Improving User Interaction with Spoken Dialog Systems via Shaping

by

Stefanie L. Tomko

December 2006

Language Technologies Institute

School of Computer Science

Carnegie Mellon University

Pittsburgh PA 15213

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Thesis Committee:

Roni Rosenfeld, Chair

Alexander Rudnicky

Alexander Waibel

Candace Sidner, MERL

Keywords: spoken dialog systems, user interfaces, speech recognition, Speech Graffiti, Universal Speech Interface, structured input, adaptation, convergence, human-computer interaction

This research was supported in part by a National Defense Science and Engineering Graduate Fellowship and funding from the Pittsburgh Digital Greenhouse and the National Science Foundation. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

Abstract

Speech recognition technology offers an attractive interface option: speak to a computer, and it will understand you. One of the most promising applications of speech recognition technology is the spoken dialog system, which offers the promise of simple, direct, hands-free access to information. However, many factors conspire to make user communication with such a system less than optimally efficient. One problem is that users often speak beyond the bounds of what the computer is programmed to understand. This can lead to misunderstandings from the perspectives of both the user and the system, and recovering from such situations can add extra turns and time to the overall interaction. In this thesis, I describe a strategy, termed *shaping*, for improving user interaction efficiency with spoken dialog systems. This strategy involves the use of a target language designed to foster more efficient communication, and within which users will be encouraged to speak. When users interact with the dialog system and speak outside the target language, the system attempts to understand their input and aims to strike a balance between helping them complete the current task successfully and helping them increase the efficiency of future interactions by learning the target language (which in this case is Speech Graffiti).

The shaping strategies have been investigated through a series of three user studies with telephone-based spoken dialog systems. Results show that shaping can improve efficiency by removing the need for a pre-use tutorial and reducing word-error rates. Users in the studies exhibited significant intrasession, intersession, and cross-domain increases in Speech Graffiti grammaticality. The studies in this thesis have demonstrated a fully-functional, non-directed-dialog system, accessing real-world data, that takes advantage of users' propensity for convergence.

In memory of Susanne Tomko and Daniel Krug, who both always believed in me.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am clearly and overwhelmingly indebted to my advisor, Roni Rosenfeld, for his support and guidance throughout my whole graduate school career. Roni took me on as his student when I was a first-year Master's student with limited computer science experience, and taught me all about research—successfully enough that I stayed on at the LTI to pursue a PhD. I would like to thank Roni for all the time, sage advice, and support he dispensed generously throughout my time at CMU. I could not have asked for a better advisor. I would also like to thank my committee members, Alex Rudnicky, Alex Waibel, and Candy Sidner, for their helpful comments on and suggestions about this work. Alex Rudnicky in particular gave me a great deal of feedback very early on in the research process, before this work even made it into proposal form.

This work would probably never have been done had it not been for the early efforts of the Speech Graffiti/Universal Speech Interface research team. Arthur Toth, Xiaojin Zhu, and James Sanders did the bulk of the early coding and software engineering, and I thank them profusely for their hard work. Thomas Harris did more recent work on the system, and extended the Speech Graffiti protocol to control small devices. I must also thank Ravi Mosur, David Huggins-Daines, and Evandro Gouvêa for their help in solving various issues with the Speech Graffiti systems.

While at CMU, I have been lucky enough to be a part of the Sphinx speech research group, and I thank everyone there for providing a fantastic support network for all things speech-related, as well as more pizza than I ever needed. In particular, I would like to thank Alan W Black for help with the speech synthesis on both this project and on some of my earliest research at CMU. In addition, I would like to thank the members of the Dialogs on Dialogs reading group, which has been an invaluable

resource for comments and suggestions about my work: Dan Bohus, Antoine Raux, Jahanzeb Sherwani, Satanjeev Banerjee, Ananlada Chotimongkol, Thomas Harris, Jason Williams, Mihai Rotaru, Ellen Campana, Greg Aist, Sergio Grau Puerto, and Joel Tetrault.

This research also could not have happened without the contributions of all of my willing user study participants over the past few years. I appreciate their time and effort much more than they likely realize. I would also like to thank Sharon Cavlovich for her help with administrative details for my studies, and Melanie McClain and Mary Ann Merranko and their respective departments for facilitating the sessions I conducted at the University of Pittsburgh.

I am fortunate to have made many great friends while at CMU, and I am grateful to them for making my life in Pittsburgh more fun and the stresses of research more bearable: Tina Bennett, Paul Bennett, Mike Seltzer, Kathrin Probst, Ariadna Font-Llitjos, Jay Wylie, and Kevin Dixon.

I would like to thank my wonderful family for their love, support and encouragement. Finally, I must mention the best friend I made while at CMU, Dan Gaugel, who is now my husband. Dan has given me unqualified love and support throughout the whole, long thesis process, and has unquestionably made the journey much easier.

Chapter 1

Introduction and Outline

Speech recognition technology offers an attractive interface option: speak to a computer, and it will understand you. One of the most promising applications of speech recognition technology is spoken dialog systems.

A spoken dialog system—defined by Glass (1999) as "an interactive system which operates in a constrained domain"—offers the promise of simple, direct, hands-free access to information, yet several factors conspire to make user communication with such a system less than optimally efficient. One problem is that users often speak beyond the bounds of what the computer is programmed to understand. This can lead to misunderstandings from the perspectives of both the user and the system, and recovering from such situations can add extra turns and time to the overall interaction. Another issue is related more directly to efficiency: the long, explanatory

system prompts that are meant to be helpful for novice users can be tiresome and time-consuming for more advanced and frequent users. Furthermore, as in all speech recognition applications, there is always the issue of misrecognitions: cases in which, even though the user may have spoken within the bounds of what the computer is programmed to understand, the system generates an incorrect hypothesis as to what was said.

An efficient modality should be effective, fast, satisfying, and easy to learn. In this thesis, I describe a strategy, termed *shaping*, for improving user interaction efficiency with spoken dialog systems. This strategy involves the use of a target language designed to foster more efficient communication, and within which users will be encouraged to speak. For the purposes of this research, the target language will be Speech Graffiti, which has been shown to have shorter task completion times, lower word- and concept-error rates, and higher user satisfaction ratings when compared to a natural language speech interface. When users interact with the dialog system and speak outside the target language, the system attempts to understand their input and aims to strike a balance between helping them complete the current task successfully and helping them increase the efficiency of future interactions by learning the target language.

The rest of this chapter discusses the advantages of Speech Graffiti for addressing the problems noted above, and the enhancements to it that comprise the core work of this project, followed by a summary of the goals of this thesis. Chapter 2 presents a summary of related work. Chapter 3 describes the process of shaping in more detail. Chapters 4 through 10 present the designs and results of three user studies conducted to assess the effectiveness of shaping on interaction efficiency, and Chapter 11 summarizes the overall findings. Chapter 11 also includes a discussion of potential future extensions to this work.

1.1 Spoken dialog systems

As one of the most common modes of human-human interaction, speech could be considered an ideal medium for human-computer interaction. Speech is natural and the vast majority of humans are already fluent in using it for interpersonal communication. Speech is portable, it supports hands-free interaction, and its use is not limited by the form factor of speech-enabled devices. Furthermore, technology now exists for reliably allowing machines to process and respond to basic human speech. Speech is currently used as an interface medium in many commercially available applications, such as dictation systems (*e.g.,* IBM® ViaVoice® and Dragon™ NaturallySpeaking®), web browsers (*e.g.,* Conversay Voice Surfer™), and spoken dialog systems (*e.g.,*1-800-555-TELL™ from TellMe). This research focuses on the latter class of speech applications.

Despite the potential advantages of speech interaction, many problems still exist in the design of user interfaces for spoken dialog systems. For example, a principal advantage of using spoken language for communication is its unbounded variability, yet speech recognition systems perform best when the speaker uses a limited vocabulary and syntax (Kamm, Walker, & Rabiner, 1997). In addition, unlike simple dictation systems that use speech recognition technology, spoken dialog systems must do more than simply identify the words that are spoken. When humans hear speech, they extract semantic and pragmatic meanings from the string of words based on their syntax, prosodic features, and the context (both spoken and situational) in which they were uttered (Searle, 1970). The challenge of spoken dialog systems is to interpret user input in order to execute the user's request correctly, while at the same time approximating the role of a conversational partner. Humans tend to follow certain implicit rules when engaging in conversations with others, such as being brief, being "orderly," and making contributions that are no more and no less informative than the situation requires (Grice, 1975). Humans also expect

both participants in an interaction to work to make the conversation succeed, especially with respect to problems that arise over the course of the conversation (Clark, 1994). Successful spoken dialog systems must, to some extent, be designed to address these expectations.

