal relationships between the propositional content and the mutual beliefs of the participants. Nakajima and Allen (1993) specifically examined the relationship between the

fundamental frequency (F0) and discourse structure in spontaneous task-oriented dialogue and found that F0 values tend to signal topic shift and topic continuation across pause boundaries.

Pitch prominence is associated with salient material (Pierrehumbert & Hirschberg, 1990) such as the contribution of new propositional content (Chafe, 1986). Because acknowledgments do not contribute new propositional content, I expect an acknowledgment to be associated with a lack of prosodic and conversational prominence, that is, to be de-emphasized and thus to be low in both pitch and energy.

Pitch and tune also offer cues as to the ending of syntactic and intonational units, and so may help signal transition-relevance points and thus places where acknowledgments might be needed. The duration of the final syllable in an intonational phrase in American English is typically lowered and compressed in pitch range and lengthened in duration (Pierrehumbert & Hirschberg, 1990; Ferreira, 1993). This accords well with Ward's (1996) findings that regions of low pitch signal appropriate places for acknowledgments to occur in Japanese conversation. I also expect that acknowledgments themselves, especially in their back-channel form, should exhibit a lower and compressed pitch range because of their status as a one-syllable turn.

2.3 Summary

In this chapter I defined the dialogue mechanism of acknowledgment and described the roles it plays in human-human conversation with regard to turn-taking and in helping to establish mutuality of understanding. I then described prior research into human conversation phenomena that suggests what acknowledgment behavior we might expect in human-computer interaction.

I described the cues that I expect to be useful in recognizing and anticipating acknowledgments in human-computer interaction. Based on our knowledge of these cues in human conversation, how might acknowledgments be recognized using these cues? The definition and role of acknowledgment in dialogue suggest that acknowledgments should

occur in the context of a contribution of information by the other conversant, particularly where the need to assure mutuality is high. The intonational contour should be low in both pitch and energy and the vowel duration lengthened, reflecting the acknowledgment's role as a one-word phrase (and often a one-word turn). The research on pause in spontaneous speech gives mixed signals; the cognitive load of producing an acknowledgment should be low, so one would expect the pause preceding an acknowledgment to be short. At the same time, pause may be a marker for a transition relevance point, and so the acknowledger may require more time to recognize the pause and decide to decline the turn.

Will subjects be able to predict where acknowledgments might occur in human-human dialogue? Acknowledgments play a role in both turn-taking and in ensuring mutuality, and their ubiquity in human conversation suggests that the role they play is important. Furthermore, cross-language differences in acknowledgment behavior can lead to breakdowns in communication (Oviatt & Cohen, 1992), suggesting that people do have culturally-specific expectations about the appropriate timing and frequency of acknowledgments.

Finally, are people willing to use acknowledgments when interacting with a computer? Human-computer interaction remains a difficult communication exercise, with speech-recognition errors, limited vocabulary, and voice synthesis imperfections creating many opportunities for misunderstanding. In situations where understanding is difficult, I expect that people would make use of human conversational techniques, including acknowledgment, to increase their assurance of mutual understanding if they were able to do so.

To test these hypotheses, a variety of methods will be required. In the next chapter, I describe and discuss the three experimental paradigms used in this project.

Chapter 3

Methodology

This chapter presents the overall method plan for this project, with discussion of the major methodology issues that arose. More detailed discussion of methodology and of specific hypothesis will be found in subsequent chapters, as each study is described.

Acknowledgments are a dialogue phenomenon that I believe to be important to a style of interaction not supported by current-generation spoken-language interfaces. Before I can directly test my hypotheses about its appearance, usability, and utility in human-computer interaction, I employ three experimental paradigms to approach the core issues from different perspectives. I first study corpora of human-human conversation to understand how acknowledgments occur in certain kinds of human conversation. I then turn to perceptual studies to probe human judgements of the appropriateness of acknowledgment based on local cues. Finally, I use a Wizard-of-Oz study to assess subjects' behavior with a simulated spoken-language interface. Each of these paradigms has strengths and weaknesses. In this chapter I identify these strengths and weaknesses and discuss the issues that arose in the course of designing these studies.

3.1 Corpus Studies

In a corpus study, one selects (or collects) an appropriate corpus of examples, labels the phenomena of interest, and then examines the labelled data to confirm or refute hypotheses. The results often are statistical in nature: how often does the hypothesis appear to account for the observed behavior?

There are several advantages to using a corpus of human conversation as a first step in studying spoken-language phenomena. Humans remain the unrivalled experts in language use, and so it makes sense to study examples of human conversation. With a recorded corpus, one can examine fleeting phenomena closely and use those observations to formulate and test initial hypotheses. Although the cost of collecting and annotating a corpus can be very high, once a corpus has been prepared one can use it for multiple studies of related phenomena. As standardized corpora such as Switchboard become available, researchers in different labs are able to directly compare results obtained against common data. Corpus studies have proven valuable in examining phoneme- and word-level phenomena, and they have been helpful in understanding the level of variation that a grammar might need to support.

At the same time, there are important limitations to corpus studies, especially for studying dialogue-level issues. Conversants' background knowledge and the conversational context play a strong role in the course of a conversation, and even very similar conversations diverge rapidly. The size, variation and complexity of even a brief conversation make it extremely expensive and difficult to collect the numbers of similar conversations needed for statistical analysis.

Because a corpus is static, it can be difficult to use it to explore alternatives. It is virtually impossible to find a "minimal pair" of conversations that differ in only one characteristic of interest, so constructing an experimental comparison of dialogue-level phenomena from material found in a corpus is problematic. We can observe what did happen in a particular conversation, but it is difficult to say anything about what would have happened if something were changed. Similarly, we can see one example of the course of a particular conversation, but we cannot determine what alternatives might have been deemed equally acceptable—or completely unacceptable—by the conversants. In other words, we can study what did happen, but it is difficult to draw conclusions about what might have happened or what could not have happened.

Finally, the use of corpora of human-human conversation limits the conclusions which can be drawn with respect to the design of human-computer interfaces. Even a fairly-focused task-oriented human conversation is likely to be more wide-ranging than current spoken-language systems can handle, particularly in terms of domain and world knowledge and often in terms of dialogue and task models as well, thus limiting the immediate applicability of the models that can be derived. Also, many studies (for example, Kennedy et al., 1988; Brennan, 1991; Okato, 1998) have demonstrated that people alter aspects of their speaking style when they believe that they are talking to a computer. It is not clear to what extent this reflects low user expectations as to the abilities of current-generation spoken-language interfaces and to what extent people simply do not feel that computers should be addressed in the same way as humans. Nonetheless, the phenomenon does require that we extrapolate from human-human to human-computer interaction with some care.

Despite these shortcomings, an examination of a corpus of human-human conversation can be a valuable first step in formulating and testing initial hypotheses.

3.1.1 Desired Characteristics

An ideal corpus would consist of spontaneous (as opposed to read) dyadic conversation, as that is the object of this research. The complete communicative signal must be captured; an audio record must certainly be present, and if the conversation took place face-to-face, the corpus must include a full video record. Because an accounting of the visual portion of the communicative signal is beyond the scope of this project, however, an audio-only conversation is preferred.

If we wish to probe some of the hypothesized relationships, such as those relating phrasal tunes and mutual beliefs of the conversants, we need conversations in which the domain and background knowledge are constrained enough to permit us to build and manipulate plausible models of conversant belief.

As explained above, it would be preferable to have a corpus of human-computer conversation if one could be found containing the dialogue phenomena of interest. Because current-generation systems are designed to discourage the use of acknowledgment, however, it is unlikely that corpora of human-computer dialogue would contain enough examples of acknowledgement to support a corpus study.

Transcriptions are needed for data analysis but are extremely time-consuming to produce, so a transcribed corpus would be helpful.

Even if we could find a corpus meeting all of the criteria outlined above, it still may not be adequate for our needs. An opportunistic examination of arbitrary conversation may not yield clear examples of the various cues to be examined.

3.1.2 Alternatives and Trade-Offs

Many of the relationships investigated in this work have been studied through examination of existing corpora such as the ATIS corpus (Hirschman, 1992). However, there are relatively few corpora of spontaneous speech and even fewer examples of dialogues where we have captured the complete communicative signal.

The studies reported in Chapter 4 make use of two existing corpora, both of human-human conversation. One, Switchboard, is an excerpt from a standard corpus that consists of short conversations on a set topic (credit cards, in the case of this particular excerpt). The other, the Vehicle Navigation System corpus, is a collection of direction-giving conversations.

Switchboard is a corpus of spontaneous telephone dialogue that was collected for the primary purpose of supporting research in speaker identification and topic-spotting, so the collections of ten-minute telephone conversations do not have the kind of strong task focus and limited domain knowledge that would encourage detailed computational modelling. The domain knowledge needed to understand these conversations is daunting, and it is not clear how to define the task and belief models involved in free-ranging

conversations about credit cards. The conversations do offer many examples of acknowledgement behavior.

The Vehicle Navigation System corpus is a collection of 93 brief direction-giving conversations between one of two experts and one of 21 travellers. The traveller, in a car and using a cell phone, drove to locate various addresses while consulting the expert for directions. The task in these conversations is clearer than that seen in the Switchboard corpus, and the amount of domain information needed to understand these conversations is much more constrained. This corpus suffers slightly from the structure of the task and the limited number of experts used, however: the two experts together account for about half of the corpus, so any conversational idiosyncrasies that they may have would tend to dominate the corpus.

Although each corpus has its shortcomings, both were deemed usable for the purpose of examining prosodic cues and speech-act-level contextual cues.

3.2 Perceptual Studies

A corpus study can suggest hypotheses—in this case, factors which may be useful in recognizing or signalling acknowledgments—but there are other questions that a corpus study cannot answer adequately. One shortcoming of a corpus study lies in the frozen nature of the data: we can examine what did happen, but it is much harder to ask questions about what might have happened. In the case of acknowledgments, for example, we can see where acknowledgments did occur, but we do not know whether acknowledgments might have occurred in other places. In the second phase of this project, then, I turned to a different method: I used a a pair of perceptual studies to probe subjects' judgments about the appropriateness of acknowledgments occurring after certain turns.

One means of investigating conversational alternatives is to generate the combination of situations that we wish to examine and ask human informants to rate or respond to the test utterances. This technique has the advantage of supporting a systematic examination of multiple factors, and it has been used successfully to examine such issues

as the interaction between syntax and prosody in read speech (Price et al., 1991). In working with spontaneous speech, however, it is not clear how to produce minimal pairs of utterances while retaining the characteristics of spontaneous speech. For example, it is very difficult for most people to deliberately insert filled or unfilled pauses into an utterance without sounding artificial. Similarly, it is not clear how to exactly reproduce or control for the context of the prior utterances in a conversation; even for very simple dialogues built around clear-cut tasks, we find that natural conversations quickly diverge. For the perceptual experiments, then, I relied on utterances drawn from an existing corpus, the Switchboard excerpt used in one of the corpus studies. Because I was interested in the effect of local cues without context, I extracted individual turns and asked subjects to rate the likelihood of an acknowledgment occurring after that turn.

One benefit of this approach is that we can better understand what cues are important to the subjects. For example, we may be able to measure some statistically significant difference between two sets of samples, but if subjects do not notice those differences then they may not prove reliable in practice in interpreting the signal. A drawback is that we are limited in the cases that we can test to those that we can find in the corpus.

3.2.1 Measuring Reliability

When asking subjects to perform judgement-based tasks such as identifying or rating dialogue phenomena, we are interested in understanding the extent to which they can agree on their judgements. If the subjects cannot agree that a particular phenomenon occurred at a particular place, then we cannot state that the phenomenon is reliably identifiable.

The kappa statistic (Carletta, 1996; Bakeman & Gottman, 1997) is accepted as a suitable method of assessing inter-rater reliability over disjoint categories, and the weighted kappa is accepted as a method of assessing inter-rater reliability over ranked categories. The kappa statistic measures the pair-wise agreement between two raters, taking into account

the agreement that would be expected by chance. Intuitively, it attempts to correct for the misleading level of agreement that occurs when some categories are rare. If one category accounts for 98 percent of the cases, for example, coders could agree on that one category while disagreeing on all others and they would still achieve a 98 percent agreement. That figure clearly does not reflect the fact that the coders agreed on the coding of only a single category and that they could not identify the rare cases reliably at all. The kappa calculation attempts to normalize for the expected chance agreement based on the number of times that each category is actually used in the coding[1]. To demonstrate that agreement exceeds levels that one would expect from chance, Bakeman and Gottman (1997) suggested that kappa should exceed 0.7; Carletta (1996) suggested that a good agreement should exceed 0.8, with tentative conclusions drawn when kappa is above 0.67.

Two refinements of the kappa calculation are used in this work, both described by Bakeman and Gottman (1997). First, when working with multiple raters—as is true of both of the perceptual studies reported in Chapter 5—the agreement between each possible pairs of raters is averaged to produce a kappa for the set of raters. Second, a weighted kappa calculation is used for the first of the two perceptual studies. The weighted kappa is appropriate when working with ranked categories; it attempts to captures the intuition that some disagreements are worse than others. In other words, a subject who judges an acknowledgment to be "Certain" is in closer agreement with a subject who selects "Very Likely" than with one who selects "Impossible" and kappa calculation should reflect that by penalizing some disagreements more than others. The use of the weighted kappa for ranked categories partially compensates for the problem that some subjects are more reluctant than others to use categories at the extremes of the scale.

---

1. More precisely, the number of times that each category is expected. In the absence of reliable expectations, however, actual counts are used (Bakeman and Gottman, 1997).

3.3 Wizard-of-Oz Studies

In the first two phases of this research, results were based either explicitly or implicitly on human-human conversation. As discussed in the first part of this chapter, people are known to change aspects of their conversational style when they believe that they are talking to a computer. This phenomenon limits the conclusions that can be drawn from examinations based on human conversation; we cannot know for certain how people would prefer to interact with a computer interface unless we put them in that position. At the same time, we cannot look to current systems to answer our questions because they are designed to discourage the phenomena that we are interested in studying.

It is difficult to build a real application that would allow us to test hypotheses relating to conversational control mechanisms such as acknowledgment. Conversational control and the turn-taking implied by mixed-initiative interaction are areas of active research (Haller et al., 1999). Furthermore, the less-constrained responses that such an interface might encourage would be likely to degrade recognition accuracy. The danger is that the interaction quality could begin to slip for reasons having little to do with acknowledgment behavior per se. At best, the subjects likely would respond by changing their interaction style to compensate for the communication difficulties; at worse, the interaction would fail entirely. In fact, that is what happened when I attempted a preliminary implementation of the Chapter 6 study using a fully-automated system: recognition errors and out-of-vocabulary responses led to a breakdown of the interaction and left the subject unable to complete the task.

I therefore employed the Wizard-of-Oz paradigm in the third phase of this research. In this approach, we simulate an interface that we are not yet capable of building by including a human in the loop, most frequently as the language understanding module. The wizard listens (in the case of an audio-only interface such as this one) to the subject inputs and selects or constructs the correct response based on a set of rules. The wizard is able to handle out-of-vocabulary and other problematic inputs in a more graceful fashion than a

fully-automated system, so that the interaction does not fail entirely over errors not related to the hypothesis.

This approach has two important limitations, though. The first is that the interaction speed of the interface is limited by human reaction time. In the case of the study reported here, the wizard's interface was a mouse-operated push-button GUI. The wizard had to hear the subject utterance, decide how to respond, move the mouse to the correct position, and click. While dialogue expectations allowed the wizard to pre-position the mouse in many cases, the total interaction time was still slow by human conversational standards. Because the entire interaction was slow, however, subjects seemed to adjust to the pace.

The other limitation is perhaps more subtle. While the human wizard can perform more reliably than does the automated interface, the wizard may make some errors and may behave slightly inconsistently in some cases. The errors and inconsistencies are unlikely to be those that a real system would make—which is, of course, the purpose in using the wizard in the first place. But while the Wizard-of-Oz interaction allows us to examine phenomena that are not easily supported by existing systems, we must bear in mind that an eventual computational implementation will have to function in the context of a system that is making more and different errors than were seen in the Wizard simulation. This difference may cause the interface as a whole to function quite differently than in the simulation, perhaps so much so as to limit the applicability of our conclusions.

Despite this caveat, the Wizard-of-Oz simulation provides the most accurate indication of the probable effect and effectiveness of an experimental dialogue model such as the one explored here.

3.4 Summary

In this chapter I presented the major methodological decisions that underlie my experiments and discussed the strengths and weaknesses of the approaches used. I showed how three experimental paradigms—corpus studies, perceptual studies, and Wizard-of-Oz studies—can be used to provide multiple perspectives on the same dialogue mechanisms: I

chose corpus studies to examine how low-level prosodic features and contextual cues might help us recognize acknowledgments; I chose perceptual studies to test whether subjects can predict whether acknowledgments might occur after particular utterances drawn from a corpus of human-human dialogue; I chose a Wizard-of-Oz study to examine subjects' willingness to use acknowledgment behavior in human-computer interaction. By employing a complementary suite of experiments, one can gain greater insight into this poorly-understood aspect of human communication and thus lay the groundwork for incorporating acknowledgment behavior into our computational dialogue models.

In the next three chapters of this dissertation, I present the experiments according to the experimental paradigm used. I begin in Chapter 4 with the corpus studies, then move to the perceptual studies in Chapter 5, and finally present the Wizard-of-Oz study in Chapter 6.

Chapter 4

Recognizing Acknowledgments

How might acknowledgments be recognized by a spoken-language system? In this chapter, I report on two corpus studies that probe the potential contribution of using certain low-level, local cues for recognizing acknowledgements. In both of these studies, I look for factors that I expect to be useful, available, and relatively robust in current systems. Thus, I was particularly interested in simple pitch and pause measures, plus the cues offered by the context in which the contribution was offered.

4.1 Pitch as a Cue to Recognizing Acknowledgments

In this experiment, I examined a potential interrelationship between a simple measure of pitch change and word usage. In particular, I hypothesized that word-level pitch information might provide exploitable cues to the usage of a word.

4.1.1 Experiment

The data for this study were drawn from the Vehicle Navigation System (VNS) corpus (Novick & Sutton, 1994), a collection of task-oriented human-human dialogues taking place over cellular telephone. In these conversations, one conversant (the traveller) is attempting to drive to several different addresses and is consulting the other conversant (the expert) for directions. There are a total of two experts and 21 travellers navigating to three destinations each, with each destination requiring from one to three conversations

between expert and traveller. The corpus comprises 93 brief conversations consisting of nine to 62 turns each. All conversants were native speakers of American English.

For this study, I focused on a single word to avoid phoneme effects on pitch and duration. In this corpus the word "right" appears frequently and is used in several distinct senses. The most common of these were

- "Right" as a direction, for example,

  "Right on 27th."

- "Right" as an affirmative answer to an explicit question, for example,

  "Are you still on your way there?"

  "Right, I'm there."

- "Right" as the acknowledgment of a contribution made by the other conversant, for example,

  "Turn left again heading north on Elizabeth."

  "Right."

"Right" as an answer and as an acknowledgment are intuitively similar, although they are technically different speech acts. Used as an answer, "right" conveys agreement. When used as an acknowledgment, as in the third example above, "right" functions as a signal that the previous utterance was understood but conveys no agreement. In the example, the speaker has not agreed to turn left on Elizabeth; it may be a one-way street that does not permit a left-hand turn. Acknowledgments and agreements may serve similar discourse functions, however, and analysis (discussed below) showed that they exhibited similar prosodic characteristics.

I hypothesized that there would be consistent prosodic differences in the way the word "right" was pronounced in each usage. To eliminate co-articulation effects, I included only occurrences that were turn or phrase-initial. That is, I considered only occurrences in which the word "right" was preceded by a pause or by the non-overlapping speech of the other conversant. Utterance-initial cases are also of particular interest from the standpoint of potential usefulness; at the point at which a system would be processing these occurrences, it would have relatively few syntactic cues from preceding words.

Table 4.1 Summary of Phrase-Initial Usage of "Right" in VNS Corpus

"Right" used as:	Example	Phrase-Initial Occurrences		
		Used by Expert	Used by Travell er	Total
Direction	E: "Right on 27th"	10	6	16
Answer	E: "Are you still on your way there?" T: "Right, I'm there."	23	2	25
Acknowledgment	E: "Turn left again heading north on Elizabeth." T: "Right."	4	10	14
Other	T: "Right now I believe I'm on Platte."	1	1	2
Total		38	19	57

With these restrictions, I identified a total of 57 occurrences. Two coders independently classified these according to the categorization described above. One coder worked from both audio tapes and a word-level transcription of the corpus. The other coder worked from transcription alone. Differences were resolved by discussion. The usage distribution is summarized in Table 4.1.

In devising a measure of intonation, I wanted a metric that would be usable in the context of a spoken language understanding system. In particular, it should be robust in the presence of pitch tracker inaccuracies caused by, for example, glottalization. After some preliminary tests, I settled on a simplified prosodic representation in which I divided the word into equal thirds and measured the change in average pitch between the first and last thirds of the word. This is an inexact measure of prosodic tune; it does not, for example, capture the variations described by Grigoriu et al. (1994). I eliminated pitch values higher than 350 Hertz and lower than 50 Hertz, but made no further effort to correct pitch-tracker errors.

Table 4.2 Comparison of Difference in Pitch Changes by Category

Comparison		p-value	Significance
"Right" used as:	Compared to:		
Acknowledgment	Directions	0.1007	Trend, not significant
Answers	Directions	0.1113	Trend, not significant
Acknowledgments plus answers	Directions	0.0375	Significant, confidence 0.95
Acknowledgments	Answers	0.4315	Cannot distinguish

4.1.2 Results

I compared the pitch changes found in the three categories of phrase-initial or utterance-initial "right" using the Welch Modified Two-Sample t-test. I found significant differences in pitch patterns; the results are summarized in Table 4.2. When "right" was used as an acknowledgment or answer, it was more likely to be pronounced with a falling intonation. When used as a direction, "right" was more likely to occur with a rising intonation. The individual comparison of the acknowledgments and answers categories with the directions category showed only a suggestive trend. When the acknowledgments were grouped with answers, however, the combined categories showed a significant difference compared to the directions category. The acknowledgments category could not be reliably distinguished from the answers category based on pitch change alone.

Although there is a statistically significant difference between the intonation of acknowledgments/answers and directions, the differences are not reliable enough to allow systems to use prosody alone to distinguish between these usages. If we choose flat intonation as the discriminator, assigning occurrences with flat or falling intonation to the acknowledgment/answer category and rising intonation to the directions category, instances are assigned to their correct category only 67 percent of the time (Table 4.3).

Table 4.3 Discrimination Based on Pitch Change Alone

True Category	Number Correct	Number Incorrect	Percent Correct
Acknowledgments plus answers	27	12	69%
Directions	11	5	69%
Other	0	2	0%
Total	38	19	67%

4.1.3 Discussion

The pitch change differences seen in this study reflect the intuitive observation that "right" used as an acknowledgment or answer is likely to be uttered with a falling pitch, which is typical of American English sentence-final intonation (Pierrehumbert & Hirschberg, 1990). In fact, 18 of the 39 acknowledgment/answer "rights" occurred as single-word turns. When "right" is used as a direction, it is usually the first word in a longer phrase and as such is more likely to exhibit a level or rising intonation.

How can this result help us build more robust systems? If pitch change alone is not an adequate discriminator, and if the prosodic differences merely reflect the tendency for acknowledgments and answers to be uttered as single-word utterances of phrases, then would a syntactic analysis serve the same purpose? In many cases, it probably would—if the recognizer is able to accurately return the words used in the utterance. When the recognizer returns an incorrect word string, however, syntactic analysis could be misleading.

Taken as one of many potential cues (for example, the context of the preceding utterance or the syntactic analysis of the utterance) a simple measure of pitch change could be helpful. Where potentially ambiguous words such as "right" occur at the beginning of a

longer utterance, for example, the direction of pitch change could serve as a confirming cue when analyzing ambiguous or erroneous recognizer output. Pitch cues potentially could be used dynamically to guide an integrated system in which acoustic analysis is interleaved with a probabilistic language model, as described by Goddeau (1992). Statistical pitch cues such as these could also be input as cues for a probabilistic parser.

Based on this experiment, I concluded that useful cues could be discovered using corpus inspection methods. Accordingly, I carried out a more detailed corpus study to expand the number and types of cues considered. This experiment is reported in the next section.

4.2 Multiple Cues for Recognizing Acknowledgments

What cues might point to the interpretation of a contribution as an acknowledgment? In this study, I hypothesized that several low-level factors might be useful in recognizing the presence of an acknowledgment in human-human conversation.

Acknowledgments can occur as back-channel contributions, that is, contributions that are offered without taking the floor. Furthermore, acknowledgments are commonly either the first or the only utterance in a turn. I would therefore expect acknowledgments to be more likely to occur in overlapped speech, either because the acknowledger does not take the floor or because the acknowledger begins the turn before the other has completely finished speaking.

As was discussed in Chapter 2, prosodic prominence is associated with the contribution of new propositional content in a conversation (Chafe, 1986). Because acknowledgments do not contribute new propositional content, I hypothesized that an acknowledgment might be associated with a lack of prosodic and conversational prominence. In informal terms, I expected acknowledgments to be de-emphasized and to be low in both pitch and energy.

From the definition of acknowledgment, I further expected that acknowledgments would be more likely to follow certain illocutionary acts and would be unlikely or impossible (in that they would result in a misunderstanding) after others. And, as suggested by the perceptual studies reported in Chapter 5, I expected explicit acknowledgments to be more likely following lengthy turns.

My hypotheses, then, were that in a corpus of human-human conversation:

- Acknowledgments are more likely to occur in overlapped speech than are other grounding acts[1].
- An acknowledgment will exhibit a smaller mean power and a lower mean pitch, normalized for speaker and word, than will the same word in other speech-act contexts[2].
- An acknowledgment will exhibit a shorter duration, normalized for speaker and word, than will the same word in other speech-act contexts.
- Acknowledgments are more likely to accept Inform speech acts than other types of speech act.
- The length of the pause preceding a turn-initial acknowledgment will be shorter than that preceding other types of grounding acts.
- Acknowledgments are more likely following lengthy turns than shorter turns.

These hypotheses are based on factors that I expect to be useful, available, and relatively robust in current systems. As in the previous experiment, I was particularly interested in cues that would be available relatively early in the spoken language understanding process, ideally in parallel with speech recognition.

4.2.1 Corpus Preparation

The corpus for this study was drawn from the Switchboard Corpus, NIST Speech Disc 8-1.1, March 1992. This corpus was originally designed to support speaker

1. The grounding acts considered are shown in Table 4.5 and discussed in Section 4.2.1.2.
2. The speech acts considered are shown in Table 4.4 and discussed in Section 4.2.1.1.

identification and topic identification research. Speakers were asked to carry on a telephone conversation about a particular topic; in the conversations used in this study, subjects discussed credit cards.

The Switchboard corpus has some drawbacks for the purposes of this study. The task is vague ("talk about credit cards for ten minutes") and has no clearly-defined goal, which makes it difficult to model belief states of conversants or to find the sort of task-related conversational structure that others have reported in more task-focused dialogues (Grosz & Sidner, 1990). Also, both conversants are human; it is possible that people will use acknowledgments differently when interacting with a computer.

Despite these problems, the corpus does offer several advantages:

- Many acknowledgments are present.
- The corpus includes both female and male speakers.
- The corpus was collected over the telephone, so non-verbal cues are not able to affect the listener's interpretation and thus need not be accounted for here.
- Each speaker's voice is (mostly) on a separate track, so overlapped speech can be analyzed separately.

4.2.1.1 Speech Act Coding

In this section, I discuss the speech act coding. Details of the corpus preparation, including the procedures used in labelling speech acts, may be found in Appendix A. Here I describe the coding scheme itself.

I elected to work with turns as the unit of segmentation, where a turn is an uninterrupted stretch of speech by a single speaker. This decision was motivated by my focus on cues that would be available early in the understanding process; turns are defined and can be detected in terms of the uninterpreted speech signal.

As discussed in Chapter 2, a turn may comprise multiple speech acts. My hypotheses, however, refer only to the speech act immediately preceding a contribution. For

this study, then, coders identified only the final speech act in the turn. They did not segment the dialogue (to identify where each speech act began) nor did they identify other speech acts that might have been present.

This relatively simple coding scheme is motivated in part by my goal of situating this work in the context of a human-computer interface; in a spoken-language system, the system would know what speech act was intended in its own contribution immediately preceding a user's response (although it generally will not know how the user interpreted it, of course). Also, the hypotheses to be tested did not depend on segmenting the corpus to that level of detail.

The speech acts used in coding this corpus (Table 4.4) were drawn from several sources, particularly Traum (1996). The set omits or combines some speech acts used by Traum that seem not to apply to this corpus[1] and includes the corpus-specific "Error in Transcript" (defined below).

Although this set of acts is intended primarily as a classification of illocutionary function, it includes a "grounding" category that Traum classifies as an entirely different level of coding. The grounding act as used here is intended to provide a classification for those utterances that consist only of an acknowledgment. This mixing of Illocutionary and Grounding Acts was necessitated by my decision to work with physically-defined turns instead of hand-labelled utterances. A turn may be incomplete due to interruptions and so may not embody a complete illocutionary act; this would be coded as "continue." Similarly, a turn may consist only of an acknowledgment, thus serving a grounding function, but because the turn has no propositional content it is difficult to classify it in terms of illocutionary function. Rather than force it into some relatively arbitrary and potentially ambiguous category of speech act, I provided the separate category of "grounding." Finally, I added the category "error in transcript" for the not-infrequent situation that the coder noticed a discrepancy between the transcript and the sound file. Those samples were skipped in subsequent processing.

1. Traum's set reflects the phenomena found in the more task-oriented TRAINS corpus.

Table 4.4 Illocutionary Acts Used in Speech Act Coding

Illocutionary Acts	Definition
Inform	The speaker provides new information (including providing requested information when answering a question).
YN Question	The speaker asks a yes-no-question
Wh Question	The speakers asks a wh-question
Expressive	The speaker expresses an attitude about the propositional content of the contribution
Performative	The speaker's utterance performs an action
Check	The speakers attempts to verify that certain information is true
Grounding	The speaker's turn consisted only of a grounding act such as an acknowledgement
Continue	The turn does not contain a complete illocutionary act
Error in Transcript	The transcript does not match the sound file.

4.2.1.2 Grounding Act Coding

Grounding acts were coded in a separate pass from illocutionary acts. The grounding act categories are summarized in Table 4.5. These categories were modified from the types of evidence of understanding suggested by Clark and Schaefer (1988).

Of the kinds of evidence of understanding defined by Clark and Schaefer, one was omitted as not relevant to a telephone interaction ("demonstration," to demonstrate understanding by performing an appropriate non-verbal action). The category of "continued attention" (the listener "shows he is continuing to attend and therefore remains satisfied with A's Presentation") also does not apply because of the selection of physically-defined turns as the coding unit. Essentially, "continued attention" is the case every place except where turn boundaries occur.

The category of "assessment" does not appear in Clark and Schaefer's hierarchy of acceptance acts. In this corpus, however, it was not uncommon to find speakers delivering an utterance that lexically might be classified as an acknowledgment but that was prosodically marked. Subjectively, the prosodically-marked utterances seem to convey additional messages beyond additional messages additional messages beyond simple acknowledgment, such as sympathy or sarcasm. Arguments could be made for:

- considering them to be acknowledgments (as with any relevant contribution, they in fact accept the previous contribution); or
- considering them to be next contributions (they seem to convey more than simple agreement); or
- considering them to be assessments (as defined by Sacks et al., 1974); or
- eliminating them from the analysis completely (it is possible that such presentations are more common in the casual human-human dialogue that is seen in the Switchboard corpus than in task-oriented human-computer dialogue).

For these reasons, I separated them in the coding so that their role could be analyzed more closely.

As with the speech act coding, two categories were added to account for particulars of the corpus: "continue" was used for samples which did not include a complete acceptance act, often due to interruption by the other speaker, and "error in transcript" was used for samples that did not match the transcript.

Table 4.5 Grounding Acts

Category	Definition
Not Heard/Ignored	The speaker did not seem to hear or chose to ignore the previous contribution.
Miscommunication[a]	The speaker indicated that the previous contribution was misunderstood (either physically or semantically).
Acknowledgment	The speaker indicated that the previous contribution was understood.
Assessment[b]	The speaker indicated understanding by offering a brief assessment or opinion of what the other has said, such as "oh, no." This category is used when an acknowledgment seems to be carrying an additional message such as agreement or sympathy.
Repetition[c]	The speaker signalled understanding by repeating all or part of the contribution.
Next Contribution[d]	The speaker accepted the contribution by continuing the conversation with a relevant next contribution.
Continue	The turn did not contain a complete grounding act, usually due to overlapping speech.
Error in Transcript	The transcript did not match the speech sample.

a. This category was termed "misunderstanding" by Clark and Schaefer (1989)
b. This category does not appear in Clark and Schaefer's hierarchy
c. This category was termed "display" by Clark and Schaefer
d. This category was termed "initiation of the relevant next contribution" by Clark and Schaefer

4.2.1.3 Measurements

In addition to the speech and grounding act coding, several measurements were made for each turn in the corpus. These were

- length in milliseconds of the gap between the end of the previous turn and the start of the current turn,
- normalized average pitch of the first words of the turn,
- normalized pitch of the first words of the turn,

- pitch range of the first words of the turn,
- normalized power of the first words of the turn,
- duration of the first words of the turn,
- pitch range of the entire turn,
- normalized average pitch of the entire turn,
- power range of the entire turn,
- normalized mean power of the entire turn, and
- length of the turn.

Pitch values were calculated using a pitch program developed by van Vuuren (1992). For the purpose of calculating pitch values, voiced and unvoiced segments were determined independently of the forced-aligned transcription. Pitch averages were calculated only over voiced segments.

Pitch and power values were normalized by dividing by the mean values for that speaker calculated across all turns for that speaker. In a speaker-independent system, of course, this figure would not be available at the start of a conversation. In such a system, however, the pitch/power means could be normalized by the cumulative values for the conversation as it progressed. Plots of the cumulative pitch/power means by turn for the speakers in this corpus suggest that by the seventh turn the cumulative means approach the mean across all utterances for that speaker.

4.2.2 Results

As will be seen in this section, this corpus did not provide sufficient numbers of samples to test all hypotheses. In particular, there were too few samples from any one speaker to test hypotheses relating to prosodic differences.

4.2.2.1 Acknowledgments in Overlapped Speech

> Hypothesis: Acknowledgments are more likely to occur in overlapped speech than are other grounding acts.

Overlapped speech was detected by comparing the start and ending time of consecutive turns. When one turn started before the end of the previous turn, that turn was considered overlapped. The extent of overlaps may be somewhat underestimated by this method, though; a contribution that occurs during a pause in the other speaker's contribution would not be counted as overlapped even though the contribution might be judged as an interruption by a human rater.

This effect was partially offset by omitting turns coded as "Continue." Continuations are artificial turns created by utterance division due to overlaps; to count them in the analysis as separate turns would be misleading, as the utterances in which they occur have already been counted.

"Assessments" are similar to "Acknowledgments" (as discussed in Section 4.2.1.2). Arguments could be made for including them with Acknowledgments, or for including them with non-acknowledgments, or for eliminating them from the analysis completely. I tried each approach; the results were unchanged in all cases.

The hypothesis that acknowledgments are more likely to occur in overlapped speech is not supported by these data. In this corpus, acknowledgments account for 29 percent of overlapped utterances and for 31 percent of non-overlapped utterances. The intuition of these proportions was confirmed with Chi Square applied to counts of turns in each category (X-squared = 0, df = 1, p-value = 1).

4.2.2.2 Acknowledgment Power and Pitch

> Hypothesis: An acknowledgment will exhibit a smaller mean power and a lower mean pitch, normalized for speaker and word, than will the same word in other speech-act contexts.

There were not enough data available to test this hypothesis. The corpus contained few instances of the same speaker using the same utterance-initial word in contrasting speech-act contexts. The most commonly-occurring cases are shown in Table 4.6.

There were not enough samples to test the original hypothesis, so I experimented with some larger groupings of the data. Pitch and power are strongly affected by syllabic stress; I therefore compared utterance-initial single-syllable words from a given speaker. Only a weak correlation was seen, however: normalized for speaker, acknowledgments were found to have a smaller power range, that is, less variability in power, p = .06. The correlation for normalized pitch range was even smaller, p = .07.

4.2.2.3 Acknowledgment Duration

> Hypothesis: An acknowledgment will exhibit a shorter duration, normalized for speaker and word, than will the same word in other speech-act contexts.

There were insufficient data available to test this hypothesis. As can be seen in Table 4.6, Speaker SW2987.A produced ten instances of turn-initial "yeah," seven of which were judge to be acknowledgments and three categorized as other types of contribution. The means of the two sets were similar (270 for the acknowledgments, 263.3 for the non-acknowledgments), and a standard two-sample t-test found no statistically-significant difference (t = 0.0987, df = 8, p-value = 0.9238). The same speaker produced eight instances of turn-initial "right," three acknowledgments and five non-acknowledgments. Although the means of the two sets showed a larger difference (236 for the acknowledgments, 300 for the non-acknowledgments), a standard two-sample t-test found no statistically-significant difference (t = -1.1217, df = 6, p-value = 0.3049). Other speakers produced even fewer occurrences.

To compensate, I experimented with more general groupings of the data. I compared the durations for all single-syllable, turn-initial phrases for one speaker, SW2987.A. This speaker produced 12 acknowledgments and 12 non-acknowledgments that fit the criteria. A standard two-sample t-test showed that the data do not support the hypothesis that these two sets differ in duration (t = -1.1793, df = 22, p-value = 0.2509, mean duration of acknowledgments = 261.6667, mean duration of non-acknowledgments = 340).

I compared the durations for the same lexical choice across all speakers in the acknowledgment and non-acknowledgments condition. For the turn-initial word "yeah" across all speakers, there were 33 occurrences that were coded as acknowledgments and 42 coded non-acknowledgments. A standard two-sample t-test showed weak significant differences at the 0.1 level (t = -1.6899, df = 73, p-value = 0.0953, mean duration of acknowledgments = 287.5758, mean duration of non-acknowledgments = 360.7143. Similarly, there were 20 turn-initial instances of the word "right" across all speakers. A standard two-sample t-test showed a weaker trend (t = -1.5029, df = 18, p-value = 0.1502, mean duration of acknowledgments = 288.75, mean duration of non-acknowledgements = 479.1667).

Finally, I compared all one-syllable turn-initial phrases (hand-coded) across all speakers in the corpus. This grouping resulted in a set of 51 acknowledgments and 111 non-acknowledgments. The results were not significant, although a standard two-sample t-test indicated a trend in that direction (p-value = 0.0781, mean of x = 312.54 ms, mean of y = 390.7207 ms). The two data sets have a high overlap, though, so it would be difficult to devise a good discriminating test based on this alone.

4.2.2.4 Speech Act Context of Acknowledgment

> Hypothesis: Acknowledgments are more likely to follow inform speech acts than to follow other types of speech act.

This hypothesis was supported by the data. Table 4.7 summarizes the types of grounding acts by the preceding speech act.

Inform speech acts conclude 54 percent of the turns in this corpus, but 92 percent of the turn-initial acknowledgments seen in this corpus follow inform speech acts, and 40 percent of inform speech acts are accepted by acknowledgments. Combining the similar assessments with acknowledgments, 87 percent of the combined acknowledgments/assess category occurs after informs, and 68 percent of Informs are accepted by either an acknowledgment or an assessment.

Table 4.6 Utterance-Initial Acknowledgment Words by Speaker

Speaker[a]	Word	Number of Occurrences as		
		Acknowl.	Assessment	Other[b]
SW1026.A	uh	1		8
	yeah	1	2	1
SW1026.B	well		2	2
SW2710.A	right	2	3	1
	yeah	3	2	1
SW2710.B	yeah	7	4	
SW2800.A	yeah	1		6
SW2800.B	yeah	5	5	2
SW2987.A	right	3	5	
	yeah	7	3	
SW2987.B	yeah	8	12	1

a. The speaker identification consists of the conversation number from the Switchboard corpus followed by A or B to differentiate the two conversants.
b. See Table 4.5 for a complete list.

Chi-square analysis confirms that the distributions differ, whether assessments are grouped with other grounding acts (chi-squared = 76.7366, df = 1, p-value = 0) or grouped with acknowledgments (chi-squared = 137.759, df = 1, p-value = 0), or considered as a separate category (chi-squared = 142.5327, df = 2, p-value = 0). Acknowledgments and assessments are more similar, but are still distinct (chi-squared = 4.5506, df = 1, p-value = 0.0329).

4.2.2.5 Acknowledgments and Preceding Pause Length

> Hypothesis: The length of the pause preceding a turn-initial acknowledgment will be shorter than that preceding other types of grounding acts.

The length of the gap (in milliseconds) between the end of the previous turn and the start of the current turn was compared for acknowledgments and non-acknowledgment using a standard two-sample t-test. Turns judged to be continuations of the previous turn were excluded from this analysis, as they do not begin with a grounding act.

In the case of a turn which begins while the other speaker is still speaking (overlapped speech), the preceding pause was considered to be 0. I considered this treatment to be consistent with the hypothesis, in that a pause is either present and has length or it is not present. I considered excluding overlaps from the analysis on the grounds that there was no pause, but that seemed to me to draw an artificial distinction between the case of a very short but mechanically measurable gap of, say, 50 milliseconds. and a similarly-short overlap over, say, a trailing nasal. Including the latter case in the analysis with a pause length of 0 seemed to capture the intuitive similarity of the two cases.

The data do not support the hypothesis, although with a p-value of 0.15 there may be a slight trend. The mean pause length for acknowledgements was 357 ms, compared to a mean of 546 for the non-acknowledgments.

Note that this hypothesis is not independent of the overlapped-speech hypothesis because of the definition of turns used in this study. By that definition, overlapped speech has a preceding pause length of 0 (or of a very small number).

4.2.2.6 Acknowledgments and Preceding Turn Length

> Hypothesis: Acknowledgments are more likely following lengthy turns than shorter turns.

The data support this hypothesis. The mean length of the 117 turns that were accepted by acknowledgments was 6.23 seconds, while the mean length of the 374 that were accepted by other grounding acts was 2.53 seconds. A Wilcoxon rank-sum test comparing these two sets confirms that these sets are distinct ($p = 0$).