In addition to these conversational requirements, spoken dialog systems must deal with issues directly related to the speech signal. They must be able to handle noise, both environmental (including persistent noise such as loud cooling fans, and intermittent sounds like door slams or a passing truck) and internal to the speaker (*e.g.,* speech to another person or coughing). They must also be able to handle between-speaker variations (*e.g.,* male or female voices, different accents, or different ages). Although some speech recognition applications are designed to be speaker-dependent and can therefore tailor their recognition parameters to a specific user's voice, spoken dialog systems are usually designed as interfaces to applications intended to be used by a large number of people. Such applications are often accessed via telephone, which has been shown to increase recognition word-error rates by approximately 10% (Moreno & Stern, 1994), or possibly at a public kiosk, which is also likely to add a significant environmental noise factor.

Finally, spoken dialog systems must deal with the serial and non-persistent nature of speech-based interaction. In contrast to face-to-face human conversation, in which a listener might express understanding problems via facial gestures or interruptions while a speaker talks, spoken dialog systems generally impose a fairly strict turn-based interaction, in which the system does not respond until the user is finished speaking (although most systems do allow users to "barge in" on the system while it is talking). This can generate significant user frustration if the speaker has uttered a long string of input only to discover at the end that the system did not understand any of it (Porzel & Baudis, 2004). Although multi-modal systems exist that incorporate both visual and spoken interface components (see Oviatt et al., 2000, for

an overview), visual displays are not always possible (*e.g.,* in telephone or other remote-access systems) or desirable (*e.g.,* in automotive systems) (Cohen & Oviatt, 1995). Spoken dialog systems must therefore give special consideration to features such as effectively presenting large blocks of information, facilitating interface navigation, and providing support for users to request confirmation of the system's state.

In summary, well-designed spoken dialog systems must take many factors into account:

- they must be able to interpret user input appropriately;
- they must be able to play the appropriate role for a participant in a conversation;
- they must be able to handle errors that result from speech recognition problems; and
- they must be able to present information effectively.

At the same time, it is worth keeping in mind Allen et al.'s Practical Dialogue Hypothesis (2001):

> *The conversational competence required for practical dialogues, while still complex, is significantly simpler to achieve than general human conversational competence.*

Thus, it is possible that successful spoken dialog systems do not need to exactly match human competencies on the above issues. The Speech Graffiti approach capitalizes on this hypothesis.

1.2 Approaches to spoken dialog systems

Spoken dialog systems can be loosely divided into three categories: command-and-control, directed dialog, and natural language (although some might argue that command-and-control systems are not true "dialog" systems, since there is often limited turn-taking and system feedback; they are included here to give a complete picture of the range of application types). One way these categories can be differentiated is in terms of what users can say to the system and how difficult it is for developers to create the system (or, conversely, how easy it is for the system to handle the user's input). In general, there is usually a trade-off between the "naturalness" of a system and the ease with which it can be developed (fig. 1.1).

Command-and-control systems severely constrain what a user can say to a machine by limiting input to strict, specialized commands or simple yes/no answers and digits. Since such systems do not require overly complicated grammars, these can be the simplest types of systems to design, and can usually offer low speech recognition word-error rates (WER). However, they can be difficult or frustrating for users since, if input is limited to yes/no answers or digits, users may not be able to perform a

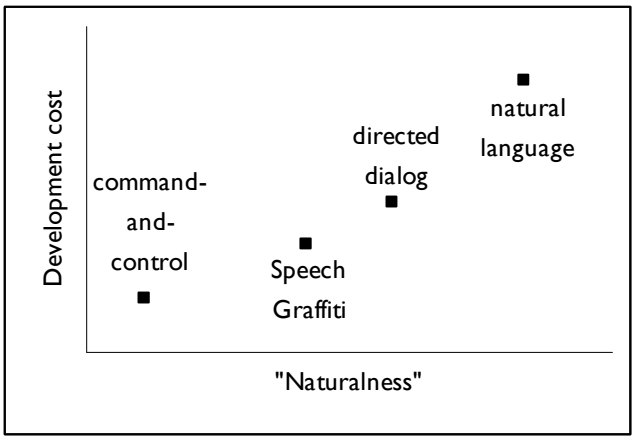

Figure 1.1. Schematic representation of development costs vs. "naturalness" trade-offs in spoken dialog system types.

desired task by using only the available choices. If specialized input is required, users will have to learn a completely new set of commands for each voice interface they come in contact with. Under this paradigm, a user might have to learn five completely different voice commands in order to set the clock time on five separate appliances. While this may not be an unreasonable solution for applications that are used extensively every day (allowing the user to learn the interaction through repeated use), it does not scale up to an environment containing dozens or hundreds of applications that are each used only sporadically.

Directed dialog interfaces are widely used in commercial applications. Such systems use machine-prompted dialogs to guide users to their goals, but this is not much of an improvement over the touch-tone menu interfaces ubiquitous in telephone-based systems (Press or say 1 for billing…)[1]. In these systems, the user is often forced to listen to a catalog of options, most of which are likely to be irrelevant to his or her goal. Directed dialog interactions tend to be slower, although error rates can be lower due to the shorter and more restricted input that is expected by the system (Meng, Lee, & Wai, 2000). When directed dialog systems allow barge-in, experienced users may be able to speed up their interactions by memorizing the appropriate sequence of words to say (as they might with key press sequences in a touch-tone menu system), but these sequences are usually not valid across different applications. Users therefore must learn a separate interface pattern and vocabulary for each new system used and for whenever an existing, familiar system is modified.

In *natural language interfaces*, users can pose questions and give directives to a system using the same open, conversational, potentially ambiguous language that they would

---

[1] Throughout this document, this typeface is used to represent system prompts. **This typeface is used to represent user input.** This typeface is used to represent speech recognition hypotheses.

be likely to use when talking to a human about the same task (*e.g.,* **When's the first flight to New York Monday?** or **Did my stocks go up?**). By giving great freedom to the user, this option avoids the issue of forcing the user to learn specialized commands and to work within a rigid access structure. However, it puts a heavy burden on system developers who must incorporate a substantial amount of domain knowledge into what is usually a very complex model of understanding, and who must include all reasonably possible user input in the system's dictionary and grammar. The large vocabularies and complex grammars necessary for such systems and the conversational input style they are likely to generate can adversely affect speech recognition accuracy. For instance, Weintraub, Taussig, Hunicke-Smith, and Snodgrass (1996) reported word-error rates of 52.6% for spontaneous, conversational speech, compared to 28.8% for read, dictation speech.

Furthermore, although the inherent naturalness of such interfaces suggests that they should be quite simple to use, this apparent advantage can at the same time be problematic: the more natural a system is, the more likely it is for users, particularly novice ones, to experience problems caused by their having overestimated the bounds of and formed unrealistic expectations about such a system (Perlman, 1984; Glass, 1999). Williams & Witt (2004) reported that in comparison with directed dialog systems, natural language, "how may I help you?"-style interactions produced lower user satisfaction and task success rates, most plausibly because of a lack of guidance as to what to say to the system. Another potential issue with natural language systems, suggested by Shneiderman (1980b), is that "natural" communication may actually be too lengthy for frequent, experienced users, who expect a computer to be a tool that will give them information as quickly as possible.

1.3 The Speech Graffiti approach

Speech Graffiti offers a middle-of-the-road approach to solving the issues discussed above. Speech Graffiti comprises a small set of standard keywords plus structural and interaction rules which can be used in all Speech Graffiti applications. By standardizing user input, Speech Graffiti aims to reduce the negative effects of variability on system complexity, similar to the way that Graffiti® handwriting recognition software for hand-held computers requires users to slightly modify their writing in a standardized way in order to improve recognition performance.[2] At the same time, the introduction of a universal structure that is intended to be used with many different applications should mitigate negative effects that might be otherwise associated with learning an application-specific command language.

1.3.1 The ATUE study

In previous work, I reported findings from a user study (hereafter referred to as the ATUE study, for "Assessing the User Experience") showing that with Speech Graffiti, users had significantly higher levels of user satisfaction ($t = 3.20$, $p < 0.003$), faster task completion times, and similar overall task completion rates compared to with a natural language spoken dialog system (Tomko, 2003). The study also showed that Speech Graffiti generated lower word- and concept-error rates compared to a natural language interface in the same domain (table 1.1).

Table 1.1. Speech Graffiti and natural language system error rates in the ATUE study.