Table 4.7 Grounding Acts by Preceding Speech Act

Speech Act ending previous turn	Total	Grounding (Accepting) Acts		
		Acknowledgment	Assessment	Other[a]
Inform	241	96	67	78
Other	202	8	16	178

a. See Table 4.5 for a complete list.

4.2.3 Discussion

Of the six hypotheses, only two—that acknowledgments were more likely to accept informs, and that acknowledgments were more likely following lengthy turns—were supported by this data. There were several cases of acknowledgments following speech acts other than Inform, although the definition of acknowledgment (and assessment) suggests that these should occur only after inform speech acts. As was seen in Table 4.7, however, that is not strictly true; a few were observed following grounding acts, and a few following expressives. These unexpected acknowledgments were not outliers due to an idiosyncratic speaker; all speakers did this at least once.

The small data set makes it difficult to draw any reliable conclusions on prosody-related issues; one needs either data from many more speakers, or much more data from each of these speakers. This also suggests that systems that attempt to do on-the-fly speaker normalization may find it difficult to make use of such measures.

It should be noted that the Chi Square test, which I use in several places, is appropriate only for independent samples. It is not clear that these samples are truly independent, in that it is possible that earlier portions of a conversation will affect the likelihood of acknowledgments occurring in later portions of a conversation. It is also quite likely that acknowledgments use will vary by speaker due to dialect and idiolect.

It also should be borne in mind that these results may be domain-specific. These are strangers that are engaged in a conversation with a broadly-defined task: they are exchanging information, getting to know each other. Would good friends use more or fewer acknowledgments? In this corpus, acknowledgements accept about 40 percent of the turns that end in informs—that is, when a person decides to speak after an inform, the person uses an acknowledgment about 40 percent of the time—and acknowledgments plus assessments account for 68 percent of the inform-ending turns. Are there local differences between the informs that are accepted by acknowledgments and those that are not? More interesting, from the standpoint of our long-term research agenda: would a person be more or less likely to use acknowledgments in a more strongly task-related interaction, and would they be willing to use acknowledgments in human-computer interaction at all? These questions are addressed in the following chapters.

Chapter 5

Predicting Acknowledgments

5.1 Introduction

This study focused on the characteristics of the turn preceding an acknowledgment. As discussed in Chapter 2, the decision to contribute an acknowledgment is assumed to depend heavily on the belief states of the conversants, particularly on the acknowledger's beliefs about the mutuality of the preceding contribution. The possibility remains, however, that some contributions may be more likely to elicit an acknowledgment than others. This possibility interested me for three reasons. First, if there are local factors that tend to elicit acknowledgments, then possibly these factors can be taken into account by a spoken language understanding system in deciding whether an acknowledgment has occurred. Second, a system that engages in mixed-initiative interaction will need to offer acknowledgments as well as understand them, and it would be helpful for such a system to know when an acknowledgment might be expected by the user. Finally, a better understanding of the factors that tend to elicit acknowledgment may be helpful in designing future studies of acknowledgment in human-computer interaction.

Do people consider some contributions to be more likely than others to elicit an acknowledgment? Can people reliably predict whether an acknowledgment will occur given only the local context of the immediate previous turn? I explored this question in two experiments. The first asked subjects to rate the likelihood of an acknowledgment occurring after an out-of-context turn. The second, a more constrained version of the first, asked subjects to judge whether an acknowledgment would occur after a particular turn.

5.2 Experiment 1: Predicting Acknowledgment Likelihood

The first experiment investigated whether subjects would consistently identify some contributions as being more likely than others to be followed by acknowledgment. This question was motivated by the hypothesis that there would be local, non-contextual differences between the two classes of samples; for example, we might expect prosodic upturns to elicit acknowledgments. In this study, then, I hypothesized that subjects would be able to reliably identify some turns as being more likely to be followed by acknowledgments than others. Subjects were presented with speech samples drawn from a human-human dyadic telephone conversation and were asked to rate the likelihood that an acknowledgment followed that sample.

5.2.1 Experiment

Six subjects—three male and three female—were presented with speech samples consisting of randomly-selected complete turns drawn from a nine-conversation subset of the Switchboard corpus. Some of samples were grammatically or prosodically "incomplete" due to interruptions by the other speaker. The subjects were asked to rate each sample on a scale of "certain," "very likely," "moderately likely," "very unlikely," or "impossible."

All subjects were native speakers of American English, and none had participated in any of my previous studies in this project. None of the subjects had a formal background in linguistics. Five of the six subjects were students or staff in the Computer Science Department of Oregon Graduate Institute. Subjects were paid $10.00 each for their participation.

Subjects read a one-page instruction sheet describing the mechanics of the study and offering a brief intuitive definition of acknowledgment. The definition and goal of the study were described as follows:

> An *acknowledgment* is a statement designed to let the other person know that you understand what they just said, but without expressing

> an opinion or answering a question. Acknowledgments often occur as a quiet little "uh-huh" or "right" in the background, but they can also be more explicit ("go on," for example).
>
> We are interested in understanding how people decide when it is appropriate to use acknowledgments in conversations. In this study you will listen to extracts of conversations in which two people chat about credit cards. We would like you to tell us whether an acknowledgment occurred immediately after that in the original conversation. In other words, in the original conversation one person said the sentence you will hear. Do you think that the other person responded to that sentence with an "uh-huh?"

The complete text of the subject instruction sheet may be found in Appendix B.

After the subject had read the instructions, the experimenter demonstrated the use of the interface and stayed in the room while subjects practiced on a separate set of ten samples. The interface is illustrated in Figure 5.1. Subjects were allowed to replay a sample or change their answers as often as they wished until they clicked the "Play Next Utterance" button to proceed to the next sample. Although records were kept of the changed answers and of the number of replays, these were not analyzed.

Subjects worked at a Sun Sparc 10 workstation. Because the volume levels of the speech files varied substantially, samples were played over the workstation's built-in audio device instead of over headphones. Subjects were provided with a separate volume control (not shown in Figure 5.1) and encouraged to adjust the volume as needed. The text of the speech sample was displayed to assist in understanding the decontextualized speech, although final punctuation was removed to avoid conveying transcriptionist judgements about the completeness of the utterance.

Samples were drawn from the Switchboard Corpus, NIST Speech Disc 8-1-1 (March, 1992). Samples were randomly-selected complete turns (as defined in Chapter 3) drawn from conversations SW1026, SW1088, SW2710, SW2800, SW2067, SW2313,

SW2409, SW2718, and SW2987 for a total of 18 speakers in the sample set. To control for order effects, the presentation order was randomized for each subject.

Each subject heard 150 samples. Due to an error, 25 responses were lost for one subject. The kappa calculations involving that subject reflect only the 125 responses for which the data were available for all subjects.

Subjects completed the task in 45-60 minutes. After coding was completed, the experimenter asked each subject to articulate any rules that they may have developed while working on the task (see Appendix E). During the post-experiment interview, the experimenter also solicited feedback about the experiment and answered the subject's questions about the experiment and the research project.

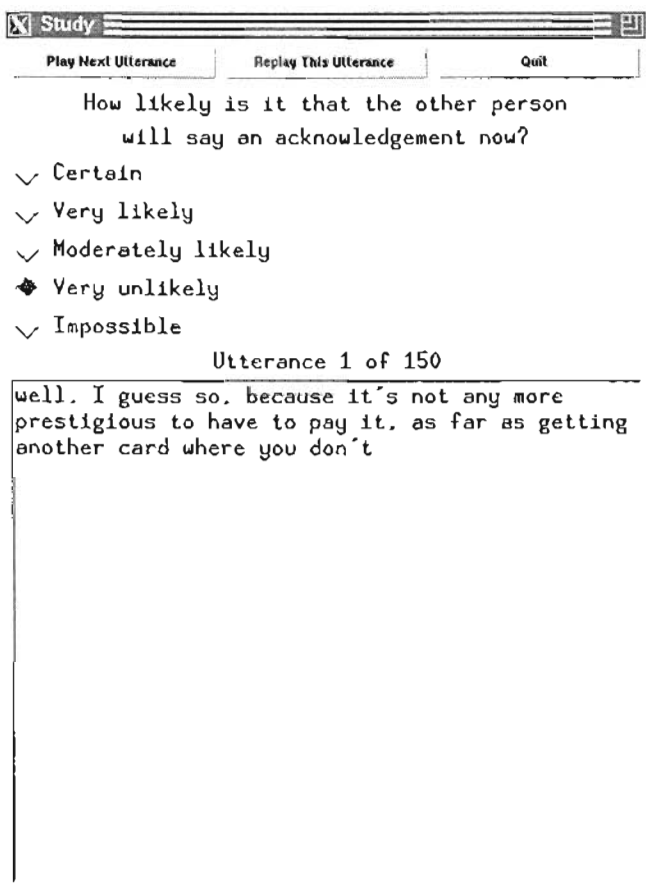

Figure 5.1 Interface for Predicting Acknowledgment Likelihood Study

5.2.2 Results

In this section, I present the results of the study without interpretation. The conclusions are presented in the Discussion section that follows.

As discussed in Chapter 3, the weighted kappa statistic is accepted as a suitable method of measuring inter-rater reliability over ranked categories (Carletta, 1996; Bakeman & Gottman, 1997). When working with multiple raters, the kappa statistic is calculated as the average of the kappa between each possible pairs of raters. A good reliability should exceed 0.8, although tentative conclusions may be found with kappa above 0.67. For this data set, average weighted kappa = 0.29 and pair-wise kappa ranged from 0.03 to 0.44, little better than chance. Detailed statistics may be found in Appendix D.

There was only one sample on which all subjects agreed exactly: the agreed-upon rating was "very likely," and the sample was a lengthy turn ending in an inform speech act in which the speaker was trying to define a particular type of restaurant. It ended with the phrase "it's, uh, you know, personal."

To determine whether a subset of samples was coded more consistently than others, the responses for each sample were ranked by the standard deviation of the subject responses. There were 52 samples (out of 125) that exhibited a standard deviation of 0.55 or less. Intuitively, these samples were coded with judgements that differed at most by one category. Twenty-three of the 52 ended in grounding or expressive speech acts (according to the previous in-context coding) or in my judgement were likely to be mistaken as groundings when presented out of context; an example would be a turn consisting of the single word "Yes" that had been offered as an answer to a question. Of the remaining 29 samples that ended in an inform speech act, 19 were coded (based on the mode of subject responses) as "moderately likely," that is, no strong opinion yes or no. Of the remaining ten cases, it appears that the following heuristic would account for the judgements: short, grammatically-incomplete presentations are coded as "very unlikely," and long, grammatically-complete presentations are coded as "very likely."

Subjects appeared to use the categories differently. Subjects 1 and 6 did not use the "certain" category at all, and no subject used it more than six times. Subjects 2, 5, and 6 rarely used "impossible," while Subject 1 used it for roughly one third of the samples.

5.2.3 Discussion

This study suggests that subjects cannot consistently judge the appropriateness of an acknowledgment from the decontextualized presentation of a turn from a conversation. In considering the results, I see several factors that may have contributed to the subjects' difficulty.

As expected, the lack of conversational context made the task challenging; out-of-context utterances are harder to understand (Novick et al., 1995). Several subjects noted that they had had difficulty interpreting the meaning of the utterances and thus difficulty in deciding whether an acknowledgment would be appropriate.

The samples which the subjects generally agreed were unlikely to be followed by an acknowledgment (standard deviation of 0.55 or less) were those that, out of context, might be interpreted as acknowledgments or assessments. This correlation would be expected on speech-act grounds and is consistent with the results seen in the previous chapter. The samples that were coded as "informs" were either judged to be an inconclusive "moderately likely" or, it appears from the post-experimental interviews, were coded based on the subjects' perception of the completeness of the utterance.

Another source of difficulty was the corpus: the prosodic upturn stereotypically associated with prompting is relatively rare among these speakers. In listening to the samples informally, I noted only a few (eight or nine) that sounded as if they ended in upturns. Would another corpus have provided samples with other prosodic markers, or is this corpus typical?

Some subjects completely avoided one or both extremes of the rating scale. Subjects 1 and 6, for example, did not rate any samples "certain." Although the weighted kappa

compensates for this somewhat—by assigning a smaller penalty to judgements that differ by only one rating category—this still adversely affected inter-rater reliability statistics.

Finally, it may be unreasonable to expect naive subjects to reach consistent conclusions about dialogue phenomena based on relatively simple coding directions. Non-linguists are not used to thinking about dialogue in terms of contribution and acceptance, so subjects may have found the task unfamiliar and hard to understand. As evidence of this, the post-experiment interviews (Appendix E) suggest that the subjects were not always answering the same question. For example, while some subjects articulated the principle that acknowledgments were more likely to be appropriate after lengthy contributions, one subject (Subject 4) concluded that acknowledgments should be avoided after "long-winded" contributions to discourage the speaker from continuing. It is possible that, with explicit training, coders could be taught to produce consistent responses, although that would not have answered the question posed by this study.

5.3 Experiment 2: Predicting Acknowledgment Occurrence

This experiment was designed as a more constrained version of the previous one. As in the first experiment, I hypothesized that subjects would consistently identify some speech samples as being more likely than others to be followed by an acknowledgment. Instead of allowing the subjects a five-point scale of likelihoods, though, I provided them only with a forced choice between "Yes" and "No." When presented with out-of-context excerpts from a corpus of human-human conversation, are human subjects able to determine whether an acknowledgement followed the excerpt in the original conversation?

5.3.1 Experiment

The experimental procedure was similar to that for the previous experiment. Conversations from the Switchboard corpus were segmented by turn. Pairs of turns were selected such that, for each pair

- both were produced by the same speaker,

- only one speaker can be heard on each sample; samples with cross-channel echo or heavily overlapped speech were excluded,
- both were originally coded (in context) as ending in inform speech acts,
- one was followed by an acknowledgment and one was followed by some other grounding act, for example, a next contribution,
- the length of the two turns were similar (within 15 percent for 85 percent of the samples).

A total of 88 samples from twelve speakers were used, as shown in Table 5.1.

Presentations were controlled for turn length because subjects in the previous experiment had indicated in post-experiment interviews that turn length was a factor in assessing the likelihood that an acknowledgment will occur. Where more than four matched pairs were available for a given speaker, short, long, and medium-length pairs were selected.

I attempted to limit the "not acknowledgment" cases to samples which had been grounded using "next contribution" in case there may have been other differences that prompted other acts. With that constraint, however, I found that I could not always achieve good match on length. I therefore relaxed that restriction and used turns which had been accepted with other kinds of grounding acts. As will be seen in the discussion of results, this did not affect the outcome of the experiment.

To control for order effects, the order of sample presentation was randomized for each subject.

5.3.2 Subjects

There were five subjects, three female and two male. Each subject was paid $10.00. All were native speakers of American English. Four had no significant background in linguistics, and one had had a course in natural language processing. None had participated in any of the other studies reported here. A summary of subject profiles may be found in Appendix F.

Table 5.1 Samples by Speaker

Speaker ID	Number of pairs
sw1026 0	4 (8 samples)
sw1026 1	4 (8 samples)
sw1088 0	1 (2 samples)
sw1088 0	1 (2 samples)
sw2710 0	4 (8 samples)
sw2710 1	7 (14 samples)
sw2718 0	7 (14 samples)
sw2718 1	1 (2 samples)
sw2800 0	1 (2 samples)
sw2800 1	4 (8 samples)
sw2987 0	6 (12 samples)
sw2987 1	4 (8 samples)

Subjects read a brief description of the task and signed a standard release form. The written instructions were similar to those used in the previous experiment. The full text may be found in Appendix B; the main instructions were

> In this study you will listen to extracts of conversations in which two people chat about credit cards. We would like you to tell us whether an acknowledgment occurred immediately after that in the original conversation. In other words, in the original conversation one person said the sentence you will hear. Do you think that the other person responded to that sentence with an "uh-huh?"

Subjects were informed that the other conversant did in fact say something at that point and that their task was to determine whether that something was an acknowledgment or something else ("yes" or "no," forced choice).

As in the previous experiment, the experimenter demonstrated the use of the interface (Figure 5.2) and remained in the room while the subject practiced with ten samples.

Samples were presented one at a time. To help the subjects understand the utterance, the display included the text of the utterance; the trailing punctuation was removed to avoid giving explicit cues about the completeness of the utterance. Subjects were allowed to replay the sample and to change their answer as many times as desired before moving on the next sample.

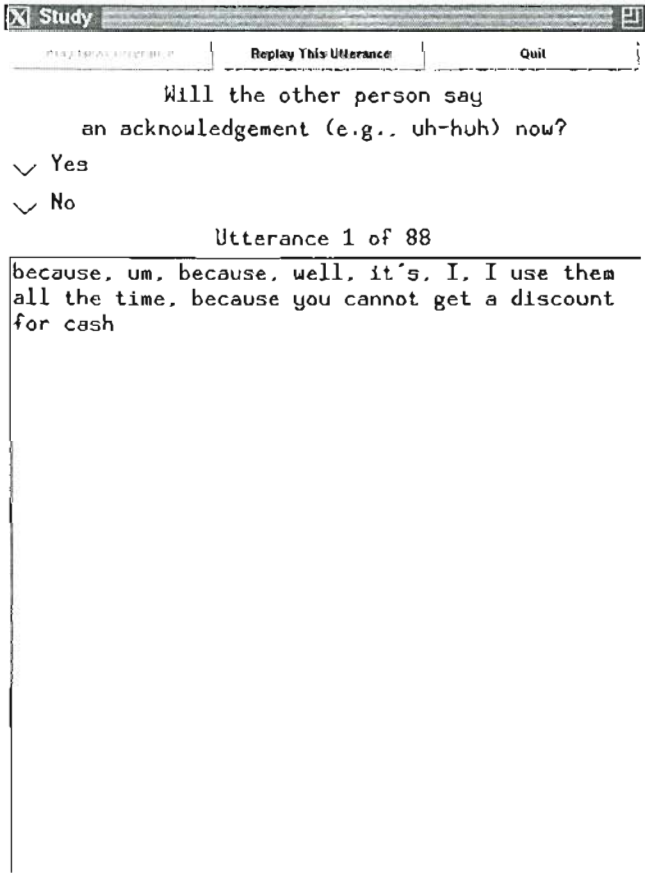

Figure 5.2 Interface for Predicting Acknowledgment Occurrence Study

Table 5.2 Percent Agreement and Kappa

Subject	Original	S1	S2	S3	S4
S1	50% (0.00)				
S2	52% (0.03)	48% (0.02)			
S3	42% (-0.16)	56% (0.11)	58% (0.18)		
S4	60% (0.16)	56% (0.11)	53% (0.04)	59% (0.18)	
S5	50% (0.00)	48% (-0.01)	57% (-0.06)	40% (-0.18)	44% (-0.12)

5.3.3 Results

In this section, I present the results of the study without interpretation. The conclusions to be drawn are presented in the discussion that follows this one.

Percent agreements and kappa were calculated for each pair of subjects. In addition, the agreement between each subject and the responses seen in the corpus were calculated. These results are shown in Table 5.2. The average kappa across all pairs of raters is 0.03, which is essentially chance. Detailed results may be found in Appendix G.

It is possible that situations in which assessments were appropriate might have a high overlap with situations in which acknowledgments were appropriate (assessments were considered "not an acknowledgment" in this study). I therefore considered three subsets of the data: those samples that were originally followed by next contributions, those that were followed by assessments, and those that were followed by acknowledgments. Kappa was not calculated on the subjects' responses for these subsets; empty categories result in kappa=0 in all cases.

Table 5.3 Subject Agreement by Grounding Act Type

Subject	Percent match for samples originally followed by:		
	Next Contributions (18 total)	Assessments (21 total)	Acknowledgments (44 total)
S1	61	52	45
S2	39	10	81[a]
S3	50	38	39
S4	50	67	61
S5	28	42	64

a. S2 coded 80% of all responses "Yes."

Finally, I considered whether there was a subset of the samples on which all subjects agreed. There were two samples that all subjects coded "yes" correctly—that is, in which the subjects' judgment matched what had been seen in the corpus—and one that all subjects coded "no" correctly. There were also two samples that all subjects coded "yes" incorrectly, that is, in which the subjects' judgement did not mach. The five samples mentioned in this paragraph were from different speakers.

In post-experiment interviews, subjects were asked to articulate how they decided on their answers. Factors mentioned in deciding in the affirmative:

- Subject said "you know" (Subjects 3 and 4), or used an inquiring tone (Subjects 2 and 4).
- Subject sounded as if needed prompting (Subject 3).
- Lower ending pitch (indicates looking for response) (Subject 3).
- Long sample (Subject 2).
- Speaker was female (Subject 1, a female).
- Speaker used frequent pauses (Subject 4).
- Statement was incomplete (Subject 1) or short (Subject 5).

Factors mentioned in deciding in the negative:

- Affirmative statement, such as "absolutely" (Subjects 3 and 4)
- Statement ended in higher pitch, indicates not finished (Subject 3).
- The last part of the phrase had a negative word in it, such as "won't" (Subject 2).
- Speaker was male (Subject 1, a female).
- Statement was complete (Subject 5).

Notes of the factors mentioned by each subject may be found in Appendix H.

5.3.4 Discussion

This study fails to show that subjects can determine whether an acknowledgment followed an excerpt from these conversations. Kappa hovered around 0 (chance agreement) for all subjects. Subjects agreed on the coding for only five of the samples, and their response in two of those cases was "wrong" in that the actual conversation had proceeded differently. There are not enough samples on which subjects agree to draw any firm conclusions about the characteristics of the data that might have elicited the agreement.

Post-experiment interviews suggest that subjects were using diverse criteria for making their decisions, with only two subjects citing prosodic considerations. It is possible that a higher agreement could be attained by specifically asking subjects to focus on prosody in making their decision or by presenting samples filtered to allow only prosodic information. Based on this study, however, I would expect the effects in such a study using samples such as these to be small at best. It is also possible that a more highly-structured, conventionalized task might identify points at which most people expect acknowledgments to occur, such as when transcribing complex information. For the unstructured, primarily-social task embodied in this corpus, however, I would expect agreement to remain near chance levels.

5.4 Conclusions

In this study, subjects were not able to judge reliably whether an acknowledgment is appropriate after an out-of-context presentation except in limited situations (the speech sample was obviously a question, for example). There may have been several reasons for this failure, including the unfamiliarity of the task and the limitations of the corpus from which the speech samples were drawn.

This study suggests that local prosodic and utterance-level cues alone are not sufficient for explaining where an acknowledgment should occur. This might be explained in terms of the collaborative view of conversation articulated by Clark and his colleagues, which would argue that the larger conversational context should play a major role in predicting the occurrence of acknowledgments in their grounding role. Subjects may have been unable to predict acknowledgments because grounding depends crucially on conversational context. What are the primary alternatives to an acknowledgment? In this corpus, the choices were usually a next contribution, an assessment, or a request for clarification. If the Clarkian view is correct, then the speaker's selection among these alternatives should depend heavily on the belief state. If instead prosody or some other local cue plays a major role in licensing an acknowledgment, then the lack of context should not matter and even should force the subjects to rely more heavily on the prosodic cues. These studies suggests that local prosody is not a definitive cue, at least not for these samples. Perhaps prosodic cues say "here's your chance" while other factors determine which option the speaker takes. In post-experiment interviews, some subjects did indeed report that they attended to prosodic cues. Despite the self-reports, their rating agreement remains poor; no pair of subjects managed to exceed 60 percent agreement or 60 percent correct (kappa $<$ 0.2). In fact, subjects reported using diverse strategies in deciding how to respond, suggesting that individual variation may play a large role in determining acknowledgment behavior.

Acknowledgments are credited with another role in conversational coordination, however: as a turn-taking mechanism (Sacks et al., 1974). It is in this role that I probe

subjects' willingness to offer acknowledgments in human-computer interaction. That study is reported in the next chapter.

Chapter 6

Eliciting Acknowledgments

6.1 Introduction

In this study I ask whether people will choose to use acknowledgments in human-computer interaction if they are given an interface that provides opportunities for and responds to acknowledgments.

6.1.1 Acknowledgments in Human-Computer Interaction

Acknowledgments occur less frequently in human-computer dialogue than in human-human dialogues for the same tasks (Okato, 1998). The reasons for this are not clear. One possibility is that, whether from popular culture or actual experience with a spoken-language interface, many people do not expect computers to understand "normal" language use and so change their contributions accordingly. Or perhaps people simply would prefer not to interact with computers in this fashion; if computers are viewed as tools, then a human-like collaborative style of interaction may not be seen as desirable by some users. Not everyone wants to chat with their refrigerator.

Regardless of the preferred style of interaction, a person who has used a current-generation spoken-language interface may not believe that they can use acknowledgment behavior. Current interfaces improve robustness by guiding the user toward short, in-vocabulary responses and by discouraging extraneous contributions (Basson et al., 1996; Hansen et al., 1996b). Acknowledgments contribute no new domain content to the

conversation, so most dialogue models are structured to discourage their use. In many systems, furthermore, turn-taking is completely controlled by one conversant; for example, the system may always prompts the user for the next piece of information or command. This rigid single-initiative dialogue model tends to eliminate the need for acknowledgments as a turn-taking mechanism: there is never any question as to whose turn it is. Also, an acknowledgment is not usually a felicitous response to a question, so the user has little opportunity to offer acknowledgments when all user contributions are responses to prompts. Other systems attempt to create interruptible system contributions through the use of barge-in technology. If the user speaks while the system is producing a response, the system contribution is cut off before the user utterance is interpreted. If that utterance was intended to signal that the system contribution should continue, the effect is exactly the opposite of the one desired.

Thus, current design practices both discourage and render meaningless the standard uses of acknowledgments. If these impediments were removed, would people choose to use acknowledgments when interacting with a computer interface?

6.2 Design Rationale

This study was designed as a first step into the effects of incorporating acknowledgement behavior in dialogue models for spoken-language interfaces. Before we can compare interfaces with and without acknowledgement behavior, we must understand the extent to which people are willing to use acknowledgments when interacting with a computer and establish a baseline for experimenting with various dialogue strategies.

6.2.1 Approach

In this study I hypothesized that subjects will choose to use acknowledgments in human-computer interaction if they are given an interface that provides opportunities for and responds to acknowledgments.

In designing the study, I assumed that it would not immediately occur to subjects that they could use acknowledgments to a computer. At the same time, I did not want to explicitly instruct or require subjects to use acknowledgment behavior, as that would tell us nothing about their preferences. I therefore decided against a comparison/control-group experimental design for this initial study and instead focused on creating a situation in which subjects would have a reason to use acknowledgments, perhaps even gain an advantage from doing so, while still keeping the behavior optional.

This study focuses on acknowledgment's role as a turn-taking mechanism. Conversants are especially likely to offer acknowledgments and repetitions when complex information is being presented, especially when the conversant is copying the information. While this is certainly explainable in terms of mutuality of understanding, this particular use of acknowledgment may be viewed from a more mechanical standpoint as regulating the pace at which information is presented. This insight suggested that a fruitful task for this study might be one in which the subject is asked to write down verbally-presented information, as when taking messages over the telephone.

6.2.2 Task

I selected the domain of telephone interface to e-mail and designed a task in which subjects were asked to transcribe items of information from the messages. Writing is slow in comparison to speaking, so I anticipated that subjects would require a slower pace of information presentation when they were writing. The messages included information not asked for on the question list to simulate "uninteresting" material that the subject would want to move through at a faster pace. In this way I hoped to motivate subjects to try to control the pace at which information was presented.

The e-mail was presented in segments roughly corresponding to a long phrase. After each segment, the system paused to give the subject time to make notes. If the subject said nothing, the system would continue by presenting the next message segment. Subjects could accept—and perhaps make use of—this delay, or they could reduce it either by

acknowledging the contribution, perhaps by saying "okay," or by commanding the system to continue by saying something like "go on." The system signalled the possibility of controlling the delay by prompting the subject "Are you ready to go on?" after the first pause. This prompting was repeated for every third pause in which the subject said nothing. I hoped this would suggest to the subjects that they could control the wait time without explicitly telling them how to do so.

I anticipated that subjects would control the pace in one of two ways; they might use a command to move the system onward ("go on," "next", "continue"), or they might use an acknowledgment ("okay," "uh-huh", or a repetition). On the surface, there is no functional difference between the two: in either case, the system responds by presenting the next message segment, and in fact it eventually presents the next segment even if the subject says nothing at all. The purpose of this design is not to probe potential advantages or disadvantages of one interaction style over another; instead, the goal is to create an interface that is neutral with respect to the use of acknowledgments or commands in order to see which the subject will prefer. Thus, the design allows the subject to choose freely between accepting the system's pace (system initiative), or commanding the system to continue (user initiative), or acknowledging the presentations in a fashion more typical of collaborative (mixed-initiative) human conversation. In this way, I hoped to understand how the subject preferred to interact with the computer.

If there is no functional difference in this case between the subject's use of an acknowledgment and a command, how can we determine which the subject intended? Acknowledgments can be recognized from the lexicalization of the utterance and from its discourse relationship with the previous utterance (Chu-Carroll & Brown, 1997). To confirm which the subject intended, the subject was asked during the post-experiment interview why they had selected the words that they had used.

6.3 Evolution of the Interface

The work reported here is intended as a baseline and proof-of-concept for a line of future work. During the course of the study, therefore, I elected to make changes in the interface to make the interaction less annoying for the subjects, and I altered some message texts to eliminate unintended ambiguity. These changes mean that between-subjects measures such as number of turns and task completion times are not meaningful, of course, and although the measurements are reported in Appendix L I draw no conclusions concerning them. In this section I describe how the interface evolved during the study.

6.3.1 Changes in Dialogue Design

The largest change involved the dialogue design. The interface was originally planned as a wholly computational system instead of a wizard-assisted one. To reduce speech recognition errors, the system was designed around two command states. In one state, it expected to receive a direction to read a message or summary or to quit; in the other, it expected to receive commands to navigate through a particular message. A subject thus had to quit one message explicitly before starting another. When preliminary tests suggested that the computational interface was still too brittle for the study, I made the minimal changes necessary to convert it to a Wizard-of-Oz interface with the wizard supplying the speech recognition. I did not redesign the dialogue structure, however.

After assessing the results and the subject feedback on the quality of the interaction for the first six subjects, I modified the interface to allow the wizard more flexibility in responding to the subject's commands. No additional functionality was provided, but the interface was changed to allow the subject to begin another message without explicitly quitting and then selecting another.

6.3.2 Changes in Prompts

I experimented slightly with the wording of the prompt used to ask the user whether the interface should continue the presentation. The first ten subjects heard the prompt "Are

you ready to go on, or should I repeat that?" This prompt is fairly long, and it was not uncommon for a subject to begin speaking just as the prompt began. Several reported some frustration at having to wait for the prompt to finish so that they could continue the task.

With the tenth and eleventh subjects, therefore, I tried a declining-length prompt designed to signal the system state while taking advantage of user experience with the system. The first time the subject heard that prompt, it was "Are you ready to go on?" The second time it was shortened to "Are you ready?", and all subsequent prompts were simply "Ready?" The two subjects who used this version of the system reported that the "Ready?" prompt was so short that it was very difficult to interpret and that it lacked a pronounced questioning inflection. I therefore changed to "Are you ready to go on?" for all prompts.

6.3.3 Changes in Message Texts

Two changes were made in the message texts in response to subject complaints that the texts were too long or difficult to understand. After the fourth subject, a particularly long segment in Message 4—"the workshop on Automatic Speech Recognition and Understanding will be held on December 12th through 15th, 1999, in Keystone, Colorado"—was divided into three segments. After the twelfth subject, the phrase "seventy-five watt light bulbs" in Message 6 was changed to "a package of seventy-five watt light bulbs."

6.3.4 Changes in Experimental Setup

For the first six subjects, the subject was seated in a conference room and the wizard was in a nearby machine room which housed several other computers and a printer. The background noise interfered with the speaker-phone, though, so that the wizard didn't always hear the subject responses. Also, the subjects complained of the noisy line (actually the machine noise). For the remaining 14 subjects, the setup was moved to a quieter pair of rooms.

6.4 Study

6.4.1 Subjects

All subjects were volunteers and were native speakers of North American English; most were staff or relatives of staff at a research university. Thirteen were female, seven were male. Ages ranged from 13 to 57. All used computers, typically office software and games, and only two—both in the first group of six subjects—had significant programming experience. One subject had used speech interfaces to airline information systems, three had seen demonstrations of research systems incorporating a spoken-language interface, and one had studied signal processing of speech. Subject profile information is summarized in Appendix J.

Each session lasted about 45 minutes total, and each subject was paid $10.00.

6.4.1.1 Instructions to Subjects

Subjects were told that the study's purpose was to assess the understandability and usability of the interface and that their task was to find the answers to the list of questions. They were given no instructions in the use of the program beyond the information that they could talk to it using normal, everyday speech. To head off possible frustration, subjects were told that some parts of the messages were somewhat hard to understand and that they should make only a reasonable effort to understand them. The subjects were asked to read the list of questions before beginning the call, ostensibly so that they would know what information they needed to find. Of course, the primary reason was to give the experimenter/wizard time to move to the wizard workstation.

Several subjects reported that the term "spoken language interface" was not meaningful to them, and that they did not expect to be able to talk to a computer over the telephone. After the first six subjects, one sentence was added to the subject instructions to emphasize that one was to talk to the program, not push buttons on the telephone handset.

The text of the subject instructions may be found in Appendix I.

6.4.2 Telephone Interface

The interface used in this study was constructed using the Rapid Application Developer in the CSLU (Center for Spoken Language Understanding) Toolkit (Sutton et al., 1998). A button panel allowed the wizard to select the appropriate response from the actions supported by the application. The application functionality was deliberately kept limited to suggest realistic abilities for a current spoken-language interface.

As described above, the initial dialogue model was designed around two command states. In one state, it expected the user to select a task. In this state, the system was able to

- begin reading a list of all messages. This list consisted of the message number, the sender, and the subject of each message;
- begin reading a particular message;
- ask the subject what to do next;
- end the program;
- play error/ help messages.

While reading a particular message or a list of messages, the system was in a state in which it expected commands relating to navigating through the message. It could

- read the next message segment;
- repeat the current message segment;
- wait (after approximately five seconds it would ask the subject whether it should continue reading the current message);
- repeat the previous message segment;
- ask the subject whether the program should continue reading the current message;
- play several error/ help messages.

The later version of the interface supported the same functions, but allowed the wizard to access any function at any time.

Message six is from Jo at teleport dot com, about, please stop by store on your way home.
I'm going to be late getting home tonight, so would you please stop by the store on your way home?
We need milk,
eggs,
a bunch of spinach,
fresh ginger,
green onions,
maple syrup,
a pound of cous-cous,
mild curry powder,
a pound of coffee,
and a package of seventy five watt light bulbs.
Thanks! See you tonight.

Figure 6.1 Text of a sample message.

This message supplied the answer to the question "What items are you supposed to pick up at the store?"

6.4.3 Message Texts

The texts of the e-mail messages were presented in phrases of varying lengths, with each phrase followed by a pause of about five seconds. Preliminary tests showed that the combined response time of the wizard and the interface was between one and two seconds, and that pauses of less than five seconds were not obviously different from the normal pace of system response. Five seconds is a long response time, uncomfortably so for human-human conversation, so I hoped that this lengthy pause would encourage the subjects to take the initiative in controlling the pace of the interaction.

The messages were divided into segments by hand. The divisions were intended to simulate a phrase-level presentation, although some short phrases were combined to make the presentation less choppy. An example of one message and its division into phrases may be seen in Figure 6.1. The complete text of all messages may be found in Appendix K.

Synthesized speech from the Festival speech synthesizer (Taylor et al., 1998) was used throughout the interface. The message texts were presented in a synthesized male

voice, while the control portions of the interface used a synthesized female voice. Default pronunciations were used except when the default was incorrect; for example, the spelling "read" defaulted to the past-tense pronunciation in all contexts and so the present-tense pronunciation was specified when appropriate. Also, I made minor use of the SABLE markup language (Wouters et al., 1999) to flatten the pitch range at the end of phrases in list items; the intent was to suggest the prosody of list continuation rather than the default sentence-final drop. To improve the understandability, both voices were slowed to 90 percent of the default speaking rate.

6.4.4 Experiment Setup

The subject was seated in a conference room and given the hand-set of a cordless telephone to use. The wizard was in a nearby room with the base of the telephone. By setting the telephone to play through the speaker on the base, the wizard could hear the subject's utterances. The conversation was recorded by setting the microphone of a tape recorder near the base of the phone.

6.5 Measures

The question to be answered is essentially binary: will the subject use acknowledgments in interacting with the program? A subject can show any of several patterns of response across the course of the dialogue:

- The subject may make no attempt to control the pacing of the interface, instead allowing the interaction to proceed through time-outs at the system's default pace.
- The subject may use only commands to control the pacing.
- The subject may use only acknowledgments to control the pacing.
- The subject may use a mixture of commands and acknowledgments.

The hypothesis was that subjects would choose to control the pacing of the interface and that some subjects would use acknowledgments (such as "okay") in preference to commands (such as "next").

The determination as to whether a particular utterance constituted an acknowledgment or a command was based primarily on word choice and dialogue context; this approach is consistent with definitions of this phenomenon, (Chu-Carroll & Brown, 1997). For example, "yes" in the context of a system inform was considered an acknowledgment, but "yes" in the context of a system question was not. The words "okay," "uh-huh," and "yes" (immediately following an inform) were taken as evidence of acknowledgments, and phrases such as "go on," "continue," "next" following an inform were taken as evidence of commands. The interpretation was confirmed during the post-experiment interview by questioning the subjects about their word choice.

A summary of dialogue behaviors by subject may be found in Appendix L. Task measures such as task completion time and numbers of items found also are reported, although no conclusions are drawn about these results.

6.5.1 Post-Experiment Interview

A post-experiment interview was conducted to gather subject feedback and to answer subjects' questions. The experimenter took notes and thus could have introduced bias in the record of responses. No tape recording was made. A transcription of the notes may be found in Appendix M.

The subject was first invited to comment on the interface and the interaction in an open-ended fashion. When the subject had finished, the experimenter asked several specific questions to assess their understanding of the interface functionality. During this portion of the interview, the experimenter reminded the subjects of the words that they had used most

frequently to prompt the system to continue during pauses and asked the subjects to explain why they had selected those words:

- Did you notice that there were two voices, male and female? When was each voice used?
- Did you notice that the program presented information in phrases and then paused? What did you make of that, that is, why do you think the interface was designed that way? Is this a good design, in your opinion?
- I noticed you used <whatever word(s) the subject used most frequently> to tell the program that it should go on when it paused. Why did you pick those words?

Finally, the experimenter explained the true purpose and hypothesis of the experiment, verified that the subject had been unaware that the purported program was a Wizard-of-Oz interface, and asked the subject to comment on the notion of using acknowledgments when interacting with a computer. The responses to this last question, especially, must be assumed to be somewhat optimistic, as it is likely that at least some subjects would be reluctant to disagree with the experimenter.

6.6 Results

As noted in Section 6.3, changes were made in the interface and the experimental setup during the course of the study. For this reason the results for the first six subjects, who used the more rigid interface and experienced the noisier phone lines, are presented separately from those of the last 14 subjects. The first six subjects will be referred to as "Phase 1" and the last 14 as "Phase 2."

The results are summarized in Table 6.1. Because the subject pool was not balanced for gender, results for male and female subjects are reported separately. One of the Phase 1

subjects guessed that the interface was a Wizard-of-Oz interface. His results are not included in the totals below.

Table 6.1 Summary of Acknowledgment Behavior

Behavior	Phase 1 Subjects			Phase 2 Subjects			All Subjects (19)
	Female 3 subjects	Male 2 subjects	Total (5)	Female 10 subjects	Male 4 subjects	Total (14)	
Used acknowledgment or repetition at least once	2 67%	0	2 (40%)	4 (40%)	4 (100%)	8 (57%)	10 (53%)
Used acknowledgment or repetition more than command	0	0	0	3 (30%)	1 (25%)	4 (29%)	4 (21%)
Used acknowledgment but no commands	0	0	0	1 (10%)	0	1 (7%)	1 (5%)
Described acknowledgment to computer as strange	1 (33%)	0	1 (20%)	2 (20%)	0	2 (14%)	3 (16%)

Two of the Phase 1 subjects and eight of the fourteen Phase 2 subjects used an acknowledgment or repetition at least once, and four of the Phase 2 subjects used acknowledgment/repetitions more frequently than they used commands. Only one subject used acknowledgments exclusively, while three Phase 1 subjects and five Phase 2 subjects never used acknowledgments. No subject relied exclusively on time-outs to allow the system to proceed at its own pace, although one Phase 2 subject did use that as her predominant method (42 times, while using acknowledgments only six times and commands three times). Only two subjects used repetition, one each in Phase 1 and in Phase 2, and both reported that they were unaware of having done so.

It is interesting to note that while all of the Phase 2 male subjects exhibited acknowledgment behavior at least once, only one preferred acknowledgment over command. One of the male subjects used acknowledgments only three times, in all cases as prefaces to commands. Conversely, although a lower percentage of women used acknowledgments (40 percent), a higher percentage of them (30 percent) used acknowledgments in preference to commands. The numbers involved are too small to establish a statistical significance between male and female preferences, though.

Table 6.2 Summary of Politeness and Meta-dialogue Behaviors

Behavior	Phase 1 Subjects			Phase 2 Subjects			Total (19)
	Female 3 subjects	Male 2 subjects	Total (5)	Female 10 subjects	Male 4 subjects	Total (14)	
Exhibited politeness at least once ("please," "good-bye")	2 (67%)	0	2 (40%)	7 (70%)	2 (50%)	9 (64%)	11 (58%)
Responded to content	1 (33%)	0	1 (20%)	3 (30%)	0	3 (21%)	4 (21%)
Made meta-comments (such as "ah, there it is!")	1 (33%)	0	1 (20%)	1 (10%)	1 (25%)	2 (14%)	3 (16%)

During the post-experiment interview, three subjects (all female) described the idea of using acknowledgments to the computer as strange and stated that they did not feel that they would do this unless directed to—and even then, they would regard it as simply an alternate command. Two other subjects, both females who had used acknowledgments two to six times during the task, each reported that she had felt silly when she had caught herself saying "please" and "okay" to a computer but had been pleased when it worked. The remainder of the subjects either expressed no strong opinion (two, both female) or expressed a positive attitude toward being able to use acknowledgments when interacting with a computer. Two subjects who had not used acknowledgments commented that they would probably be more likely to use human-like conversation if the synthesized voice were more human-like.