	Speech Graffiti (ATUE version)		Natural language system			
Error measure	M	SD	M	SD	*t*	*p*
Mean per-user % concept error	26.62	17.77	50.74	13.60	6.03	<.0001
Mean per-user % word error	34.99	16.28	50.33	10.84	5.11	<.0001

Such benefits come with a lower overall system development cost, since a toolkit is available to facilitate the development of new Speech Graffiti applications (Toth, Harris, Sanders, Shriver, & Rosenfeld, 2002). I also showed that task success and user satisfaction with Speech Graffiti was significantly correlated with grammaticality (how often users spoke within the grammar) (Tomko & Rosenfeld, 2004b). This suggests that it is very important to help users learn to speak within the grammatical bounds of spoken dialog systems. In the ATUE study, nearly all participants who used correct Speech Graffiti grammar in at least 80% of their utterances gave the system positive user satisfaction ratings, and more than half of the ATUE participants achieved this level. Furthermore, users with grammaticality above 80% completed an average of 6.9 tasks (out of eight), while users with grammaticality below 80% completed an average of only 3.5 tasks. Based on these results, 80% grammaticality appears to be a reasonable preliminary target for supporting successful, efficient interactions.

1.3.2 Issues with Speech Graffiti

However, even after engaging in a pre-use training session, some users have difficulty using the Speech Graffiti system. At the conclusion of the ATUE study, users (all of whom had interacted with both systems) were asked whether they preferred Speech Graffiti or the natural language system, and six of the 23 participants chose the natural language system. The experience of these six users provides a snapshot of frustrating communication. In their Speech Graffiti interactions, they accounted for the six highest word- and concept-error rates, the six lowest task completion rates, and the four lowest grammaticality rates. One defining characteristic of these six participants was that all but one of them belonged to the group of thirteen study participants who did not have computer programming backgrounds. As a very minimal proof-of-concept exercise, one of these six participants returned to the lab a year after the ATUE study to interact with a Speech Graffiti application again. She

was given a more intensive, interactive, pre-use training session and was encouraged to ask questions both during the training and while working on tasks. This time, her interaction was much more successful and less frustrating. While only a single data point, this experiment suggests that, given the right help, successful, efficient interactions *are* achievable with the Speech Graffiti system for a broad range of users.

1.4 Shaping

The strategy for increased interaction efficiency via shaping has been designed with such users in mind. The main obstacle to success with Speech Graffiti in the ATUE study appeared to be learning and remembering the language. Although participants had received a tutorial before using the system, I often observed a pattern in which users would work through the tutorial (an HTML-based guide that users read through on their own for ten to fifteen minutes), declare that they understood the concepts and were ready to work on the experimental tasks, and then promptly forget what to say once they were on the telephone with the working system. This indicated that pre-use tutorials were not the most effective tool for helping users become proficient with Speech Graffiti. Additionally, a pre-use tutorial can, in many situations, be impractical due to the user's time or environmental constraints. This suggested that I should investigate helping users learn Speech Graffiti at run time. This approach would have the benefit of helping users learn the more efficient, Speech Graffiti interaction style while at the same time achieving their current task goals. Thus, the system would *shape* users towards the target way of speaking. The specifics of shaping are discussed thoroughly in Chapter 3.

A note about terminology: To describe the process of users' adapting their input to match the system's prompts, this work borrows the term *shaping* from the cognitive psychology concept of successive conditioning of new responses (Domjan, 2005). This term has also been used in the context of adaptation in computer-mediated

dialog by Ringle and Halstead-Nussloch (1989). I use the term *convergence* to describe successful, grammatical Speech Graffiti shaping (*i.e.*, if the system *shapes* successfully, the user *converges*). Other researchers have used such terms as modeling (Zoltan-Ford, 1991), alignment (*e.g.*, Pickering & Garrod, 2004), coordination (*e.g.*, Branigan, Pickering, & Cleland, 2000), and entrainment (*e.g.*, Brennan, 1996) to describe similar processes or results (see Section 2.2).

1.5 Why Speech Graffiti?

The shaping strategies implemented in this work rely on the implementation—in parallel with the target, Speech Graffiti grammar—of an expanded grammar that can accept more natural language input than canonical Speech Graffiti (see Section 3.2). Although this expanded grammar is not as comprehensive as one that would be used in an interface specifically designed as a natural language spoken dialog system, a frequently raised question with this approach is why user input should then be shaped towards a more constrained target language. Several aspects of Speech Graffiti are presented here as support for this approach.

1.5.1 Universality

From the user's perspective, Speech Graffiti's structures and keywords are universal. That is, the structures and keywords learned while interacting with one Speech Graffiti system can be reused when interacting with other Speech Graffiti systems. This is in contrast to traditional command-and-control systems, in which separate applications generally have unique interaction protocols. It might be suggested that natural language systems offer the ultimate in universality: one speaks as naturally to one natural language system as to any other. But in fact linguistic coverage is not likely to be exactly the same across different natural language systems. For instance, some systems may support anaphora resolution, allowing users to say things like **`tell me more about that`**, while other systems may require users to be more

explicit about what "**that**" is. Even within the expanded grammar that is implemented in this research, syntactic and functional capabilities vary across domains. Shaping user input to the Speech Graffiti target language helps ensure that users have a universal set of skills for using with all Speech Graffiti applications.

1.5.2 Efficiency

Compared to natural language interfaces, Speech Graffiti interactions tend to be more brief. In the ATUE study, median task completion time was about 21% shorter for Speech Graffiti than for natural language. Speech Graffiti should also facilitate more efficient interactions compared to those of directed dialog systems. Because Speech Graffiti is essentially a user initiative system, users do not have to work through menus or listen to a series of lengthy prompts before creating queries for the exact information for which they are searching. Speech Graffiti's restricted language also generated significantly lower speech recognition word- and concept-error rates compared to a natural language system (see table 1), thus reducing the chance that interaction-lengthening, error correction turns will be introduced.

1.5.3 Transparency

Two key features of Speech Graffiti promote interface transparency: orientation keywords and explicit confirmation. Speech Graffiti includes keywords that help users orient themselves within an interaction. **`Options`** allows users to find out what they can say next at any point in an interaction and **`where was I?`** prompts the system to repeat all of the information it has stored for the current query. Both keywords are easy to implement in a structured system like Speech Graffiti. Although a keyword like **`where was I?`** should be fairly simple to include in natural language systems as well, **`options`**-type functions can be difficult to implement since the space of things that users can say at any point can be very large or difficult to explain. To further enhance Speech Graffiti's transparency, the response that is

generated for an **`options`** request includes a comprehensive list of available slots. Therefore, by saying **`options`** the user can easily learn the functional and domain boundaries of the system. This kind of information is notably hard to convey in natural language systems (Ogden & Bernick, 1997).

Furthermore, Speech Graffiti's explicit confirmation strategy provides feedback on recognized items at every input turn. Grounding—the process of participants "com[ing] to the mutual belief that they understand one another sufficiently well for the purpose at hand"—is a key component of human interactions (Brennan, 1998), and previous research with the system has shown that users perceive Speech Graffiti's explicit confirmations as a beneficial, grounding step that is missing in some natural language systems (Shriver et al., 2001). It has also been demonstrated that explicit confirmation messages facilitate faster recovery from error incidents compared to implicit confirmations (Shin, Narayanan, Gerber, Kazemzadeh, & Byrd, 2002).

1.5.4 Portability

Speech Graffiti was designed to support the creation of interfaces to new information-access domains with minimal language engineering and effort. A web application generator has been created that allows developers to create new applications by providing basic information such as vocabulary-to-database-column mappings via an HTML form (Toth et al., 2002). Of course, some amount of domain knowledge is still required for the creation of Speech Graffiti applications. For instance, developers may want to predict common synonyms for slot and value names. In some cases, for "smarter" interactions, they may also want to program some domain-specific default constraints for database queries (for example, in the Speech Graffiti MovieLine, a default "date = today" constraint is added to the user's query unless the date is otherwise specified or the query is not date-dependent, such as a request for the address of a theater). Although there have been efforts to

simplify and modularize the creation of new natural language spoken dialog systems (*e.g.,* Nakano et al., 2000; Glass & Weinstein, 2001; Denecke, 2002), I believe that the Speech Graffiti approach substantially minimizes the amount of syntactic and deeper semantic knowledge and analysis required.

1.5.5 Flexibility

As noted in the previous subsection, Speech Graffiti is not designed to require intensive domain knowledge and concept mapping. Speech Graffiti slots are simply matched to database columns, thus allowing users to customize queries to their needs. From the user's point of view, natural language systems may be assumed to be highly flexible, but such systems have functional limitations based on what input-to-concept mappings have been encoded in the dialog manager by developers. For instance, in the ATUE study, the natural language system allowed users to query genre information only in terms of specific movies (*e.g.,* **`What kind of movie is Star Wars?`**). The system did not support queries like **`What kind of movies are playing at the Manor Theater?`**, even though the backend database was capable of retrieving such information. In contrast, in Speech Graffiti any permutation of slots can generate queries.