Again, this report of the subjects' attitudes should be interpreted with caution; at this point in the interview they knew the experimenter's hypothesis and so may have been reluctant to express an opinion at odds with the experimental hypothesis.

6.7 Other Dialogue Behaviors

Although I had not formed any hypothesis about other dialogue behaviors, I noticed several interesting dialogue behaviors that I had not anticipated. These results are summarized in Table 6.2.

A number of subjects exhibited politeness behavior toward the interface, either saying "please" when issuing commands to the computer or responding to the program's "good-bye" at the end of the session. One subject used "please" throughout the interaction, but a more common pattern was to use "please" at the beginning of the session and to drop the behavior as the interface became more familiar. Politeness did not seem to be strongly associated with a willingness to use acknowledgments, however; four of the nine subjects who exhibited politeness did not use any acknowledgments in their interaction, and the subject mentioned in the previous section as having used acknowledgments only as prefaces to commands also said "please" eleven times.

Despite the deliberately-limited interface, several subjects responded at least once to the message content as if they were talking to the message sender. In the excerpt shown in Figure 6, for example, the subject replied "thank you" to the message text's "thank you." This did not appear to be a matter of misunderstanding the capabilities of the interface; the

System: I could come to your office now or at any of the following times. one thirty
SUBJECT: continue
System: three o clock
SUBJECT: continue
System: or five fifteen
SUBJECT: continue
System: thank you. I look forward to your prompt reply
SUBJECT: thank you- uh ((laugh)) continue

Figure 6.2 Excerpt of transcript.

In this excerpt, the subject thanks the interface. The system is reading the text of one of the messages.

subject later reported that despite the synthesized voices she had briefly forgotten that she was talking to a computer instead of to her secretary.

One Phase 1 subject and three Phase 2 subjects also made one or more meta-comments, such as "ah, there it is" when finding a particular piece of information. These may have been at least partially an artifact of the "treasure hunt" nature of the task. When questioned in the post-experiment interview, all subjects seemed unaware that they had made these comments. All but one of these instances were followed immediately by a command, so the wizard responded to the command and ignored the meta-comment. The one instance of a stand-alone meta-comment was treated as an unrecognized command (an error message was played).

6.8 Discussion

Subjects were provided with three methods for controlling the pace at which information was presented: silence, command, or acknowledgment/repetition. Over half of the subjects used commands more than they used acknowledgments, but over one half used an acknowledgment or repetition at least once during their interaction and nearly 30 percent used acknowledgments in preference to commands. This occurred despite the fact that subjects were given no reason to think that this behavior would be effective: the interface was deliberately limited in functionality, and voice synthesis was used instead of recorded

voice to emphasize the artificial nature of the interaction. Furthermore, the interface did not offer acknowledgments to the subjects, and the subjects were given no instructions suggesting that the interface understood acknowledgments. In fact two subjects who did use acknowledgments expressed surprise that they had worked, and two who had not used acknowledgments reported that they would probably have used them had they known it would work.

It is interesting to consider these results in light of those reported by Okato et al. (1998). They describe a Japanese-language Wizard-of-Oz study in which the subjects were given some instruction on using the system, and in which the system both presented and accepted back-channel feedback. They found that, even when the interface offered back-channels itself, the rate of subject back-channels was somewhat lower in human-computer interaction than in comparable human-human conversation. This makes the fact that our interface elicited acknowledgments without offering them even more encouraging. Clearly, some people are willing to utilize this human conversational convention in human-computer dialogue. Our post-experiment interviews suggest, however, that some people may find the use of acknowledgements strange or uncomfortable in human-computer interaction. While self-reports of attitudes toward hypothetical situations must be treated with some caution, it seems reasonable to assume that even when such interfaces become available there will be users who will prefer to interact with computers using commands.

Will attitudes and conversational behavior change as people gain experience with more advanced spoken-language interfaces? Despite the relatively short duration of this test—most subjects completed the task in 15-20 minutes—some changes in behavior could be observed over the course of the dialogue. In particular, politeness behaviors were likely to be seen early in the dialogues and then diminish as the subjects became more comfortable with their interaction. It is possible that the use of politeness words did not reflect a strong underlying politeness toward the computer so much as a falling back on human conventions when faced with an unfamiliar dialogue situation. One subject who had used "please" 21 times during the interaction, for example, simply hung up without warning when she had finished. This contrasts, however, with the findings of Nass et al.

(1999) that people do offer socially desirable behavior to computers. Perhaps we as a society are still unsure how we wish to interact with these new tools we have created.

Chapter 7
Conclusions

7.1 Summary

In this dissertation I presented research designed to lay the foundations for incorporating the human grounding and dialogue control mechanism of acknowledgment in spoken language understanding systems. I reported on a three-stage research program in which I probe the characteristics of acknowledgments from three perspectives: Recognizing Acknowledgment, a corpus study; Predicting Acknowledgments, a perceptual study; and Eliciting Acknowledgments, a Wizard-of-Oz study.

The Recognizing Acknowledgments study comprised two corpus experiments in which I asked how acknowledgments might be recognized from low-level prosodic features. Although the first focused on simple pitch measures and showed promising results, a more ambitious experiment incorporating pause, power, speech-act context, and turn length found inconclusive results due to the small corpus size. The corpus studies did show, however, that acknowledgments were more likely to follow lengthy turns.

While a mutuality-of-belief model would argue that the decision to offer an explicit acknowledgment depends on the mental models of the conversants—and thus on the context of the conversation as a whole—I hypothesized that local contextual cues such as turn length or prosodic upturn might serve a "prompting" role that would tend to license an acknowledgment after a particular contribution. In the Predicting Acknowledgments phase of the project, I made use of the perceptual-study paradigm to probe human intuitions of

where acknowledgments should occur. In two experiments, I asked human subjects to judge the appropriateness of an acknowledgment appearing after a given out-of-context turn. Although some subjects did report attending to turn length and prosodic cues, the lack of agreement among the subjects suggests that other factors will be more important in guiding the placement and recognition of acknowledgments in conversation.

In the Eliciting Acknowledgments portion of the project, I probed the issue of users' willingness to use acknowledgment to guide their interacting with a computer. Do people prefer to use commands over acknowledgments to control the pace at which a computer interface presents information? For this study, a Wizard-of-Oz approach was selected to provide robustness while presenting the illusion of a spoken-language interface. When offered an interface that allowed the subject to use either acknowledgment or commands to control the pace at which information was presented, nearly 30 percent of the twenty subjects used acknowledgements in preference to commands. This is particularly impressive in light of the fact that the subjects were given no reason to think that acknowledgment would work at all. In post-experiment interviews, however, some subjects reported that they found the idea of using acknowledgments to a computer to be strange. This results of this study suggest that some people are willing to apply conversational control mechanisms to their computer interfaces, but that others may indeed prefer a more command-oriented interaction.

7.1.1 Limitations of the Studies

Studies of human-human conversational corpora are an important method of investigating conversational interaction that is beyond the capabilities of current spoken

language systems and as such are valuable for establishing initial parameters for the behavior of a spoken-language interface. They have several limitations, however:

- People change their interaction style when they speak to computers;
- Statistically-significant correlations may not reflect the practical utility of the test in question;
- For dialogue-level phenomena such as acknowledgment, it can be impossible to find examples of all combinations of factors about which one wishes to hypothesize.

Perhaps the most severe limitation lies in the differences between human-human and human-computer discourse: people speak differently when they believe they are talking to a computer. Even when the interface is identical, people tend to use shorter, simpler constructions when they believe that they are conversing with a computer (for example, Kennedy et al., 1988), possibly due to well-founded low expectations of the communicative competency of computer interfaces. Furthermore, dialogue designers deliberately reinforce these tendencies in order to guide users toward the limited vocabulary and shorter responses that current speech recognition systems are likely to understand (for example, Hansen et al., 1996a).

The differences between human-human discourse and human-computer discourse are simultaneously a strength and a limitation of corpus studies. With corpus studies we can examine phenomena that do not occur in current human-computer interaction. This leaves the researcher with a chicken-and-egg problem, however. We study corpora of human-human conversation because we wish to improve the communicative competence of our human-computer interfaces, yet we cannot be certain that the results will be applicable to human-computer dialogue precisely because of the lack of communicative competence in our interfaces.

Corpus studies are also inherently limited by the fact that researchers are essentially overhearers to the conversation and thus are at a disadvantage in understanding what took place (Schober & Clark, 1989). While we can observe that certain phenomena occurred, it

may be difficult to determine exactly how important they were to the successful completion of the conversational task. Will an ability to correctly recognize acknowledgments actually result in a quantifiable difference in the utility and acceptability of a human-computer interface?

Another serious limitation of both corpus studies and perceptual studies lies in the difference between dynamic nature of conversation and the static corpus. When the corpus is of a size appropriate for use in a perceptual study, it may contain relatively few alternative approaches to dialogue-level phenomena. It generally is not possible to explore the various factors that might affect the course of a conversation. For example, what might conversants have done at various points in their dialogue had some different communicative action occurred?

The Wizard-of-Oz study addresses some of these limitations in that it allows a dynamic interaction between user and computer. The experimenter is better able to arrange the combination of factors that are of interest, for example. Also, the subjects are (or believe they are) interacting with a computer, so the interaction style will be that seen in human-computer interaction. This is not an unalloyed advantage, however, in that subject expectations about the capabilities of a computational interface may limit the dialogue phenomena that they are willing to employ during the test. If it does not occur to them that a computer "should" be capable of understanding acknowledgments, for example, subjects are unlikely to use them. This expectation limitation hampers our ability to test interaction styles not supported by current interfaces.

7.1.2 Conclusions

My response to the difficulty of studying the mixed-initiative dialogue control mechanism of acknowledgment has been to employ a three-pronged approach: corpus studies, perceptual studies, and Wizard-of-Oz studies. By combining the three approaches, I was able to probe various aspects of the larger issue of understanding how and whether we should incorporate acknowledgments in spoken-language interfaces. Both the corpus

studies and the perceptual studies suggest that local cues may not provide sufficient information to recognize or to predict (or generate) acknowledgment behavior in human-computer interfaces. Clark and Schaefer's model suggest that dialogue-level factors such as the belief states of the participants may be important; post-experimental interviews in the perceptual studies suggest that different people may use different strategies. The Wizard-of-Oz study shows that some subjects are willing to use acknowledgment as a turn-taking mechanism even in a fairly limited interface, although other subjects report resistance to the idea; more study is needed to understand the strength and implications of this resistance.

I view this work as representative of a larger class of problems: how can we understand how and whether to incorporate various human dialogue processes into our spoken-language systems? Arguably we should not attempt to duplicate every human speaking trait; an interface that peppers its contributions with filled pauses, for example, is likely to annoy. At the same time, we would like to make our interfaces easier and more transparent for the user, and it seems reasonable to look to human dialogue for ways to accomplish this. The problem is one faced at times by nearly every spoken language interface researcher: we cannot test our theories because they depend on an infrastructure that does not yet exist and that, we believe, depends on such theories. That infrastructure is more than a matter of programming; it is a matter of expectation and experience on the part of the user. In this work I illustrate an approach to answering such questions in the absence of such an infrastructure.

7.2 Future Work

Several open issues remain. Although Ward (2000) and Aist (1998) report promising results in predicting and generating acknowledgments with rules based on local cues such as turn length and prosody, the corpus and perceptual studies reported here suggest that local factors may not provide an adequate basis for a complete model of acknowledgment behavior.The collaborative view of conversation articulated by Clark and his colleagues would argue that the larger conversational context, particularly the

conversants' models of the shared understanding established in the dialogue, should play an important in determining where acknowledgments should appear. Future work should investigate the utility of incorporating larger dialogue context and conversant belief into the acknowledgment model.

Next steps also should include building upon the Wizard-of-Oz study described in Chapter 6 to better understand subjects' preferences in interaction style. Extending the study to a larger and gender-balanced group of subjects would allow firmer quantitative conclusions to be drawn about the percentage and profiles of people who chose the acknowledgment style of interaction over the command style. In particular, we cannot conclude from the current study's small sample how strong the preference for using or avoiding acknowledgment might be, especially among male subjects. Other experiments might include a comparison of how the use of recorded voice instead of synthesized voice affects the choice of interaction style.

Building on this research, future work might focus on comparing the usefulness and user acceptability of a spoken language interface with and without acknowledgment behavior. The interface used in the Eliciting Acknowledgments experiment was carefully designed to offer no functional advantages for using acknowledgments over using commands; using a mixed-initiative interface and a more complex task such as negotiating an appointment time, for example, we might ask whether a spoken language interface which accepts and responds to acknowledgments might offer advantages over one that does not in terms of standard metrics for assessing the effectiveness and usability of an interface such as task completion success and time and user satisfaction.

Bibliography

Aist, G. (1998). "Expanding a Time-Sensitive Conversational Architecture for Turn-Taking to Handle Content-Driven Interruption." In *Proceedings of ICSLP 98 Fifth International Conference on Spoken Language Processing*, 412-417.

Allen, J. and M. Core (1997). "Draft of DAMSL: Dialog Act Markup in Several Layers," October 18, 1997.

Argyle, M. and M. Cook (1976). *Gaze and Mutual Gaze*. Cambridge: Cambridge University Press.

Austin, J. L. (1962). *How to Do Things with Words*, London:Oxford University Press.

Bakeman, R. and J. M. Gottman (1997). *Observing Interaction: An Introduction to Sequential Analysis*, second edition. Cambridge: Cambridge University Press.

Basson, S., S. Springer, C. Fong, H. Leung, E. Man, M. Olson, J. Pitrelli, R. Singh and S. Wong (1996). "User Participation and Compliance in Speech Automated Telecommunications Applications," In *Proceedings of ICSLP 96 Fourth International Conference on Spoken Language Processing*, 1676-1679.

Brennan, S. E. (1991). "Conversation With and Through Computers." *User Modeling and User-Adapted Interaction*. **1**:67-86.

Cahn, J. (1992). "An Investigation into the Correlation of Cue Phrases, Unfilled Pauses and the Structuring of Spoken Discourse." In *Proceedings of the IRCS Workshop on Prosody in Natural Speech*, University of Pennsylvania, 1992, 19-30.

Carletta, J. C. (1996). "Assessing Agreement on Classification Tasks: The Kappa Statistic." *Computational Linguistics*, **22**(2):249-254.

Chafe, W. (1985). "Some Reasons for Hesitating." In *Perspectives on Silence*, D. Tannen and M. Saville-Troike (eds.), 77-89. Norwood, N.J.:Ablex.

Chafe, W. (1986). "Cognitive Constraints on Information Flow." In *Coherence and Grounding in Discourse*, R. Tomlin (ed.), 21-52. Amsterdam:J. Benjamins.

Chafe, W. (1993). "Prosodic and Functional Units of Language." In *Talking Data: Transcription and Coding in Discourse Research*, Jane A. Edwards and Martin D. Lampert (eds.), 33-43, Hillsdale, NJ: Lawrence Erlbaum Associates.

Chu-Carroll, J. and M. K. Brown (1997). "Tracking Initiative in Collaborative Dialogue Interactions." In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 262-270.

Clark, H. H. and C. R. Marshall (1981). "Definite Reference and Mutual Knowledge." In *Elements of Discourse Understanding*, A. K. Joshi, B. Webber, and I. A. Sag (eds.), 10-63. Cambridge:Cambridge University Press.

Clark, H. H. and D. Wilkes-Gibbs (1986). "Referring as a Collaborative Process." *Cognition*, **13**:259-294.

Clark, H. H. and E. F. Schaefer (1989). "Contributing to Discourse." *Cognitive Science*, **13**:259-294.

Cohen, P. R. (1984). "The Pragmatics of Referring and the Modality of Communication." *Computational Linguistics*, **10**(2):97-146.

Cole, R. A., and L. Hirschman (1992). "Workshop on Spoken Language Understanding," Oregon Graduate Institute Technical Report No. CS/E 92-014.

Cole, R. A., D. G. Novick, P. J. E. Vermeulen, S. Sutton, M. Fanty, L. Wessels, J. de Villiers, J. Schalkwyk, B. Hansen, B. and D. Burnett (1997). "Experiments with a Spoken Dialogue System for Taking the U.S. Census." *Free Speech Journal*, **1**:3.

Dowding, J., J. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore and D. Moran (1993). "Gemini: A Natural Language System for Spoken-Language Understanding." In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, 54-61.

Dragon Systems, Inc. (2001). Dragon NaturallySpeaking Preferred (Version 5) Key Features. [Online], Undated. Available: http://www.lhsl.com/naturallyspeaking/pref/features.asp, [Viewed June 3, 2001].

Dybkjaer, L., N. O. Bernsen and H. Dybkjaer (1996). "Reducing Miscommunication in Spoken Human-Machine Dialogue." In *Proceedings of the AAAI-96 Workshop on Detecting, Repairing, and Preventing Human-Machine Miscommunication*, August 4, 1996, Portland, OR, 29-36.

Eiselt, K., K. Mahesh and J. Holbrook (1993). "Having Your Cake and Eating It Too: Autonomy and Interaction in a Model of Sentence Processing." In *Proceedings of the Eleventh National Conference on Artificial Intelligence (AAAI'93)*, 380-385.

Ferreira, F. (1993). "Creation of Prosody During Sentence Production," *Psychological Review*, **100**:2, 233-253.

Goddeau, D. (1992). "Using Probabilistic Shift-Reduce Parsing in Speech Recognition Systems." In *Proceedings of ICSLP 92 International Conference on Spoken Language Processing*, 321-324.

Goldman-Eisler, F. (1961). "A Comparative Study of Two Hesitation Phenomena" *Language and Speech*, **4**:18-26.

Goldman-Eisler, F. (1958). "The Predictability of Words in Context and the Length of Pauses in Speech." *Language and Speech*, **1**:226-2331.

Goodine, D., L. Hirschman, J. Polifroni, S. Seneff and V. Zue (1992). "Evaluating Interactive Spoken Language Systems," In *Proceedings of ICSLP 92 International Conference on Spoken Language Processing*, 197-200.

Grigoriu, A., J. P. Vonwiller and R. W. King (1994). "An Automatic Intonation Tone Contour Labelling and Classification Algorithm," In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '94)*, II-181-II-184.

Gross, D., J. F. Allen and D. R. Traum (1993). "The TRAINS 91 Dialogues." Technical Reports TN92-1, Department of Computer Science, University of Rochester.

Grosz, B. J. and C. L Sidner (1990). "Plans for Discourse." In *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack (eds.), 417-444, MIT Press.

Haller, S., S. McRoy and A. Kobsa (eds.) (1999). *Computational Models of Mixed-Initiative Interaction*. Dordrecht, The Netherlands:Kluwer Academic Publishers.

Hansen, B., D. G. Novick and S. Sutton (1996a). "Systematic Design of Spoken Prompts." In *Conference on Human Factors in Computing Systems (CHI'96)*, Vancouver, BC, April, 1996, 157-164.

Hansen, B., D. G. Novick and S. Sutton (1996b). "Prevention and Repair of Breakdowns in a Simple Task Domain." In *Proceedings of the AAAI-96 Workshop on Detecting, Repairing, and Preventing Human-Machine Miscommunication*, August 4, 1996, Portland, OR, 5-12.

Heeman, P. A., D. Byron and J. F. Allen (1998). "Identifying Discourse Markers in Spoken Dialog." In *AAAI Spring Symposium on Applying Machine Learning and Discourse Processing*, Stanford, March 1998, 44-51.

Heeman, P. A., M. Johnston, J. Denney and E. Kaiser (1998). "Beyond Structured Dialogues: Factoring Out Grounding." In *Proceedings of ICSLP 98 Fifth International Conference on Spoken Language Processing*, 863-867.

Hirschman, L. (1992). "Multi-Site Data Collection for a Spoken Language Corpus." *Proceedings of ICSLP 92 International Conference on Spoken Language Processing*, 903-906.

Howells, T., D. Friedman and M. Fanty (1992). "Broca, An Integrated Parser for Spoken Language." *Proceedings of ICSLP 92 International Conference on Spoken Language Processing*, 325-328.

Hunt, A. (1994). "A Generalised Model for Utilising Prosodic Information in Continuous Speech Recognition." *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '94)*, II-181-II-184.

Issar. S., and W. Ward (1993). "CMU's Robust Spoken Language Understanding System." In *Eurospeech '93*, 2147-2150.

Iwase, T., and N. Ward (1998). "Pacing Spoken Directions to Suit the Listener." In *Proceedings of ICSLP 98 Fifth International Conference on Spoken Language Processing*, 1203-1207.

Kendon, A. (1978). "Looking in Conversations and the Regulation of Turns at Talk: A Comment on the Papers of G. Beattie and D. R. Rutter et al." *British Journal of Social and Clinical Psychology*, **17**:23-24.

Kennedy, A., A. Wilkes, L. Elder and W. Murray (1988). "Dialogue with Machines." *Cognition*, **30**:37-72.

Kita, K., Y. Fukui, M. Jagata and T. Morimoto (1996). "Automatic Acquisition of Probabilistic Dialogue Models." In *Proceedings of ICSLP 96 Fourth International Conference on Spoken Language Processing*, 196-199.

Krueger, G. and A. Chapanis (1980). "Conferencing and Teleconferencing in Three Communication Modes as a Function of the Number of Conferees." *Ergonomics*, **23**:2, 103-122.

Kai-Fu Lee, Hsiao-Wuen Hon and Raj Reddy (1990). "An Overview of the SPHINX Speech Recognition System." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **38**:1, 35-45.

Johnstone, A., U. Berry, T. Nguyen and A. Asper (1995). "There was a Long Pause: Influencing Turn-Taking Behaviour in Human-Human and Human-Computer Spoken Dialogues." *International Jounral of Human-Computer Studies*, **42**:83-411.

Litman, D. J. (1996). "Cue Phrase Classification Using Machine Learning." *Journal of Artificial Intelligence Research*, **5**:53-94.

McCandless, T. (1992). "The Use of Blackboard Architectures in User Interface Design," U S WEST Science and Technology Research Report ST 04-01.

McGee, D. R., P. R. Cohen and S. Oviatt, 1998, "Confirmation in Multimodal Systems," *Proceedings of the International Joint Conferenc of the Association for Computational Linguistics and the International Committee on Computational Llinguistics (COLING-ACL '98)*, Association for Computational Linguistics Press: Montreal, Quebec, Canada, 823-829.

Nakajima, S. and J. Allen (1993). "A Study on Prosody and Discourse Structure in Cooperative Dialogues," Rochester Tech Report No TRAINS-TN93-2, Sept. 1993.

Nakatani, C. and J. Hirschberg (1993). "A Speech-First Model for Repair Detection and Correction." In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, 46-53.

Nakatani, C. H. (1997). *The Computational Processing of Intonational Prominence: A Functional Prosody Perspective*, dissertation, Harvard University, Cambridge, Massachusetts, May 1997.

Nakazato, Shu (2000). "Japanese Dialogue Corpus of Multi-Level Annotation." In "Proceedings of the First SIGdial Workshop on Discourse and Dialogue" [Online], October 7-8, 2000. Available: http://www.sigdial.org/sigdialworkshop/proceedings/nakazatu.pdf, [Viewed June 3, 2001].

Nass, C., Y. Moon and P. Carney (1999). "Are Respondents Polite to Computers? Social Desirability and Direct Responses to Computers." *Journal of Applied Social Psychology*, **29**:5, 1093-1110.

Noguchi, H. and Y. Den (1998). "Prosody-Based Detection of the Context of Backchannel Responses." In *Proceedings of ICSLP 98 Fifth International Conference on Spoken Language Processing*, 487-490.

Novick, D. G. (1990). "Modeling Belief and Action in a Multi-agent System." In *AI, Simulation and Planning in High Autonomy Systems*, B. Zeigler and J. Rozenblit (eds.), Los Alamitos, CA: IEEE Computer Society Press, 34-41.

Novick, D. G. and S. Sutton (1994). "An Empirical Model of Acknowledgment for Spoken-Language Systems," in *Proceedings of the 32nd Annual meeting of the Association for Computational Linguistics*, 96-101.

Novick, D. G., K. Ward and B. Corliss (1995). "The effect of context on the intelligibility of dialogue." In *Proceedings of the Fourth European Conference on Speech Communication and Technology (EuroSpeech'95)*, Sept.18-21, Madrid, Spain, 1995, 1235-1238.

Novick, D. G., B. Hansen, K. D. Rubesh and K. Ward (1996a). "Coordinating Turn-Taking with Gaze." In *Proceedings of ICSLP 96 Fourth International Conference on Spoken Language Processing*, Philadelphia, PA, October, 1996, 1888-91.

Novick, D. G., L. Walton, and K. Ward (1996b). "Contribution Graphs in Multiparty Discourse." In *Proceedings of the International Symposium on Spoken Dialogue (ISSD-96)*, Philadelphia, PA, October, 1996, 53-56.

Okato, Y., K. Kato, M. Yamamoto and S. Itahashi (1998). "System-User Interaction and Response Strategy in Spoken Dialogue System." *Proceedings of ICSLP 98 Fifth International Conference on Spoken Language Processing*, 495-498.

O'Shaughnessy, D.(1993). "Locating Disfluencies in Spontaneous Speech: An Acoustical Analysis," *Eurospeech'93*, 2187-2190.

Oviatt, S. L. and P. R. Cohen (1988). "Discourse Structure and Performance Efficiency in Interactive and Noninteractive Spoken Modalities." Technical Note 454, SRI International.

Oviatt, S. (1992). "Pen/Voice: Complementary Multimodal Communication." In *Proceedings of Speech Tech'92*, New York, February 1992, 238-241.

Oviatt, S. L. and P. R. Cohen (1992). "Spoken Language in Interpreted Telephone Dialogues." *Computer Speech and Language*, **6**(3):277-302.

Oviatt, S. L., P. R. Cohen and M. Wang (1994). "Toward Interface Design for Human Language Technology: Modality and Structure as Determinants of Linguistic Complexity." *Speech Communication*, **15**(3-4):283-300.

Oviatt, S. L. and R. VanGent (1996). "Error Resolution During Multimodal Human-Computer Interaction." In *Proceedings of ICSLP 96 Fourth International Conference on Spoken Language Processing,* 204-207.

Pierrehumbert, J., and J. Hirschberg (1990). "The Meaning of Intonational Contours in the Interpretation of Discourse." In *Intentions in Communication*, P. Cohen, J. Morgan, & M. Pollack (eds.), Chapter 14, 271-311, Cambridge, MS:MIT Press.

Price, P., M. Ostendorf, S. Shattuck-Hufnagel and C. Fong (1991). "The Use of Prosody in Syntactic Disambiguation." In *Proceedings of the Fourth DARPA Workshop on Speech and Natural Language*, ed. Patti Price, looseleaf, unpaged.

Pruitt. D. G. and S. A. Lewis (1975). "Development of Integrative Solutions in Bilateral Negotiation." *Journal of Personality and Social Psychology*, **31**(4):621-633.

Sacks, H., E. Schegloff and G. Jefferson (1974). "A Simplest Systematics for the Organization of Turn-Taking in Conversation." *Language*, **50**:696-735.

Schegloff, E. A. (1982). "Discourse as an Interactional Achievement: Some Uses of 'uh huh' and Other Things that Come Between Sentences." In *Analyzing Discourse: Text and Talk*, D. Tannen (ed.), 71-93. Wahington, D.C.:Georgetown University Press.

Schober, M. and H. H. Clark (1989). "Understanding by Addressees and Overhearers." *Cognitive Psychology* **21**:211-232.

Scon, J. (1993). "Guiding an HPSG Parser Using Semantic and Pragmatic Expectations." In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, 295-297.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*, Cambridge:Cambridge University Press.

Searle, J. R. (1975). "Indirect Speech Acts." In *Syntax and Semantics, Volume 3: Speech Acts*, J. L. Morgan (ed.), New York:Academic Press, 59-82.

Searle, J. R. and D. Vanderveken (1985). *Foundations of Illocutionary Logic,* Cambridge: Cambridge University Press.

Seneff, S. (1992). "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, **18**(1):61-86.

Seneff, S., H. Meng and V. Zue (1992). "Language Modelling for Recognition and Understanding Using Layered Bigrams." In *Proceedings of ICSLP 92 International Conference on Spoken Language Processing*, 317-320.

Siegel, S. and N. J. Castellan, Jr. (1988), *Nonparametric Statistics for the Behavioral Sciences, Second Edition*. Singapore:McGraw-Hill.

Stubbs, M. (1983). *Discourse Analysis*. Chicago:Chicago Press.

Sutton, S., R. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki, P. Hosom, A. Kain, J. Wouters, D. Massaro and M.l Cohen (1998). "Universal Speech Tools: the CSLU Toolkit." In *Proceedings of the International Conference on Spoken Language Processing*, 3221-3224.

Swerts, M., H. Koiso, A. Shimojima and Y. Katagiri (1998). "On Different Functions of Repetitive Utterances." In *Proceedings of ICSLP 98 Fifth International Conference on Spoken Language Processing*, 483-487.

Taylor, P., A. W. Black and R. Caley (1998). "The Architecture of the Festival Speech Synthesis System." In the *Third ESCA/COCOSDA Workshop on Speech Synthesis*, 147-151.

Traum, D. R. (1991). "Towards a Computational Theory of Grounding in Natural Language Conversation." Technical Report 401, The University of Rochester, Computer Science Department.

Traum, D. R. (1996). "Coding Schemes for Spoken Dialogue Structure," unpublished manuscript, Department of Computer Science, University of Maryland, May 1996.

Traum, D. R. (1999a). "20 Questions on Dialogue Act Taxonomies." [Online] Invited paper presented at the Amstelogue '99 Workshop on the Semantics and Pragmatics of Dialogue, May 1999. Available: http://www.ict.usc.edu/~traum/Papers/amstel.ps, [Viewed June 3, 2001].

Traum, D. R. (1999b). "Computational Models of Grounding in Collaborative Systems,." In *Working Notes of AAAI Fall Symposium on Psychological Models of Communication*, 124-131, November, 1999.

Traum, D. R. and P. A. Heeman (1996). "Utterance Units and Grounding in Spoken Dialogue." In *Proceedings of ICSLP 96 Fourth International Conference on Spoken Language Processing*, 1884-1887.

van Vuuren, S. H. J. (1992). "Pitch Detection." Technical Report, Department of Electrical and Electronic Engineering, University of Pretoria, October 1992.

Waibel, A. (1988). *Prosody and Speech Recognition*, San Mateo, CA:Morgan Kaufman.

Waibel, A. and K-F. Lee (1990). *Readings in Speech Recognition*. San Mateo, CA:Morgan Kaufman Publishers, Inc.

Walker, M. A. (1993). *Informational Redundancy and Resource Bounds in Dialogue*. Doctoral dissertation, University of Pennsylvania.

Wang, M. Q. and J. Hirschberg. "Automatic Classification of Intonational Phrase Boundaries," *Computer Speech and Language*, **6**:175-196.

Ward, K. (1992). *A Speech Act Model of Air Traffic Control Dialogue*. Thesis, Oregon Graduate Institute of Science & Technology, 1992.

Ward, K. and D. G. Novick (1994). "On the Need for a Theory of Integration of Knowledge Sources for Spoken Language Understanding," *Proceedings of the AAAI-94 Workshop on the Integration of Natural Language and Speech Processing*, July, 1994, 23-30.

Ward, N. (1996). "Using Prosodic Clues to Decide When to Produce Back-Channel Utterances," In *Proceedings of ICSLP 96 Fourth International Conference on Spoken Language Processing*, 1724-1727.

Ward, N. (1997). "Responsiveness in Dialog and Priorities for Language Research," *Systems and Cybernetics*, Special Issue on Embodied Artificial Intelligence and Artificial Life, **28**:521-533, 1997.

Ward, N. (2000). "Prosodic Features which Cue Back-Channel Responses in English and Japanese." *Journal of Pragmatics*, **23**:1177-1207.

Wierzbicka, A. (1987). *English Speech Act Verbs: A Semantic Dictionary*. Sydney, Australia:Academic Press.

Wouters, J., B. Rundle and M. W. Macon (1999). "Authoring Tools for Speech Synthesis using the Sable Markup Standard." In *Proceedings of Eurospeech '99*, Budapest, Hungary, September 1999, 963-966.

Appendix A

Corpus Preparation and Coding

A.1 Pre-Processing

Conversations SW2800, SW2987, SW1026, and SW2710 from the Training directory on NIST Speech Disc 8-1.1 were used for this study. These were selected because they exhibited less disruptive line noise than some of the others and because they seemed to have fewer transcription errors.

The Switchboard distribution includes word-level alignment information, but examination indicated that these alignments were often in error. In particular, back-channel utterances were in some cases so badly mis-aligned that the tags completely missed the speech signal. The first step, therefore, was to align the transcriptions using the OGI Speech Toolkit (Sutton et al., 1998). The transcriptions were checked against the recordings, corrected when errors were found, then force-aligned against hand-extracted turns using the OGI Speech Toolkit.

A.1.1 Speech Act and Grounding Act Coding

As proposed in the DAMSL (Dialog Act Markup in Several Layers) annotation method (Allen & Core, 1997), the speech acts (from a contribution standpoint) and the grounding acts (acceptance) are coded in separate passes through the data. The procedures for each were similar, though, and are described in this section.

The coding interfaces (A.1) were implemented in Tk using the OGI Speech Toolkit (Sutton et al., 1998). The coders were able to select either headphones or the workstation speakers for playing the speech samples, and they could control the volume level (these controls aren't shown in these illustrations). The coder clicked a button to play the next turn. The text of a turn was displayed only after the speech file was played; this was done to encourage the coder to base the judgement on the speech and not the text while still providing the text for clarification. By clicking on the text of the turn, the sample could be replayed as often as the coder wished. The sample included 500 milliseconds of the recording preceding the utterance. This was done to provide context and to allow the coder to make use of cues from the end of the previous utterance and the intervening pause (if any).

At first, coders were asked to identify which turn contained the contribution that was being accepted. As discussed in 2, a contribution may explicitly or implicitly accept, or ground, the previous contribution in addition to (optionally) making a new contribution to the conversation. While one might argue that a single speaker's turn may embody

multiple speech acts, or even multiple levels of speech act, for this study I was interested only in the lowest-level illocutionary act that immediately precedes the other's turn. I had hoped in this way to identify structure, as my preliminary hypotheses had included a prediction that accepts of other than the previous might be marked prosodically. Preliminary analysis, however, suggested that the structure of contributions and acceptances was relatively flat in this corpus. Accordingly, this hypothesis was not tested and the coding was simplified to assume that the grounding acts accepted the last utterance in the previous turn.

A.1.1.1 Speech-act coding

The interface for coding the speech acts is shown in Figure A.1. Table A.1 shows the list of acts from which the coders could select. As discussed in 2, a turn may comprise multiple speech acts; for this study, coders were asked to identify only the final speech act in the turn. They were not asked to identify where the speech act began, nor were they asked to identify other speech acts that might have been present.

Fifteen categories are arguably too many for reliable coding (Pruitt and Lewis, 1975). In fact, coders made use of only 8 speech-act categories (Check, Continue,

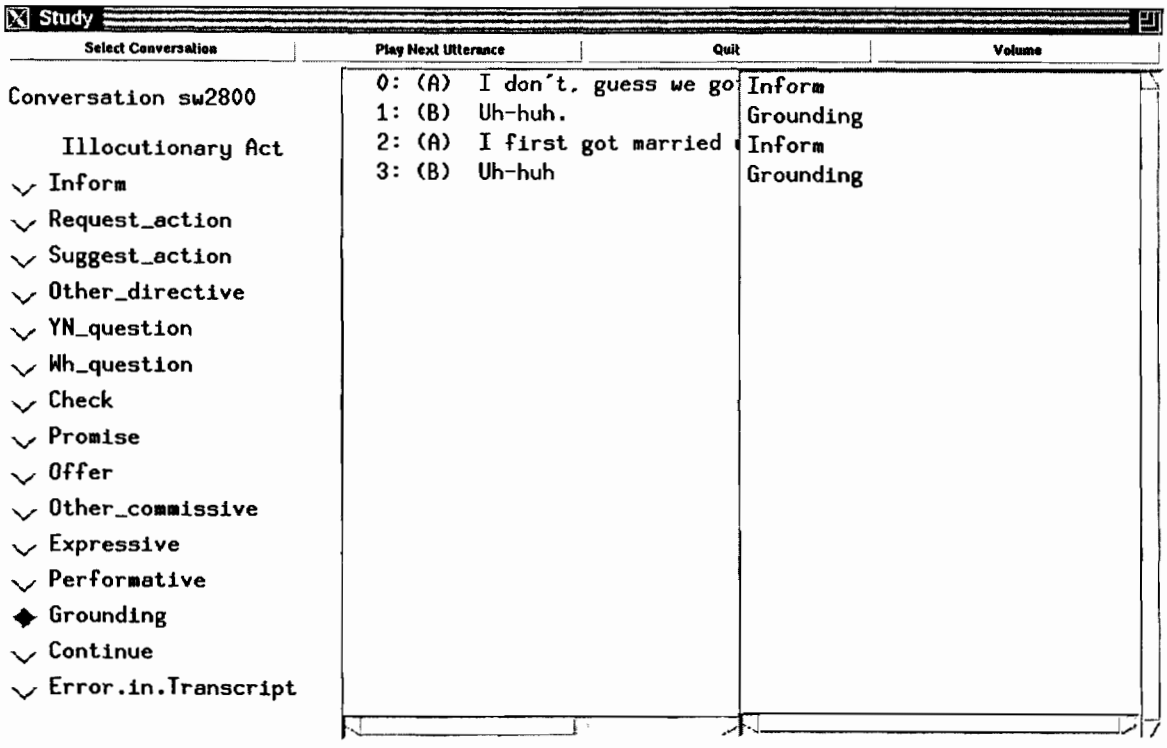

Figure A.1 Interface for Coding Speech Acts

Grounding, Expressive, Inform, Performative, Wh-Question, YN-Question) plus Error in coding this corpus.

Table A.1 Illocutionary Acts Used in Speech Act Coding

Illocutionary Acts	Definition
Inform	The speaker provides new information (including providing requested information when answering a question)
Request action	The speaker attempts to get the hearer to perform some action
Suggest action	The speaker proposes a new action
Other directive	
YN question	The speaker asks a yes-no-question, trying to determine the polarity of a proposition
Wh question	The speakers asks a wh-question, trying to determine the value of some term in a proposition
Check	The speaker attempts to verify that a certain proposition is true
Promise	The speaker commits to a future action
Offer	The speaker proposes to commit to a future action
Other Commissive	
Expressive	The speaker reacts to a previous contribution
Performative	The speaker performs an action by what is said, such as making a bet
Grounding	The speaker offers an explicit grounding act, such as an acknowledgment
Continue	The turn does not contain a complete illocutionary act
Error in Transcript	The transcript does not match the sound file

A.1.1.2 Grounding Act Coding

The interface for coding the grounding acts was similar to that used for coding the illocutionary acts. As discussed in Chapter 4, the acts used were Not Heard, Miscommunication, Acknowledgment, Assessment, Repetition, Next Contribution, Continue, and Error in Transcript.

As discussed in 2, a contribution may explicitly or implicitly accept, or ground, the previous contribution in addition to (optionally) making a new contribution to the conversation. While a single speaker's turn may embody multiple speech acts, or even multiple levels of communicative act, for this study I was interested only in the lowest-level illocutionary act that immediately precedes the other's turn.

A.1.2 Inter-rater Reliability

All samples used in this study were coded by the author. Approximately half were coded by a second person and inter-rater reliability was assessed on those samples. Most categories were coded with good agreement. The major source of disagreement between the raters was between the categories of acknowledgment and assessment. As discussed in Chapter 2, acknowledgments and assessments are similar in definition and some researchers do not distinguish between them. Continues hard to detect with original presentation method

Table A.2 Inter-rater Reliability on Grounding Act Coding

	Coder 1					
Coder 2	Ackn.	Assess.	Miscomm.	Next Cont.	Not Heard	Continue
Ackn.	22	10		4		
Assess.		26		2		
Miscomm.			2	2		
Next Cont.		2		68	2	
Not.Heard					4	
Continue						50

Total: 194
Percent agreement: 172 /194 = 89%
kappa: 0.85

Appendix B

Predicting Acknowledgments Study

Forms

This appendix contains the subject forms used in the Predicting Acknowledgment Likelihood and the Predicting Acknowledgment Occurrence studies. These are:

- Subject Agreement and Release
- Subject Profile Information
- Instructions for Predicting Acknowledgment Likelihood study
- Instructions for Predicting Acknowledgment Occurrence study
- Post-experiment Interview

Forms for both studies were the same, with the exception of the subject instructions.

Note: at the time that these studies were run, they were provisionally named Acknowledgments Study and Acknowledgment Study B respectively, and the forms reflect the provisional names.

Acknowledgments Study B
Agreement and Release

Description of Study

This study is under the direction of Dr. David Novick. Subjects will listen to excerpts from a recorded conversation and will judge whether an acknowledgment occurred after that excerpt. The task is expected to require no more than 30 minutes to complete. Subjects will be paid $10.00 for their participation.

Subject Statement and Signature

I understand and agree that:

1. I may end my participation at any time for any reason. I will be paid $10.00 for my participation whether or not I complete the task.

2. The experimental results are confidential; my name will not be associated with the experiment or with the profile information I provide. The results of the study may be used by members of the Laboratory and other persons designated by them for reasonable education, scientific, and technical purposes.

Signature: _____

Name: _____

Address: _____

_____

Telephone: _____

Date: _____

I have received $10.00 for my participation in this experiment:_____ (initial)

Acknowledgments Study B
Profile Information: Subject __

We need this information to interpret the results of the data you provide to us. Your profile will not be associated with your name.

Subject profile for: ____

Age: _____

Gender: _____.

In what part of the country did you spend most of your childhood?

_____

In what parts of the world have you lived for more than 5 years?

_____

_____

Have you studied (formally or informally) linguistics, spoken language understanding, or natural language processing? If so, what were your interests within those fields?

Acknowledgments Study
Instructions

An *acknowledgment* is a statement designed to let the other person know that you understand what they just said, but without expressing an opinion or answering a question. Acknowledgments often occur as a quiet little "uh-huh" or "right" in the background, but they can also be more explicit (e.g., "go on").