1.6 Thesis statement

Shaping can be used to induce more efficient user interactions with spoken dialog systems. The shaping strategy can improve efficiency by increasing the amount of user input that is actually understood by the system, leading to increased task completion rates and higher user satisfaction. This strategy can also reduce upfront training time, thus accelerating the process of realizing more efficient interaction.

1.7 Research contributions

The main contributions of this work, which will be discussed in Chapter 11, can be summarized as:

- An understanding of which factors shape user input most effectively in spoken dialog systems and when such shaping should be done. These findings should have broader applicability for all spoken dialog systems, not just subset language ones.
- A strategy for increasing the efficiency and effectiveness of user interaction with spoken dialog systems.
- A functional system that exploits the phenomena of shaping and entrainment observed in human-human and human-computer interactions to a greater extent than has been done in previous research.

Chapter 2

Related Work

2.1 Restricted languages

Despite the interest in and research challenges posed by conversational natural language interfaces, various studies and researchers have suggested that restricted or subset languages such as Speech Graffiti are indeed a reasonable approach to spoken interaction with computers and that such input is not necessarily unnatural. For instance, Shneiderman (1980a) suggests that using a small, well-defined language may actually make interactions easier for novices, since it clarifies what is and what is not accepted by the system. Structured interactions tend to generate significantly fewer parses per utterance (Oviatt, Cohen, & Wang, 1994). In contrast, longer, unconstrained utterances have been shown to generate more disfluencies (Oviatt, 1995), thus making the speech recognition process less accurate for more conversational spoken dialog systems.

Kelly (1977) conducted a study in which users completed tasks (via typed input) using either unlimited vocabularies or restricted vocabularies of 500 or 300 words. The restricted vocabularies were chosen based on frequency of use in solving similar tasks in the same domain. Prior to working on a set of tasks, participants studied the vocabulary until they could pass a recognition test with 75% accuracy. Kelly found no significant differences in the time required to solve problems with different vocabulary sizes, and noted that subjects easily adjusted to the restricted vocabularies they had to work with. In simple memory experiments, Black and Moran (1982) also found that sets of command words prescribed by system designers were no more difficult for users to remember than ones generated by users themselves.

Hendler and Michaelis (1983) conducted a study in which participants completed problem-solving tasks with a partner by sending text messages over linked terminals. Participants in one condition were told that the system only accepted a strict, limited grammar, although they were not actually told what that grammar was. When users in this group sent an ungrammatical message, the message was blocked and sent back to the sender marked as ungrammatical. Participants had one hour to complete each of three tasks. Although users in the limited grammar condition took nearly twice as long to complete the first task compared to participants in the non-limited condition, completion times did not differ significantly for the second and third tasks, indicating that users soon became comfortable with the grammar limitation. Jackson (1983) has also shown successful user adaptations to syntax restrictions in text-based interactions.

More recently, Sidner and Forlines (2002) conducted a study on a restricted language interface for a home entertainment system and showed that users were able to complete tasks successfully. They found that participants' performance did not decline when attempting tasks the following day, thus demonstrating that users were able to retain their knowledge of the restricted language. However, only three out of

21 participants achieved near-perfect grammaticality, and users of this system had visual help available, either continuously or upon request. The restricted language used in this study was designed to use simple, common English grammatical structures and a limited vocabulary, but unlike Speech Graffiti it was not necessarily designed to be adaptable to different domains.

These studies demonstrate that users can indeed interact successfully with restricted language interfaces. However, most of these experiments used modalities with a visual component, thus avoiding some of the problems of non-persistence and asychronicity inherent in speech communication. The ATUE study confirmed the potential for restricted languages in speech-only human-computer interaction, but also established the need for further research on making such languages and interfaces more habitable for all users.

2.2 Convergence and shaping

Convergence—"the process of interaction adaptation whereby one partner adopts behavior that is increasingly similar to that of the other partner" (Burgoon, Stern, & Dillman, 1995)—has been well documented in human-human interactions.

For instance, in a classic study, Matarazzo, Weitman, Saslow, and Weins (1963) found that in interpersonal interview settings, the duration of interviewee utterances was significantly affected by the duration of interviewer utterances. This affect was bi-directional, such that when interviewer utterances grew longer, interviewee utterances also lengthened; when interviewer utterances became shorter, interviewee utterances shortened.

Convergence has also been observed in syntax (Branigan, Pickering, & Cleland, 2000) and linguistic style (Niederhoffer & Pennebaker, 2002). In fact, Pickering and

Garrod (2004) have theorized that participants in a dialog align their interaction at *all* linguistic levels.

In experiments with computer-mediated human-human dialog, Ringle and Halstead-Nussloch (1989) explored whether more formal responses could be used to shape user input to be syntactically simpler. In this study, users sought assistance via typed input from a remote human tutor on a document editing sub-task. Participants knew that they were interacting with a human tutor (albeit through a computer-mediated channel) and were told that there were no restrictions on the content or length of their input. In the "natural" condition, the tutor handled and replied to all input as in normal human interaction, but in the "formal" condition, the tutor attempted to simulate a limited, rule-based system with regard to parsing and handling input. The authors found that users in the formal condition produced input with significantly fewer parsing problems and with significantly lower complexity compared to the natural condition.

Adaptation and convergence have been found in human-computer interactions as well. It has been shown that users converge on prosodic features such as the amplitude and speed of computer systems' text-to-speech output (Coulston, Oviatt, & Darves, 2002; Darves & Oviatt, 2002; Suzuki & Katagiri, 2003; Bell, 2003). For this research, I am particularly interested in syntactic and stylistic adaptation and convergence.

The mere belief that one is interacting with a computer as opposed to with a human has been shown to affect users' input style. In the UNIX help domain, Chin (1984) found that users in a control group who were told they were interacting with a human operator used more anaphora and ellipsis than participants in the (simulated) intelligent help group. Guindon, Shuldberg, and Conner (1987) found that people used more formal language in typed, Wizard-of-Oz interactions with an interactive

help program, apparently under the impression that the computer system could not handle more complex input. In the ATUE study, I observed a natural restriction effect in the scope of natural language constructions. Considering items like movie and theater names to be equivalence class members, the utterances used by participants when speaking to the natural language MovieLine reduced to approximately 580 different syntactic patterns. In contrast, in the Speech Graffiti MovieLine, when users spoke outside the Speech Graffiti grammar and used natural language instead, their utterances reduced to only 94 syntactic patterns. One of the main differences between the syntactic patterns in the two systems was the lack of conversational phrases like **`can you give me…`** and **`I would like to hear about…`** in natural language speech to the Speech Graffiti system. Thus the use of a restricted language system influenced users to speak in a simpler way, even though they did not always speak in exactly the "correct" simplified manner. Shriberg, Wilder, and Price (1992) have observed that input simplification also occurs when speech recognition word-error rates are high.

Zoltan-Ford (1991) conducted a Wizard-of-Oz study to determine the effects of several potentially influential variables on user input: conversational vs. terse system output, restricted vs. unrestricted user input, familiar vs. unfamiliar vocabulary, and keyboard vs. voice input. She found that terse system outputs generated user inputs that were 60% shorter than those generated by conversational system output, and that restricted-input users were much more likely to match their input to the system's output characteristics. Overall, she found that explicit shaping, in which errors occur if the user does not adapt, was a more effective influence on user input than modeling, in which the user adapts naturally to the computer's style. However, she found that the effectiveness of shaping came at the cost of increased number of messages sent to the system by users.

Brennan (1996) found strong evidence for lexical entrainment, or the shared use of the same term to refer to the same object, in human-computer interaction. In a speech-based database manipulation task, she found that users adopted the computer's term 88% of the time when it was explicit or exposed (*e.g.,* By college, do you mean school?) and 58% of the time when it was implicit or embedded. She also found an effect of memory: users adopted the system's term 87% of the time when the object needed to be re-referred to immediately, and 59% of the time when the object was re-referred to later.

Gustafson, Larsson, Carlson, and Hellman (1997) and Bell (2003) found strong entrainment effects for both vocabulary and syntax in directed-dialog systems. Branigan, Pickering, Pearson, McLean, and Nass (2003) found evidence of syntactic convergence (for object attachment) for both assumed human-human and human-computer typed interactions (in actuality, the partner was a computer in both cases). Pearson, Hu, Branigan, Pickering, and Nass (2006) reported on lexical convergence in typed human-computer interactions, with the interesting finding that the effect was significantly stronger in the condition where the computer partner was presented as a "basic" model as opposed to a more up-to-date, advanced one (although in reality the systems were both exactly the same).

A few caveats should be considered about the human-computer convergence findings noted above. First, many of them were based on typed rather than spoken interactions. Of those that did involve speech, most were Wizard-of-Oz studies or experimental situations that did not represent typical spoken dialog system tasks. The results reported by Gustafson et al. (1997) involved spoken dialog systems, but in the context of directed dialog questions. Bell (2003) observed lexical convergence in functional spoken dialog systems, but without the goal of actually encouraging convergence to a specific form. In Zoltan-Ford's (1991) study, users were told that

they were working with a system that could communicate in "ordinary, everyday English," which could perhaps have weakened shaping effects that occurred.

Speech Accommodation Theory posits that one reason speakers converge is to improve communication efficiency (Giles, Mulac, Bradac, & Johnson, 1987). Although all of the studies discussed above demonstrate the existence of the phenomena of adaptation and convergence in human interaction, both with other humans and with computers, no studies that could be found that specifically address the idea of exploiting these phenomena to improve the quality of human-computer speech interaction, as I explore in this work.

2.3 Error identification and handling in spoken dialog systems

Because of the uncertainty inherent in the automatic speech recognition procedure, error identification and handling has been an area of particular interest in spoken dialog systems research. Significant work has been done on identifying and predicting error situations (*e.g.,* Walker, Langkilde, Wright, Gorin, & Litman, 2000; van den Bosch, Krahmer, & Swerts, 2001; Litman, Hirschberg, & Swerts, 2001) and designing and evaluating error handling and repair strategies (*e.g.,* Goldberg, Ostendorf, & Kirchoff, 2003; Bousquet-Vernhettes, Privat, & Vigouroux, 2003; Bohus, 2004).

Many common error handling strategies involve two aspects of spoken dialog systems that are not present in Speech Graffiti: conversationality and system- (or at least mixed-) initiative. Systems that strive to be conversational generally include a variety of error handling strategies, in order to mimic human-human communication and to avoid potentially boring repetition in the human-computer interaction. Thus, the same error conditions may generate different error-handling prompts on different occasions. Systems that support system- or mixed-initiative interactions can handle errors by wording their prompts to elicit specific information (*e.g.,* I'm sorry, on

what day did you say you wanted to travel?). They can also design their prompts to include implicit confirmations (*e.g.,* OK, when do you want to fly to *San Diego*?)

In contrast to the strategies employed by more conversational systems, the aim of this research is to create simple, domain-independent strategies that increase overall interaction efficiency. These strategies will not be concerned with merely correcting individual errors and moving on, but with helping users understand how to interact better with the system. Therefore, in this environment, explicit confirmations and regular interaction structures are more important than variety or naturalness.

2.4 Shaping and help

Nearly all human-computer interfaces incorporate some kind of help facility, which can be categorized from the system's point of view as either passive or active (Fischer, Lemke, & Schwab, 1985). Speech Graffiti has always included passive help. Since the shaping strategies implemented in this work could be considered a form of active help, research in that area is reviewed here.

Passive help is most commonly available: a user, realizing that he or she has made an error or is not sure how to perform some action, explicitly says **`help`** (or clicks a button in a physical or visual interface), and the system responds with some presumably informative help prompt. This prompt could be interactive, asking the user a series of questions to determine more precisely what the problem is; it could provide information based on what the system automatically knows about the task context and the current system state; or it could provide a general help prompt that may or may not address the user's specific problem.

Active help is provided automatically when the system determines that there is a problem with the interaction. Active help has also been variously described as knowledge-based (Fischer et al., 1985), advice-giving (Carroll & McKendree, 1987),

intelligent (Hockey et al., 2003), or targeted (Gorrell, Lewin, & Rayner, 2002) help, since in order for the system to provide useful information in these cases it must somehow form a hypothesis about what the specific problem is, without explicit input from the user. Perhaps the most well-known example of active help was Microsoft's Office Assistant ("Clippy"), but evidence suggests that its well-intended interventions were not always viewed as helpful by users (Swartz, 2003). Carroll and McKendree (1987) present a thorough discussion of research and design issues for advice-giving expert systems and note that active help offers a potentially powerful strategy for managing the tradeoff between learning a system and working on a task; ideally, systems should allow users to do both at the same time.

Most work on active help has been conducted in the area of text-based systems with visual components. When it appears in spoken dialog systems, active help is generally fused with a system's error handling strategy; the help is provided as a way to recover from dialog errors that have been identified. Two approaches to active help in spoken dialog systems are presented in Gorrell et al. (2002) and Hockey et al. (2003).

Gorrell et al. created a system to provide targeted help in the context of a mostly user-initiative, natural language system for device control in a home. This help system was based on the use of two language models: a grammar-based model for general use, and a statistical language model (SLM) for when the grammar-based model failed or had very low confidence. When a back-off to the SLM was triggered, a decision tree was used to classify the utterance as to what user was most likely trying to do. A targeted help message was then delivered based on the decision tree result, which could be one of twelve classifications. The help messages generally took the form of To do X, try saying Y. In a user study comparing targeted help to a control help condition (in which system non-understandings first generated a Sorry, try again message and then a short, standard help message on consecutive non-understandings) the targeted help system generated significantly lower word-error

rates overall and in the first five utterances, and significantly higher grammaticality rates.

Hockey et al. created a similar intelligent help system, differing from Gorrell et al.'s approach mainly in its classification strategy. When back-off to the SLM recognizer was triggered in this system, a targeted help message was created if the SLM result was not parsable by the dialog system. This help message comprised one or more of the following components:

A. a report of the SLM recognition hypothesis

B. a description of the problem with the user's utterance

C. a similar in-coverage example to suggest what the user might say instead.

A hand-built, rule-based system was used to determine the exact content of parts B and C. In constructing part C, the system tried to use words and the dialog-move type (*e.g.,* wh-question, yes/no question, answer or command) from the user's original utterance. In a user study comparing their targeted help to a no-help condition, Hockey et al. found that significantly fewer targeted help users gave up on tasks and that targeted help had a positive effect on task completion times. However, it is not clear if user-initiated (passive) help was available at all or used by participants in either condition. As in Gorrell et al.'s work, this research did not investigate the effects of targeted help activation over time, nor did it report on how often the systems provided inappropriate help.

These active help systems are similar to the shaping strategies implemented in this research, but with some key differences. Although the systems in which Gorrell et al. implemented targeted help used limited grammars, they did not necessarily use grammars designed to support transfer to different domains. Furthermore, they did not report any longitudinal user study results, such as how targeted help use changes over time. Therefore it is not clear that their targeted help assists users in learning a

specific interaction *style* as opposed to fixing one-time errors. Also, in both of these systems, the active help was designed to be executed only in the case of system non-understandings, whereas shaping makes use of intelligent help strategies in cases where the user's input *is* understandable, but is not Speech Graffiti. It is also not clear from the experimental results whether there was a particular performance or portability advantage to the decision tree classification strategy vs. the rule-based one. Finally, neither group systematically explored the performance implications of the specific content of their targeted help messages.

Chapter 3

Improving User Interaction via Shaping

As originally envisioned for this research, the implementation of shaping strategies has three main components, as depicted in the flow chart in fig. 3.1. First, an expanded grammar (A) allows the system to accept more natural language input than is allowed by the canonical Speech Graffiti language. The hypothesis is that the use of the expanded grammar will reduce training time and allow the system to be more forgiving for novice users, which should increase user satisfaction. Separate language models are be constructed based on the Speech Graffiti and expanded grammars so that each user input is decoded twice. Second, shaping confirmation provides an appropriate system response (B) to non-Speech Graffiti input that is accepted by the expanded grammar. Finally, an error classification and response strategy provides