We are interested in understanding how people decide when it is appropriate to use acknowledgments in conversations. In this study you will listen to extracts of conversations in which two people chat about credit cards. We would like your opinion as to how likely it is that the other person will respond with an acknowledgment. In other words, if you were having a conversation with someone and the other person said this "sentence" to you, do you think it likely that you would respond with an acknowledgment?

A few comments:

- Note that we aren't asking whether the extract you listen to is an acknowledgment; instead, we're asking whether you think that the next thing that will happen might be an acknowledgment.
- Acknowledgments are often expressed using the same words as those used to answer yes/no questions (e.g., "yes," "uh-huh"). Answers to questions aren't usually acknowledgments, though, so "Are you listening to me?" "Uh-huh." is not an example of an acknowledgment.
- These extracts are drawn from nine different conversations, and they are presented in random order. Consecutive utterances don't necessarily make sense.
- Because the extracts are drawn from several different conversations, the volume levels and recording quality vary greatly. Feel free to adjust the volume as needed.
- You may play each extract as many times as you like, and you may change your mind about the likelihood category as many times as you like until you play the next extract.
- If you discover that you've made a mistake after you've gone on to the next extract, please make a note of utterance number and the correct answer on a piece of paper. I'll fix it later by hand.

Acknowledgments Study B
Instructions

An *acknowledgment* is a statement designed to let the other person know that you understand what they just said, but without expressing an opinion or answering a question. Acknowledgments often occur as a quiet little "uh-huh" or "right" in the background, but they can also be more explicit (e.g., "go on").

We are interested in understanding how people decide when it is appropriate to use acknowledgments in conversations. In this study you will listen to extracts of conversations in which two people chat about credit cards. We would like you to tell us whether an acknowledgment occurred immediately after that in the original conversation. In other words, in the original conversation one person said the sentence you will hear. Do you think that the other person responded to that sentence with an "uh-huh?"

A few comments:

- Note that we aren't asking whether the extract you hear is an acknowledgment; instead, we're asking whether you think that the next thing that happened was an acknowledgment.
- Acknowledgments are often expressed using the same words as those used to answer yes/no questions (e.g., "yes," "uh-huh"). Answers to questions aren't usually acknowledgments, though, so "Are you listening to me?" "Uh-huh." is not an example of an acknowledgment.
- These samples are drawn from nine different conversations, and they are presented in random order. Consecutive utterances don't necessarily make sense.
- Because the extract are drawn from several different conversations, the volume levels and recording quality vary greatly. Feel free to adjust the volume as needed.
- You may play each extract as many times as you like, and you may change your mind about your answer as many times as you like until you play the next extract.
- If you discover that you've made a mistake after you've gone on to the next extract, please make a note of the utterance number and the correct answer on a piece of paper. I'll fix it later by hand.

Acknowledgments Study B
Post-experiment Interview: Subject __

As you worked on the task, did you find yourself making rules about how to decide your answer? Can you articulate any of these rules?

Do you have any suggestions?

Do you have any questions about the experiment or the study?

Appendix C

Predicting Acknowledgment Likelihood Study

Subject Profiles

Table C.1 Subject Profile Information for Predicting Acknowledgment Likelihood

Subj	Age	Gender	Childhood Location	Other Location	Studied Linguistics
S1	23	Female	Washington (state)	Canada (5 yrs.), New York (3-4 yrs.)	No
S2	24	Female	Oregon		No
S3	43	Male	Oregon	USA	Ph.D. student in speech recognition
S4	33	Male	Oregon		No
S5	12[a]	Female	Oregon		No
S6	25	Male	Pennsylvania (Pittsburgh)	Oregon	No

a. Because of the age of this subject, I checked the study results with and without her data. As can be seen from the kappa calculations in Appendix D, her responses were not noticeably different than those of the other subjects and excluding this subject's data would not have caused me to change any of my conclusions.

Appendix D

Predicting Acknowledgment Likelihood Study

Results

Detailed results of the Predicting Acknowledgment Likelihood are documented in this section. These consist of tables showing, for each pair of subjects, their agreement on the coding task and the kappa statistic for that pair.

Table D.1 Agreement between Subjects S6 and S5

		Subject S5				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S6	Impossible	1	8	4	0	0
	Very Unlikely	7	30	18	7	3
	Mod. Likely	0	6	18	18	0
	Very Likely	0	2	12	13	3
	Certain	0	0	0	0	0

Total samples: 150
S6/S5 percent agreement: 62 / 150 = 41%
S6/S5 weighted kappa = 0.32

Table D.2 Agreement between Subjects S6 and S4

		Subject S4				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S6	Impossible	2	2	2	2	5
	Very Unlikely	19	25	13	5	3
	Mod. Likely	4	16	19	3	0
	Very Likely	0	6	17	6	1
	Certain	0	0	0	0	0

Total Samples: 150
S6/S3 percent agreement: 68 / 150 = 45%
S6/S3 weighted kappa: 0.14

Table D.3 Agreement between Subjects S6 and S3

		Subject S3				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S6	Impossible	9	1	3	0	0
	Very Unlikely	13	21	20	10	1
	Mod. Likely	1	3	18	15	5
	Very Likely	1	0	9	20	0
	Certain	0	0	0	0	0

Total samples: 150
S6/S3 percent agreement: 68 / 150 = 45%
S6/S3 weighted kappa: 0.40

Table D.4 Agreement between Subjects S6 and S2

		Subject S2				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S6	Impossible	6	6	1	0	0
	Very Unlikely	3	41	6	15	0
	Mod. Likely	0	7	9	25	1
	Very Likely	1	1	8	19	1
	Certain	0	0	0	0	0

Total samples: 150
S6/S2 percent agreement: 75 / 150 = 50%
S6/S2 weighted kappa: 0.43

Table D.5 Agreement between Subjects S6 and S1

		Subject S1				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S6	Impossible	8	1	0	0	0
	Very Unlikely	27	12	12	6	0
	Mod. Likely	4	6	23	6	0
	Very Likely	2	2	10	6	0
	Certain	0	0	0	0	0

Total samples: 125
S6/S1 percent agreement: 49 / 125 = 39%
S6/S1 weighted kappa: 0.34

Table D.6 Agreement between Subjects S5 and S4

		Subject S4				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S5	Impossible	4	3	0	1	0
	Very Unlikely	19	14	6	1	6
	Mod. Likely	1	18	22	9	2
	Very Likely	1	12	20	4	1
	Certain	0	2	3	1	0

Total samples: 150
S5/S4 percent agreement: 44 / 150 = 29%
S5/S4 weighted kappa: 0.17

Table D.7 Agreement between Subjects S5 and S3

		Subject S3				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S5	Impossible	5	2	1	0	0
	Very Unlikely	16	17	8	3	2
	Mod. Likely	2	2	24	21	3
	Very Likely	1	3	14	19	1
	Certain	0	1	3	2	0

Total samples: 150
S5/S3 percent agreement: 65 / 150 = 43%
S5/S3 weighted kappa: 0.38

Table D.8 Agreement between Subjects S5 and S2

		Subject S2				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S5	Impossible	0	8	0	0	0
	Very Unlikely	6	27	7	6	0
	Mod. Likely	3	13	8	28	0
	Very Likely	1	5	9	22	1
	Certain	0	2	0	3	1

Total: 150
S5/S2 percent agreement: 58 / 150 = 39%
S5/S2 weighted kappa: 0.33

Table D.9 Agreement between Subjects S5 and S1

		Subject S1				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S5	Impossible	6	0	1	0	0
	Very Unlikely	28	3	7	1	0
	Mod. Likely	5	10	17	7	0
	Very Likely	2	8	18	7	0
	Certain	0	0	2	3	0

Total samples: 125
S5/S1 percent agreement: 33 / 125 = 26%
S5/S1 weighted kappa: 0.28

Table D.10 Agreement between Subjects S4 and S3

		Subject S3				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S4	Impossible	9	11	4	1	0
	Very Unlikely	4	8	23	12	2
	Mod. Likely	1	4	19	24	3
	Very Likely	2	1	4	8	1
	Certain	8	1	0	0	0

Total samples: 150
S4/S3 percent agreement: 44 / 150 = 30%
S4/S3 weighted kappa: 0.14

Table D.11 Agreement between Subjects S4 and S2

		Subject S2				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S4	Impossible	0	19	2	4	0
	Very Unlikely	0	21	8	20	0
	Mod. Likely	1	10	10	28	2
	Very Likely	0	5	4	7	0
	Certain	9	0	0	0	0

Total samples: 150
S4/S2 percent agreement: 38 / 150 = 0.25
S4/S2 weighted kappa: 0.033

Table D.12 Agreement between Subjects S4 and S1

		Subject S1				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S4	Impossible	19	2	4	0	0
	Very Unlikely	11	13	13	5	0
	Mod. Likely	3	5	23	10	0
	Very Likely	3	1	5	2	0
	Certain	5	0	0	1	0

Total samples: 125
S4/S1 percent agreement: 57 / 125 = 46%
S4/S1 weighted kappa: 0.31

Table D.13 Agreement between Subjects S3 and S2

		Subject S2				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S3	Impossible	8	14	1	1	0
	Very Unlikely	1	18	3	3	0
	Mod. Likely	1	19	8	22	0
	Very Likely	0	4	9	30	2
	Certain	0	0	3	3	0

Total samples: 150
S3/S2 percent agreement: 64 / 150 = 0.43%
S3/S2 weighted kappa: 0.45

Table D.14 Agreement between Subjects S3 and S1

		Subject S1				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S3	Impossible	18	0	0	2	0
	Very Unlikely	16	3	3	1	0
	Mod. Likely	4	13	19	7	0
	Very Likely	1	4	21	8	0
	Certain	2	1	2	0	0

Total samples: 125
S3/S1 percent agreement: 48 / 125 = 38%
S3/S1 weighted kappa: 0.37

Table D.15 Agreement between Subjects S2 and S1

		Subject S1				
		Impossible	Very Unlikely	Moderately Likely	Very Likely	Certain
Subject S2	Impossible	5	0	0	1	0
	Very Unlikely	28	6	15	3	0
	Mod. Likely	3	6	9	0	0
	Very Likely	5	9	20	13	0
	Certain	0	0	1	1	0

Total samples: 125
S2/S1 percent agreement: 33 / 125 = 26%
S2/S1 weighted kappa: 0.25

Appendix E

Predicting Acknowledgment Likelihood Study

Qualitative Results (Subject Feedback)

This appendix contains transcriptions of notes from the post-experiment interviews. The notes were taken by hand; no tape recording was made. These comments were made in response to the post-interview question "Can you articulate these rules [about how to code the utterances]."

E.1 Subject S1

- All yeahs depend on context: "Moderate"
- All questions: "Impossible"
- Most short ones: "Impossible" because no reason to comment
- Long ones seemed like narrative: "Very Unlikely"
- "Yeahs" are confusing

E.2 Subject S2

- Used inflection
- Question: "Impossible," otherwise avoided "Impossible" or "Certain"
- A one-word utterance is probably an Acknowledgment, but sometimes one Acknowledges an Acknowledgment!
- If speaker is subdued, probably one would just keep talking
- "Wow" would probably be acknowledged
- An inquisitive-sounding utterance might be acknowledged
- On the phone, must wait for pauses, so you don't know when pauses will be on these short samples
- Context matters

E.3 Subject S3

- A rising intonation invites acknowledgment: "Very Likely"
- When utterance ends in the middle of a factual statement: "Unlikely"
- When utterance ended at the end of factual statement: "Likely"
- Short statements and obvious acknowledgments: "Unlikely"
- Questions: "Impossible"
- Should clarify what to do about questions invoking "uh-huh," some in the sense of "understand" and some in the sense of "yes." (subject expressed confusion over the difference between "uh-huh" functioning as acknowledgment and as answer to yes-no question)

E.4 Subject S4

- Long-winded or stuttering utterances: low likelihood, because an acknowledgment would encourage them to continue
- Cheery voice: more likely
- Short, one-word utterances: Seemed like answer, no acknowledgment. The short ones were difficult at first; were they answering a question, or asking one? Decided based on intonation

E.5 Subject S5

- Didn't use "Certain" or "Impossible" much
- Most ratings in the middle
- Utterances that aren't complete sentences: tried to complete sentence and then respond

E.6 Subject S6

- "You know" asks for an acknowledgment
- Short utterances: "Unlikely" (because they sound like an acknowledgment)
- Questions: "Unlikely" (you'd answer the question instead)
- Inflection upward but not question: "Likely" (subject noticed 5 or 6 of these)
- Utterance ended in conjunction: "Unlikely"
- Situation relevant to own experience, identified with what was being said: "More Likely"
- "Impossible" and "Certain" were too strong

Appendix F

Predicting Acknowledgment Occurrence Study

Subject Profiles

Table F.1 Subject Profile Information for Predicting Acknowledgment Occurrence Study

Subj.	Age	Gender	Childhood Location	Other Location	Studied Linguistics
S1	30	Female	Oregon		No
S2	17	Female	Oregon, Arizona		No
S3	24	Female	Texas	Alaska, Kentucky	No
S4	31	Male	New York, Massachusetts	Japan, Oregon	Ph.D. student in speech recognition, has taken NLP class, has listened to short extracts from Switchboard corpus in testing forced-alignment program
S5	26	Male	Minnesota	West Virginia	No

Appendix G

Predicting Acknowledgment Occurrence Study

Results

Detailed results of the Predicting Acknowledgment Occurrence are documented in this section. These consist of tables showing, for each pair of subjects, their agreement on the coding task and the kappa statistic for that pair.

Table G.1 Agreement Between Subjects S5 and S4

		S4	
		No	Yes
S5	No	12	19
	Yes	30	27

Total samples: 88
S5/S4 percent agreement: 39 / 88 = 44%
S5/S4 kappa: 0.13

Table G.2 Agreement Between Subjects S5 and S3

		S3	
		No	Yes
S5	No	12	19
	Yes	34	23

Total samples: 88
S5/S3 percent agreement: 35 / 88 = 40%
S5/S3 kappa: 0.19

Table G.3 Agreement Between Subjects S5 and S2

		S4	
		No	Yes
S2	No	5	26
	Yes	12	45

Total samples: 88
S5/S2 percent agreement: 50 / 88 = 57%
S5/S2 kappa: 0.05

Table G.4 Agreement Between Subjects S5 and S1

		S4	
		No	Yes
S5	No	17	14
	Yes	32	25

Total: 88
S5/S1 percent agreement: 42 / 88 = 48%
S5/S1 kappa: 0.01

Table G.5 Agreement Between Subjects S4 and S3

		S3	
		No	Yes
S4	No	26	16
	Yes	20	26

Total samples: 88
S4/S3 percent agreement: 52 / 88 = 59%
S4/S3 kappa: 0.18

Table G.6 Agreement Between Subjects S4 and S2

		S2	
		No	Yes
S4	No	9	33
	Yes	8	38

Total samples: 88
S4/S2 percent agreement: 47 / 88 = 53%
S4/S2 kappa:.04

Table G.7 Agreement Between Subjects S4 and S1

		S1	
		No	Yes
S4	No	26	16
	Yes	23	23

Total samples: 88
S4/S1 percent agreement: 49 / 88 = 56%
S4/S1 kappa: 0.12

Table G.8 Agreement Between Subjects S3 and S2

		S2	
		No	Yes
S3	No	13	33
	Yes	4	38

Total samples: 88
S3/S2 percent agreement: 51 / 88 = 60%
S3/S2 kappa: 0.18

Table G.9 Agreement Between Subjects S3 and S1

		S1	
		No	Yes
S3	No	28	18
	Yes	21	21

Total samples: 88
S3/S1 percent agreement: 49 / 88 = 56%
S3/S1 kappa: 0.11

Table G.10 Agreement Between Subjects S2 and S1

		S1	
		No	Yes
S2	No	10	7
	Yes	39	32

Total samples: 88
S2/S1 percent agreement: 42 / 88 = 48%
S2/S1 kappa:.02

Appendix H

Predicting Acknowledgment Occurrence Study

Qualitative Results (Subject Feedback)

This appendix contains transcriptions of notes from the post-experiment interviews. The notes were taken by hand; no tape recording was made. These comments were made in response to the post-interview question "Can you articulate these rules [about how to code the utterances]?"

H.1 Subject S1

- Tried not to think about rules, tried to use gut instinct
- Tended to code "Yes" more for female speakers, "No" for male
- Male voices seemed more authoritative
- Tried to picture self in conversation
- Needed more practice, perhaps 20 samples instead of 10

H.2 Subject S2

- If the last part of the phrase included negative words, e.g., "won't," then "No"
- Inquiring tone: "Yes"
- Long utterances: "Yes," especially if it was a story or personal experience

H.3 Subject S3

- If they were continuing on, sounded like they needed prompting, or said "you know": "Yes"
- If they rambled on, sounded like they didn't need much prompting: "No"
- Pitch: when lower voice, looking for response: "Yes"
- Pitch: when higher pitch, means not finished: "No" (unless "you know")
- Statement such as "absolutely", I wouldn't respond to: "No"
- Hard to decide with no context
- A speaker may or may not allow you to respond, but you can talk over them if necessary

H.4 Subject S4

- "You know" occurred near end of utterance: "Yes"
- Question-like intonation: "Yes"
- Intonation more like straight affirmative statement: "No"
- Paused a lot: "Yes"
- The rest were random guesswork

H.5 Subject S5

- Speaker completed statement: "No"
- If not, other likely to elicit more information with an acknowledgment: "Yes"
- Subject assumed at first that these were sales calls, may have biased responses toward "No." Realized otherwise around sample 20.
- Short utterance: "Yes" (although not hard-and-fast, and took completeness of utterance into account)
- Tried to imagine contexts, the flipped a coin
- May have responded instinctively to pitch, but can't quantify

Appendix I

Eliciting Acknowledgments

Subject Forms and Instructions

This appendix contains the subject forms used in the Eliciting Acknowledgments study. These are:

- Subject Agreement and Release
- Subject Profile Information
- Subject Instructions

Dialogue Understanding Study
Agreement and Release

Description of Study

This study is under the direction of Dr. David Novick and Dr. Peter Heeman. Subjects will answer simple questions about the content of sample electronic mail and about the effectiveness of the spoken language interface used to access the mail. An audio record of the session will be made. The task is expected to require about 45 minutes to complete. Subjects will be paid $10.00 for their participation.

Subject Statement and Signature

I understand and agree that:

1.I may end my participation at any time for any reason. I will be paid for my participation whether or not I complete the experiment.

2.The experimental results are confidential; my name will not be associated with the experiment or with the profile information I provide. The results of the study, including audio records, may be used by members of the Laboratory and other persons designated by them for reasonable education, scientific, and technical purposes.

Signature: _____

Name: _____

Address: _____

_____

Telephone: _____

Date: _____

I have received $10.00 for my participation in this experiment:_____ (initial)

Dialogue Understanding Study
Profile Information

We need this information to interpret the results of the data you provide to us. Your profile will not be associated with your name.

Subject profile for: ____

Age: _____

Gender: _____

In what part of the country did you spend most of your childhood?

_____

In what parts of the world have you lived for more than 5 years?

_____

_____

Have you ever used a telephone interface to read email? Which one?_____

Have you studied (formally or informally) linguistics, spoken language understanding, or natural language processing? If so, what were your interests within those fields?

Dialogue Understanding Study Instructions

We are evaluating the effectiveness of various methods of presenting information. In this experiment, you will interact with a simple spoken-language interface to "read" email over the phone. That means that you talk to it, you don't push buttons. This program is designed to understand normal, everyday speech, so you can speak to it conversationally. It only knows how to read email, however; it does not, for example, understand the content of the email messages.

To help assess the understandability and usability of the interface, we'd like you to use the system to find the answers to the questions on the next page; the information is in the email messages. Please write the answers in the space provided. If you can't find the answer, write "not found." If you can't understand the message well enough to determine the answer, write "can't understand."