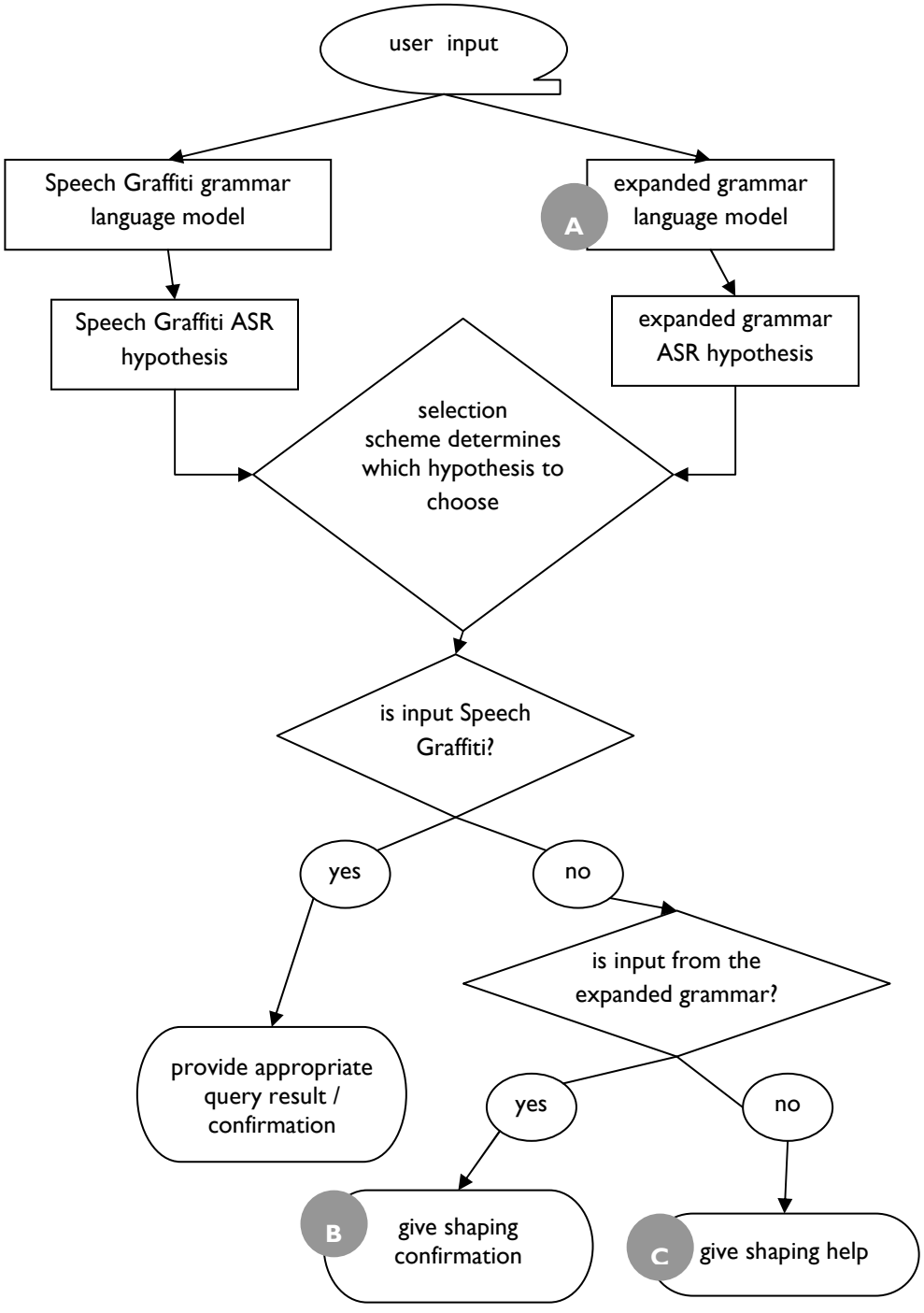

Figure 3.1. Flow chart showing proposed utterance handling process for shaping and its three main components: an expanded grammar (A), shaping confirmation (B), and shaping help (C).

context-appropriate, shaping help (C) in situations in which the recognized input string is accepted by neither the Speech Graffiti nor the expanded grammars. This chapter will describe each component as originally proposed, and later chapters will discuss modifications to the scheme and the performance of each aspect.

3.1 Speech Graffiti

The shaping strategies discussed in this work have been implemented within the framework of the Speech Graffiti system for spoken interaction with information access applications. The Speech Graffiti approach to dialog systems is built on the principles of portability, universality, flexibility, and transparency, and as such offers a system-level attempt at increasing interaction efficiency. As noted in Chapter 1, Speech Graffiti takes a middle-of-the-road approach to handling several common issues that arise with spoken dialog systems.

Speech Graffiti users learn a small set of standard structure rules and keywords that can be used in all Speech Graffiti applications. The structure rules are principles governing the regularities in the interaction, such as *input is always provided in phrases, each conveying a single information element*. Each application designer can specify how flexible the grammar should be for individual phrases unique to an application. Although in theory this could range from a tightly prescribed format to nearly unconstrained natural utterances, the more regular the input format is, the more easily portable the system should be to new domains, from both the developer's and the user's point of view. The current Speech Graffiti applications therefore use a fairly restricted input format in which all phrases must contain either a keyword or both a *slot* element and a *value* element. Phrases can be used to either *specify* constraints (**`[slot] is [value]`**) or to query a slot (**`what is [slot]?`**), and multiple phrases can be concatenated in a single utterance.

Table 3.1. Speech Graffiti keyword summary.

Keyword	Function
repeat	Replays the system's last utterance
more	Lists the next chunk of items from a list
scratch that	Cancels the effect of the user's last utterance
start over	Erases all accumulated context
where was I?	Tersely restates the accumulated context
options	Lists what can be said next at any point
what is…?	Queries the value of a specific slot

The Speech Graffiti keywords (table 3.1) are designed to provide regular mechanisms for performing interaction universals: actions which are performed by users at one time or another in nearly all speech user interfaces (Shriver & Rosenfeld, 2002). The set of universals addressed by Speech Graffiti was derived by analyzing several domains and application categories prior to developing the Speech Graffiti vocabulary.

For a given application, the complete Speech Graffiti lexicon comprises these keywords plus a set of domain-specific words corresponding to the application's slots and values.

For example, the Speech Graffiti movie information application (MovieLine) has nine slots: movie titles, ratings, genres, show times, and dates, and theater names, addresses, phone numbers, and neighborhoods. The lexicon for each slot includes a small set of synonyms (*e.g.,* [**area**|**city**|**neighborhood**|**location**]; [**movie**|**title**]). For each slot's values, the size of the lexicon varies widely, and can include standardized value types like times and dates, small enumerated sets (*e.g.,* **G**, **PG**, **PG-13**, **R**, and **NC-17** for movie ratings), and larger enumerated sets like

movie titles. Fig. 3.2 shows an example of a Speech Graffiti MovieLine interaction, while the full Speech Graffiti MovieLine grammar is included as Appendix A.

3.1.1 System architecture

The Speech Graffiti implementation is modular, with its various components residing on multiple machines spanning two platforms (Linux® and Windows NT®). The dialog manager consists of an application-independent Speech Graffiti engine and an application-specific domain manager. The Speech Graffiti engine calls on a Phoenix parser (Ward, 1990), and the domain manager interacts with commercial database packages. These components together constitute a stand-alone, text based version of the system, which can be developed and tested independently of the speech recognition, speech synthesis, and telephony control components. Fig. 3.3 shows the system architecture used throughout this research.

Automatic speech recognition is performed by the CMU Sphinx-II engine (Huang et al., 1993), using acoustic models based on data from previous Speech Graffiti interactions. Statistical language models for the Speech Graffiti system were created using the CMU-Cambridge Toolkit (Clarkson & Rosenfeld, 1997). The language models were built by using the system grammars to generate corpora of 100,000 utterances upon which to base the models. The Good-Turing method was used as the language model discounting strategy. The system's speech output is unit selection-based, limited domain speech synthesis (Black & Lenzo, 2000) using the Festival system (Black, Taylor, & Caley, 1998). All the components are integrated using the Microsoft Visual Basic® development system and a socket interface where needed.

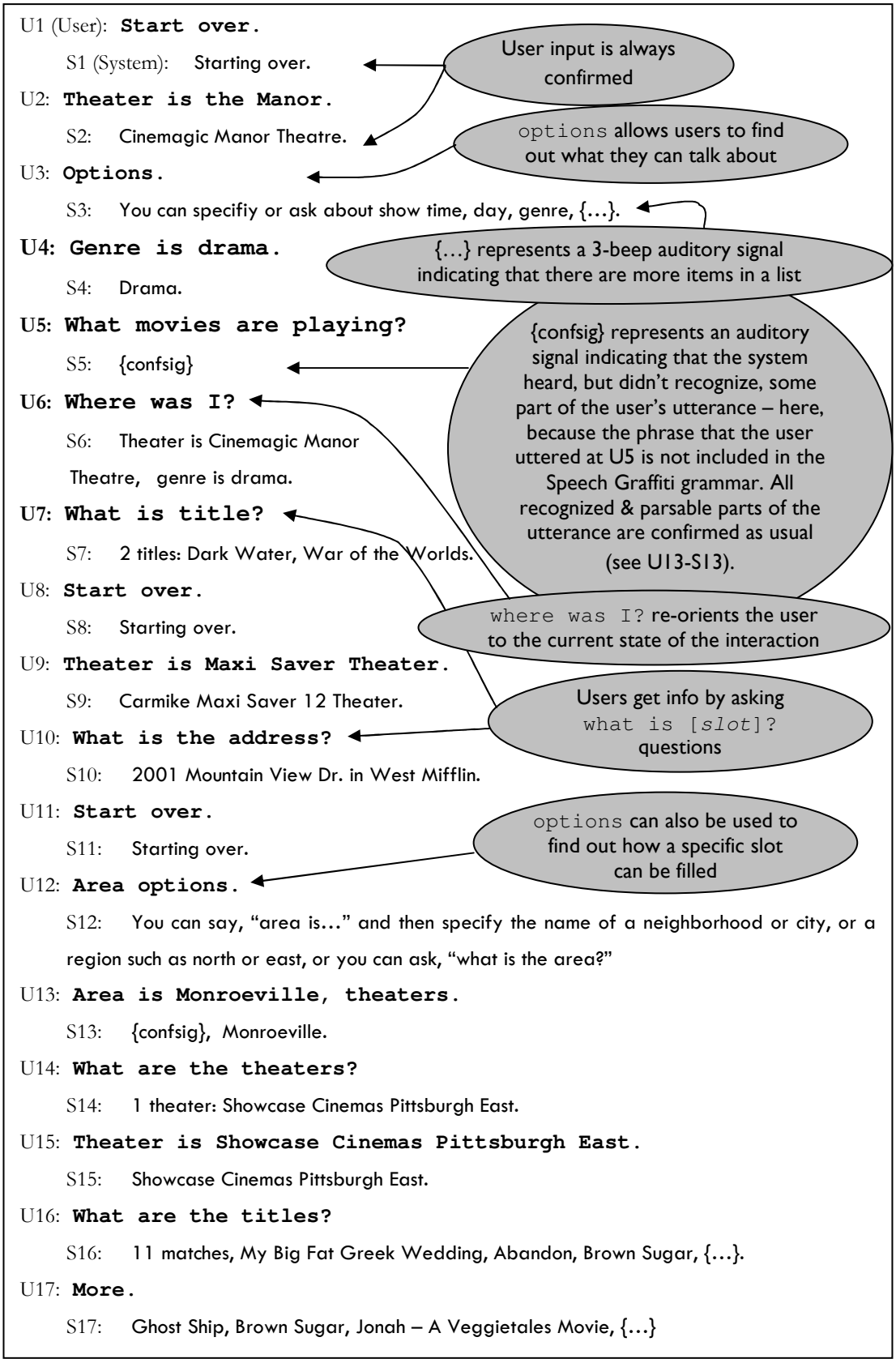

Figure 3.2. Sample Speech Graffiti dialog in the movie domain. This dialog is intended to convey the essence of a Speech Graffiti interaction and the function of the system's keywords. The strategies investigated in this work somewhat alter the input and output structure shown here; examples and discussion of the changes will be provided in later chapters.

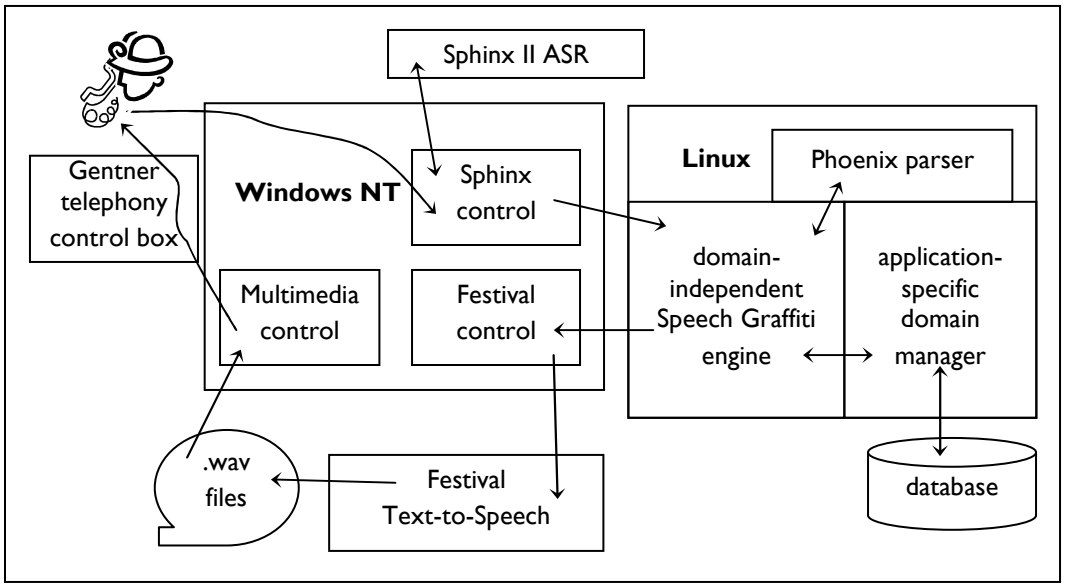

Figure 3.3. Speech Graffiti system architecture.

Movie data for the MovieLine system is stored in an Oracle 9*i*™ database. This database is semi-automatically populated with current information about theaters and movies playing in the Pittsburgh area on a several-times-weekly basis. The update process involves running a script that scrapes relevant data from the Google™ movie search site[3], inserts that data into the database, and then interactively assists the administrator with the process of updating the grammars, dictionaries, speech synthesis, and language model files for the system.

3.2 Expanded grammar

The use of both a more natural language and a restricted language in the same system is intended to allow us to investigate the following research questions: Do users exhibit convergence when interacting with a functional spoken dialog system? Given

[3] http://www.google.com/movies

the option of speaking with more natural language, will users allow their input to be shaped to a more efficient form?

The expanded grammars are designed to take advantage of the phenomenon that users tend to restrict and simplify their speech when they are aware that they are speaking to a system with restricted understanding capabilities (see Section 2.1). This effect has been observed in two studies in the Speech Graffiti lab. First, during the ATUE study, when users spoke to the Speech Graffiti MovieLine but did not actually use Speech Graffiti, they still used a much smaller set of natural language syntactic constructions than users did when speaking to the natural language MovieLine.

To further investigate this effect, a Wizard-of-Oz study was conducted in which users were provided with brief instructions indicating that they should "speak simply" to the telephone-based system (Tomko & Rosenfeld, 2004a). The guidelines for the wizard role were to reject user input that contained any of the following items:

- non-task, conversational words (*e.g.,* **could you tell me…**);
- task-based, non-content items (*i.e.,* those that would be extraneous in a Speech Graffiti slot+value phrase, like **what movies are showing in West Mifflin?**); or
- task-based vocabulary that was not in the current Speech Graffiti versions of these database systems (*e.g.,* using the term **films** instead of **movies**).

If the wizard determined that none of these rejection flags were present, the input was tersely confirmed. Any input containing at least one rejection flag was responded to with a simple non-understanding message, which on consecutive rejections cycled between a) Excuse me?, b) Sorry, I didn't get that, and c) and a replay of the original

"speak simply" instructions. The high overall task completion rate—96%—indicates that users are adept at simplifying their input, since tasks could not be completed without simplified input.

The initial MovieLine expanded grammar constructed for this research handled approximately 85% of the non-Speech Graffiti input spoken to the Speech Graffiti system in the ATUE study and the input collected in the Wizard-of-Oz study described above. This domain-specific grammar was created by hand by one person in about two days, significantly longer than the few hours that it takes to create a Speech Graffiti grammar using the Speech Graffiti web application generator (Toth et al., 2002). For simplicity of processing, I imposed a critical structural limitation to the expanded grammar such that all legal utterances must map linearly to a Speech Graffiti grammar equivalent, as demonstrated in fig. 3.4.