Appendix J

Eliciting Acknowledgments

Subject Profile Information

Table J.1 Subject Profile Information for Eliciting Acknowledgments Study

Subject		Gen-der	Age	Dialect Influences	Experience with Telephone Email	Linguistics/ NLP/SLS Background	Computer Experience
	Nbr[a]						
1	W1	F	49	Oklahoma Pacific Northwest	No	Some linguistics in college (1968/9)	professional programmer/ analyst
2	W2	F	42	Midwest Northwest	No	No	Office software
3	W3	omitted (subject guessed was wizard)					
4	W4	M	28	California Utah, Rhode Island	No	Speech from signal processing perspective (speech enhancement)	Grad student in computer science
5	W5b	M	48	Oregon	No	No	
6	W6b	F	56	Midwest	No	No	Office software

Table J.1 Subject Profile Information for Eliciting Acknowledgments Study

Subject		Gender	Age	Dialect Influences	Experience with Telephone Email	Linguistics/ NLP/SLS Background	Computer Experience
	Nbr[a]						
7	2W1	F	13	Oregon	No	No	Games, classes in school (Jr. High)
8	3W1	F	36	Ohio, Idaho Oregon	No	No	Office software
9	3W2[b]	F	55	Pacific Northwest	No	Intro. to Linguistics (30 years ago)	Office software
10	4W1	M	44	Southern California Oregon	No	No	Office software, database use
11	4W2	F	49	California Oregon	No	No	
12	4W3	M	27	West Coast New Jersey	No	No, but saw demo of research multimodal system	Management software, some Basic programming about 15 years ago
13	4W4	F	42	Boston Northwest	No	No	Office software
14	4W5	F	28	West Coast Alaska	No	No	Office software, programming class (high school)
15	4W6	F	50	Washington State Oregon	No	B.A. Degree in speech pathology and audiology	Office software

Table J.1 Subject Profile Information for Eliciting Acknowledgments Study

Subject		Gender	Age	Dialect Influences	Experience with Telephone Email	Linguistics/ NLP/SLS Background	Computer Experience
	Nbr[a]						
16	4W7	F	57	Oregon	No	No	Office software, drawing, landscaping packages
17	4W8	M	40	Texas Wisconsin	No	Has used spoken interface for airlines	Office software, layout
18	W49	F	47	Oregon	No	Has seen demos of early versions of CSLU Toolkit	Office software
19	4W10	M	24	Pacific Northwest	No	No	Games, web
20	4W11	F	24	Oregon Texas, Massachusetts, Georgia	No	Has seen Baldi demo	Office software, a little Visual Basic

a. This was the number that was used on the forms filled out by the subject.
b. Subject had some hearing loss and reported some trouble understanding synthesized speech

Appendix K

Eliciting Acknowledgments

Message Texts and Questions

This appendix contains the message texts and questions used in the Eliciting Acknowledgments study reported in Chapter 6. Unusual spellings and spacing were used to guide the synthesizer to the desired pronunciation. The spellings "c s e" and "sea ess ee" give the same auditory output.

K.1 Text of Message 1

message one is from c s l u at c s e dot o g i dot edu
because of the many people out of town
this week's staff meeting is cancelled
starting next week we will meet on thursdays
from twelve oh clock to one o clock
in room four zero one

K.2 Text of Message 2

message 2 is from heather at c s e dot o g i dot edu.
hi. here are the addresses you wanted.
Jim's address is jim at sea ess ee dot o g i dot edu.
and my mailing address is: 196 thousand North west Walker Road,
Beaverton, Oregon, 9 7 0 0 6.
I'm looking forward to receiving your report. See you at the conference next month!

K.3 Text of Message 3

message three is from p s u 1 2 3 4 5 at odin dot p d x dot edu about help with

homework
I'm having trouble with assignment three. do you have time to meet with me today?
I could come to your office now or at any of the following times. one thirty
three o clock
or five fifteen
thank you. I look forward to your prompt reply

K.4 Text of Message 4

message four is from david at sea ess ee dot o g i dot edu, about, hay ess are you 99
Call For Papers.
forwarded from Dana M Miller, d m m at research dot hay T T tee dot com.
the workshop on Automatic Speech Recognition and Understanding
will be held on December 12th through 15th, 1999,
in Keystone, Colorado.
the workshop will focus on recent advances and new paradigms and systems for
automatic speech recognition and understanding.
important deadlines are as follows:
electronic submission of photo-ready papers by August 15,
super early registration, before September 1,
notification of acceptance by October 1,
advance registration, before October 15.
for additional information, please check the workshop website at hay ess you are 99,
dot research, dot hay tee tee dot com.

K.5 Text of Message 5

Message five is from kathy, at see ess ee dot o g i dot edu, about, survey you are ell.
Thanks again for helping me with my research project.
My survey can be found at w w w dot see ess ee dot o g i dot edu, slash,
projects, slash,
project two, slash,
survey dot h t em ell.
Most people complete the survey in about 20 minutes.
I need to have the survey done by Friday. Thanks again, for your help!

K.6 Text of Message 6

Message six is from Jo at teleport dot com, about, please stop by store on your way
home.
I'm going to be late getting home tonight. so would you please stop by the store on

your way home?
We need milk,
eggs,
a bunch of spinach,
fresh ginger,
green onions,
maple syrup,
a pound of coos coos,
mild curry powder,
a pound of coffee,
and, a package of seventy five watt light bulbs.
Thanks! See you tonight.

Dialogue Understanding Study
Questions: Subject__

Please answer the following questions. The information will be found in the email messages.

1. How many email messages are there?

2. What is Jim Anderson's email address?

3. What is Heather's mailing address?

4. At what times can the student (psu12345@odin.pdx.edu) meet with you?

5. Which homework assignment is the student (psu12345@odin.pdx.edu) having trouble with?

6. What is the title of the workshop, and when and where will the workshop be held?

7. What is the deadline for submission of papers to the workshop?
   By what date will authors be notified of the acceptance of their papers?

8. What is the URL for the workshop website?

9. What is the new regular time and place for the staff meeting?

10. What is the URL for the survey?

11. By when does the survey need to be completed?

12. What items are you supposed to pick up at the store?

Appendix L

Eliciting Acknowledgments

Quantitative Results

Subjects were asked to find the answers to 12 questions with a total of 30 items to be reported. Results are reported in two tables. Table L.1 reports the task performance measures, as follows:

- "Items Correct" is the number of items reported correctly on the subject's sheet.
- "Items Located" is the number of items located but reported incorrectly (misunderstood, not understood, mis-transcribed).
- "Items Not Found" is the number of items for which the subject did not locate the answer (missing or wildly incorrect).
- "Task Completion Time" is the time to complete the task, rounded to the nearest minutes
- "Turns" is the number of subject utterances. False starts were counted as part of the following turn. Few turns with lengthy Task Completion Time may mean that subject frequently allowed system to proceed at the default pace.

No conclusions are drawn on these data due to interface changes during the course of the study (see Chapter 6).

Table L.2 reports the gender and the dialogue behaviors observed for each subject:

- "Acknowledgments" reports the type and number of acknowledgment or repetition behavior used by the subject. For acknowledgments, the word choice is reported, followed by the number of instances. For repetition, the total number of repetition instances is reported.
- "Commands" reports the numbers of commands by word choice.
- "Politeness" reports the numbers of occurrences of politeness behaviors.
- "Other" describes other unusual behaviors.

Table L.1 Task Performance for Eliciting Acknowledgments Study

Subject	Task Results (30 items total)			Task Completion Time (Minutes)	Turns
	Items Correct	Items Located	Items Not Found		
W1	29	1	0	15	121
W2	29	1	0	14	71
W3	omitted (subject guessed was wizard)				
W4	28	2	0	15	119
W5b	28	1	1	21	155
W6b	30	0	0	17	130
2W1	25	5	0	15	94
3W1	28	2	0	10	74
3W2	24	5	1	a	78
4W1	26	3	1	16	113
4W2	28	2	0	18	152
4W3	13	2	15	a	67
4W4	22	2	6	10	68
4W5	21	1	8	12	71
4W6	28	2	0	14	67
4W7	20	2	8	13	40
4W8	29	1	0	14	112
4W9	26	2	2	10	81
4W10	25	4	1	16	99
4W11	29	1	0	8	72

a. Tape recorded at wrong speed. Time not recoverable.

Table L.2 Dialogue Behaviors in Eliciting Acknowledgments Study

Subj.	Gender	Acknowledgments	Commands	Politeness	Other
W1	F	repetition: 15	continue: 34 next: 5 read: 1 (as next)	please: 1	meta[a]: 7 silence: 16 uh/um: 3
W2	F		continue: 38		silence: 28
W3	M	omitted (subject guessed was wizard)			
W4	M		go on: 60 next: 6 (all in context of summary)		silence: 3
W5b	M		continue: 63		silence: 23
W6b	F	okay: 3	continue: 14 next: 49	please: 4 goodbye: yes	silence: 12 responded to message content 4 times
2W1	F		go on: 49		silence: 16
3W1	F		continue: 42		silence: 6 uh: 4
3W2	F	all right: 1 okay: 1 yes: 4	next: 3	please: 51	silence: 42 uh: 3 responded to message content: 1
4W1	M	okay: 52 yes (ackn): 1	go on: 4		silence: 17 responded to content: 2
4W2	F		continue: 103		(note: ran through msgs twice) silence: 12 responded to content (thank you): 1
4W3	M	okay: 3 all right: 1	next: 12 go on, go ahead: 10		and: 1

Table L.2 Dialogue Behaviors in Eliciting Acknowledgments Study

Subj.	Gender	Acknowledgments	Commands	Politeness	Other
4W4	F		go on: 41	please: 3	
4W5	F	okay: 30	continue: 3 next: 7 go ahead: 1	please: 9	
4W6	F		continue: 39 next: 10 go ahead: 1	goodbye: yes	
4W7	F	okay: 1 all right: 1	continue: 18 next: 9 go on: 1 (but part of turn with "next message")	please: 21	meta[a]: 2 responded to message content: 3
4W8	M	okay: 2 (prefaced commands) yeah: 1 (ackn, prefaced command)	continue: 65 next: 16	please: 11 (in first 23 turns, none after that)	
4W9	F		continue: 52 next: 3 (all at message level)	please: 8 (all but 1 at message level)	
4W10	M	repetition: 4	continue: 49 go on: 3	thank you	meta[a]: 2
4W11	F	okay: 12 uhhuh/umhum: 28 yes: 9			meta[a]: 1

a. "Meta" includes comments on the task, e.g., "found one!" and laughter

Appendix M

Eliciting Acknowledgments

Qualitative Results (Subject Feedback)

This appendix contains transcriptions of notes from the post-experiment interviews. The notes were taken by hand; no tape recording was made.

M.1 Subject W1

Hard to understand, wouldn't use - but got nearly everything eventually.

Losing patience - hard to control at first, figured out commands, then OK.

Told it to go on, continue command.

Shopping list - perceived continued, but actually repeating items, more at ease, didn't connect continuing with speaking

M.2 Subject W2

NOTE: Subject initially pushed buttons. Restarted

Wasn't expecting to talk, make that clearer. Like it. Repeat option handy

Confusion trying to quit out of message early on

Realized didn't have to wait for it to prompt, could prompt it command continue

Did notice that sometimes it went on without saying anything, but felt like prompting most of the time.

When talking to a machine, act like a machine. don't expect to recognize "uh huh"

M.3 Subject W3

Omitted. Subject guessed was wizard.

M.4 Subject W4

relatively easy once got the hang of it, figured out commands

frustrating, couldn't say "read message 4". also, summarizing doesn't save place when reading

reading part pretty good - didn't realize at first that the first segment wasn't the entire message. phrased info relatively well, but conference title was too long.

two voices not helpful, but did recognize partition. strange meaning of "should I go on" on prompts - you weren't reading the mail. agreed might be confusing otherwise.

didn't realize would go on without prompt. only happened once or twice, decided had times out. liked being able to say do on, didn't want wait - had the impression could say anything, suspected "uh huh" or "yeah" would worked, but didn't experiment. didn't feel like talking with unconstrained enough system to risk experiment with it. its gives impression of system sophistication, concluded not robust. (explained experiment) a matter of how much confidence in the system.

reread earlier messages just to play around a little more. If people use for a month, may become more comfortable & do it. longer experiment? another session?

long conference title reduced confidence - consider splitting

M.5 Subject W5b

need to be explicit about *spoken* interface.

hard to understand. not hearing word fragments correctly. "and help" heard as "m l" on read help

very noisy, much static

noticed hh leading ay

at first didn't understand "do you want to continue" realized that voice would continue without commands, used that on shopping list quite deliberately slash chunked forward on URL presentations realized that system needed interaction, used "continue"

M.6 Subject W6b

couldn't hear at first, noisy. had some trouble figuring out how to get message. "message" can stop where need to instead of listening to whole thing, much easier. really like that.

wasn't sure how to get the number of messages

didn't notice that it would go on without any prompt, thought it asked every time. (did do this, didn't notice). used back and repeat to buy time while writing, next and continue to speed it up. good thing, to be able to control speed.

(asked about "okay" at beginning) "like I was talking to a person?" was unsure how to make it work.

M.7 Subject 2W1

TTS is very fast

perceived spacing between utterances as approximately one second (is 5)

female voice easier to understand. understood why two voices used.

noticed would go on, thinks it should wait. liked when asked whether it should go on

totally a computer

used only "go on" "repeat" "end" because thought computer would understand better.

M.8 Subject 3W1

took a while to realize had to prompt it for next line, thought required, but noticed that went on without prompt near end when taking longer to write. It no prompt, got message sometimes.

wait period felt too long, annoying. annoying to have to prompt, until things that had to write down were longer.

noticed chunks were of variable size, makes sense for a list of items

no way to check spelling of cathy

further got into, thought this would be useful - to check E-mail over phone.

understandable.

found chunking annoying eventually, because knew what was looking for. might work better if don't know what looking for.

once found something that works, stuck to it.

M.9 Subject 3W2

took a while to get the hang of it. Amazed at ability to understand different ways told to go on, very nice.

hard to understand long messages, URL, mild curry powder, spinach

"ready" hard to understand. once understood, thought meant system was waiting for a response.

didn't get message time, hard to understand CSE vs. ESE (subject reported some hearing loss)

really liked it, ease with telling to do things. liked response, much nicer then menus. liked chunking, once figured out could control pace. good that it went on without asking, noticed that it asked sometimes.

used "yes" near the end, like in conversation, caught self and quit doing because talking to machine. would probably use if used every day. became comfortable with it.

M.10 Subject 4W1

fun. wished for instructions to know system commands, but found it pretty intuitive. not as hard to understand as expected.

noticed gave time to understand. recognized would go on if said nothing, but felt had to say something to make it go faster. pros: wait while writing. con: time spent waiting to go on. but can say something, so ok. use this to control pace. useful.

okay meant okay, I'm done. what would say to computer.

M.11 Subject 4W2

"edu" hard to understand, also URL for workshop. no problem with program. experience with computers made commands "logical" only difficulty in listening and writing, went smoothly.

"thank you" responding to reader, as if on phone and getting message from secretary.

noticed would go on if nothing said. "enjoyed" telling it to go on, felt psychologically satisfying to be in control, more active, less passive. also faster. liked it. gave time to write, didn't have to repeat entire message.

continue just seemed like logical choice, clarity, exactly what I wanted it to do.

Very workable program, could use quite easily.

would you use acknowledgment to a computer? if knew would work. Yes, would be more satisfying, enjoyable, doesn't matter whether computer cares, can be more comfortable, pleasant. would be kind of fun.

M.12 Subject 4W3

NOTE: something wrong with the tape speed. too fast. No time estimate.

didn't find sixth message.

you end up confirming to it. simpler - just use the verb. would hope that can program to respond to own customer ID/personalized.

short segments. at first, wasn't sure whether segment or whole message. short segments useful, but would be better if whole thing with cutoff (barge-in) and rewind.

wasn't really aware of two voices, but clear on message vs. control

shopping list question - thought was a separate memo function.

didn't wait for me. was ok, because could go back. wait a little longer.

tried umhm, like to person (also ok). misunderstood - didn't work. decided to try several in a row. good, don't want to meet it halfway. know it looks for keywords.

M.13 Subject 4W4

cool. Liked that it stopped between items on list, so was easy. hard to understand accent "eddy". Subject didn't know what URL meant.

no directions as to how to work interface. "what words would I use if I was a computer?" easy to use. logical words. once found something that worked, stuck to it.

went on without saying anything, figured out could go back. figured was probably a certain passed time. might has been a problem if writing a lot of info

(noticed two voices) female voice was "instructor" male voice was E-mail itself. male accent more difficult.

used "go on" as prompt because ready to continue.

(described experiment)

think that "okay" would be most common form, but not knowing...

M.14 Subject 4W5

had some trouble.

noticed pause, sometimes thought was over, was taking notes on other info in messages.

segment - pause - was waiting for you to say something, or was over? confusing. may be too strong a word - if knew how working, would be ok

felt hurried when it went on without prompt. Not sure why it works this way.

okay - because that's what would say to a person, meant I understood. didn't seem strange

felt awkward at beginning, and when confusion near end. rest of it felt very natural, like someone telling me my messages.

M.15 Subject 4W6

"edu" possibly confusing, conference URL hard to understand. Subject used "continue": thought could probably say something else with fewer syllables, but had found something that worked. Eventually realized could speed up by saying something. "was this useful?" at first seemed useful, then annoying because was too slow. Would prefer one long unbroken message, like voice mail "even when copying?" I suppose that was why. Sender might divide message differently if they know it will be presented like this. Found URL messages unrealistic, because would use search engine to find a URL (described experiment and hypothesis) would use "ok, uhhuh" if had known that understood and less stress on voice, but it would just be an alternate command. related anecdote about use of "uhhuh" in chat room conversation - seemed weird.

M.16 Subject 4W7

NOTES: subject tried to negotiate meeting time with student

NOTES: despite extensive use of "please," subject hung up without warning

NOTES: "please" usage more consistent at beginning, dropped off near end

hard to understand. workshop title too fast, also URL.

subject asked for message nine: assumed question numbers mapped to message numbers. Didn't realize could repeat. didn't hear all messages to end, and didn't realize this. Background noise was disturbing, "clanging" noise NOTES: (may have been blower

coming on, quite audible in Room 118, where wizard was located. the recording was done by putting phone on speaker, so room noise may be audible. I didn't notice it when listening to the tape, though.)

expected six messages to be from six different people. interpreted voice as being the person's. Female/mail voice clear, but voice mail "conditioning" signalled message from same person. Voice too fast, kept hoping that next message would be different voice and slower.

(asked about message chunks) that was ok, but some too long intervals, and confusing in the beginning. (why did I do that?) to give people time to write and the opportunity to ask for repeat. felt nervous at first because parts of message not needed to answer questions. (notice that it would sometimes go on without prompting? good or bad?) good. confusing at first, didn't realize at first would go on automatically, until needed time to write more and used up time.

(reaction?) seemed silly, uncomfortable to say "please continue" to machine, haven't done this before. "oo, cool, it understood me" when it worked.

M.17 Subject 4W8

"edu" pronunciation confusing would be easier to use with experience. got easier as used it. easy to navigate. hard to understand voice at first. started out using complete sentences, simplified as went on. Figuring out easiest way to do it. would be nice if speech more natural (why?) would tend to treat it as he or she instead of it. Might be nicer, not lose patience, e.g., if in a hurry.

(what did you make of chunking) segment length varied. made easier for me, but would have preferred larger pieces. have more options that way, e.g., try to write down 6 things from shopping list at a time, repeat if necessary. User should be able to set as preference. Too small, slowed me down, but others might like pace. Noticed would go on, usually wanted to move faster than that. voice dichotomy "organized subliminally" clear. doesn't recall (segment) prompts, didn't understand.

(explained experiment) continue used because simple command. Said please at first, stopped because felt silly. would use acknowledgment if voice less mechanistic and if knew that it worked.

M.18 Subject 4W9

took a minute to figure out how long pauses were, when to say continue. believed had to prompt to go on, prompt usage suggested had to say something. (why "continue") made sense, leaves open whether done or not.

synthesis hard to understand, especially male voice. (why two voices?) didn't think about it. represent different levels, navigator vs. inside E-mail.

if not writing, better to just read. if transcribing or getting info, then helpful because can write.

((explained experiment)) would probably not use unless directed to.

M.19 Subject 4W10

hard to understand "edu." Pretty good system, liked being able to repeat. Could say "continue." like response to simple one-work commands, could skip end of message. Liked no buttons. (notice chunking?) what it said in middle of message different than between messages. liked. pleasant. less offensive than any systems. (why used 'continue') seemed like word to use. Star Trek influence. Annoyed (slightly) when it went on without command, but OK because could repeat. (should always wait for command?) no. (annoying to always say continue?) no, more interactive, more like doing it together, more lifelike. (two voices?) female was introducing, male voice read message. female easier to understand. (explained experiment) If computer more human, voice quality, would probably forget and do it.

M.20 Subject 4W11

cool thing. Really neat. Easy to use. Can I have one? Would possibly be useful under some circumstances. responded well: yes, uh huh, okay experimenting, probably use yes, uh-huh wished for pause before "what to do next" so didn't have to wait for it to finish. seemed to pause after punctuation (why?) makes it more conversational, separates info into pieces so can assimilate and understand synthetic voice, take notes. (liked?) yes, but for list of short items would prefer to have several items presented. (two voices?) noticed, one read messages, other was menu. very clear. (did notice would go on?) yes, but quicker (explained experiment. reasonable to talk to computer this way?) yes, fine. great not to have to use standard commands, makes it a little more natural.

cool

Biographical Sketch

Despite being a native Oregonian, Karen Ward was born in Lexington, Kentucky on July 17, 1956. A few months later her Oregonian parents realized the error of their ways and returned to Oregon to stay, so Karen grew up a few miles north of Portland, Oregon in the town of St. Helens.

In 1973, Karen moved to Eugene, Oregon to attend the University of Oregon. She returned to the Portland area in 1978 with a B. Sci. in Computer Science and accepted a position at Consolidated Freightways, first as a business applications programmer and later as a trainer and software standards specialist. In 1979, she moved to Portland General Electric and remained there until returning to school in 1991. While at PGE, Karen worked with teams responsible for developing and maintaining major software systems in support of PGE's engineering and business activities. During this time, her professional interests focused on database implementations, on software quality and on designing and implementing large systems and in expanding and enhancing existing systems.

Karen entered the Ph.D. program at Oregon Graduate Institute in 1991. She was awarded the M.S. degree in Computer Science in 1992. She has taught as an adjunct professor at Oregon Graduate Institute, at Oregon Institute of Technology-Portland and at the University of Portland. She is currently a member of the faculty of the Computer Science Department at the University of Texas at El Paso.