The expanded grammar allows both task-related, non-slot, non-value words (like **playing** and **showing**) and non-task, conversational words and phrases (like

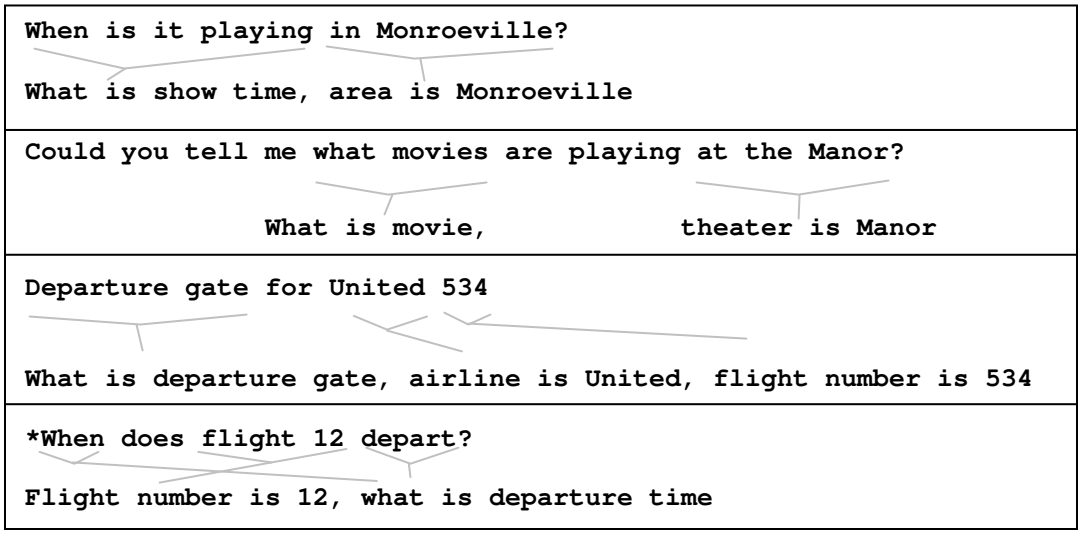

Figure 3.4. Sample expanded grammar utterances in the movie and flight information domains and their Speech Graffiti grammar equivalents. The final example is not allowed by the expanded grammar, since its components do not map in a strictly linear manner to a target language input.

please and **could you tell me**). It allows anaphoric references, but does not actually resolve them. Although the expanded MovieLine grammar contains only about 8% more words than the Speech Graffiti MovieLine grammar, it allows more syntactic variation. The expanded grammar also introduces ambiguities that are not present in the Speech Graffiti grammar. For instance, in the movie domain "Squirrel Hill" is both the name of a theater and of a neighborhood. Thus an input string consisting of only **Squirrel Hill** will be interpreted as either "neighborhood = Squirrel Hill" or "theater = Squirrel Hill" (the choice depending on the convention of the parser), making an efficiency-reducing correction step necessary in some cases. The full expanded MovieLine grammar is included as Appendix B.

3.2.1 Hypothesis selection

For users to realize the greatest advantage from using Speech Graffiti, the Speech Graffiti and expanded grammars should be used to construct separate language models so that Speech Graffiti input can benefit from having a smaller language model (*i.e.,* the smaller, Speech Graffiti language model should have lower perplexity and thus lower word-error rates). Therefore, I created a system in which each user utterance makes two passes through the Sphinx automatic speech recognition (ASR) decoder, once for each language model. This two-pass system then selects one of the hypotheses to respond to.

The baseline hypothesis selection scheme simply chooses the hypothesis with the best decoder score from each of the two passes. In preliminary testing, this approach was found to provide marginally better word-error rates (by about 4%) compared to always choosing the Speech Graffiti hypothesis, always choosing the expanded hypothesis, or choosing a hypothesis at random. However, with the two-pass system, there are now *two* hypotheses to choose from rather than one, and in preliminary tests the correct hypothesis was available as one of the two options in the two-pass system in about 30% more cases than when only one ASR pass was used. Therefore,

in the two-pass case, additional selection heuristics can be added to improve performance relative to a single-pass strategy. Based on preliminary test data, the following heuristics were added to the hypothesis selection process:

- If the selected hypothesis has no parsable phrases, use the alternate hypothesis.
- If the selected hypothesis is only semi-parsable, compare the number of parsable words in the selected and alternative hypotheses and use the one with the higher number of parsable words. If they are equal, use the originally selected hypothesis.
- If the alternative hypothesis is under consideration (due to the unparsability of the original selection), and it comes from the expanded grammar model and contains multiple, identical query phases, simply issue a {confsig} error beep rather than using either hypothesis.

3.3 Shaping confirmation

The goal of shaping confirmation is to handle user input that is accepted by the expanded grammar, but not by the Speech Graffiti grammar, in a way that balances current task success and future interaction efficiency. As previously noted, one problem with natural language spoken dialog systems is that it can be difficult for users to perceive the boundaries of what those systems understand, making it all too easy for users to inadvertently speak outside the grammar or outside the conceptual or functional limits of the system. The design of the shaping confirmation should be such that when the system is recognizing expanded language input, it is actively shaping users towards the target, Speech Graffiti language.

For the initial evaluation, I implemented a baseline shaping confirmation strategy that simply confirmed all expanded-language user input with the Speech Graffiti

equivalent (this was also the same confirmation users received when they successfully used Speech Graffiti). This implicit strategy was intended to serve the dual purposes of grounding interactions via explicit confirmation and shaping users to speak in a more efficient way. The baseline shaping strategy is further described in Section 4.2.

The choice to simply confirm expanded grammar with the Speech Graffiti equivalent, but to not give any other instruction, was based on the theories and results discussed in Section 2.2. Speech Accommodation Theory and Pickering and Garrod's work (2004) predict that participants in human-human interaction will adapt their speech on various levels to match that of their conversational partner. Several studies have shown evidence of adaptation in human-computer interaction as well. I was interested in whether simple, dialog-based exposure to implicit confirmations in the Speech Graffiti-language format would influence users to adapt their input style to match the target, Speech Graffiti language.

3.4 Shaping help

Shaping help is intended to provide assistance to users in cases in which their input is parsable by neither the Speech Graffiti nor the expanded grammars. The original intent was to use features from both the interaction context and the ASR hypothesis to generate a prompt that might help the user say the right thing in the future. Because I was uncertain about what such unparsable input might actually look like and where and how frequently it might occur, I planned to implement this component later on in the research program, based on analysis of the interactions from the first user study. For the initial version therefore, the system simply responded to input that was not parsable by either grammar with a {confsig} error beep.

3.5 Evaluation plan

The core of this research program is a series of three user studies designed to evaluate shaping strategies within functional spoken dialog systems. User Study I evaluates the baseline shaping strategy compared to the original, non-shaping Speech Graffiti system. User Study II evaluates the baseline shaping strategy in comparison to two more-explicit shaping strategies. The results from Study II suggested the potential of an adaptive shaping approach, and User Study III evaluates such an approach in comparison with one of the more-explicit approaches and in the contexts of longer-term and cross-domain use.

Chapter 4

User Study I Design: Baseline vs. Simple Shaping

For the first evaluation of shaping in Speech Graffiti, a user study was conducted to compare the baseline, implicit shaping confirmation strategy to the original, non-shaping Speech Graffiti system that was used in the ATUE study. This evaluation had two goals:

- to determine the effectiveness of the simple, implicit shaping strategy on user input and interaction efficiency, and
- to collect a corpus of interactions to inform further decisions about shaping confirmation and help strategies.

4.1 Participants

Thirty-three adults participated in the study. They were recruited from the neighborhood around Carnegie Mellon University and the University of Pittsburgh via small public signs and by postings on Carnegie Mellon's electronic bulletin boards. After all the study sessions were completed, data from four participants was discarded. One participant's age was outside of (above) the study range; one was determined to be a native speaker of a non-American variety of English (West Indian); one participant abandoned the experiment early without attempting all the tasks; and one had an extreme amount of line noise on her telephone connection during the session. This left a total of 15 male and 14 female participants, all of whom were native speakers of American English and who were between the ages of 23 and 54. All of the participants reported having completed at least some undergraduate coursework, and about half of the participants had received graduate degrees or completed some graduate coursework, as summarized in table 4.1.

Based on the results from the ATUE study, I specifically tried to recruit participants for this study who were more representative of "average" adults. Thus, in this experiment, no participants had significant experience with computer programming (as measured by a reply of *never* or *rarely* to the survey question *I do computer programming: never - rarely - sometimes - often*), and all were new to the Speech Graffiti interface. About two-thirds of the participants reported using telephone-based information services five times a month or less.

4.2 Setup

A between-subjects experiment was designed in which participants were randomly assigned to one of two basic conditions: original or shaping. Participants in the shaping condition were then further divided into two sub-groups: tutorial and no-tutorial. Due to the necessity of pre-use training in the original Speech Graffiti

Table 4.1. Selected demographic characteristics of participants (N = 29) in User Study I.

Characteristic	*N*	%
Age at time of survey (years)		
20-24	1	3.4
25-34	19	65.6
35-44	5	17.2
45-54	4	13.8
Highest education level completed		
Some high school or less	0	0
High school graduate	0	0
Some college	5	17.2
2-year college/technical school	0	0
4-year college	8	27.6
Some postgraduate work	7	24.1
Postgraduate degree	9	31.0
Reported frequency of computer programming		
Never	24	82.8
Rarely	5	17.2
Fairly Often	0	0
Very Frequently	0	0

system, all users in the original condition received a tutorial. This resulted in a total of three sub-groups: original+tutorial (N = 9), shaping+tutorial (N = 9), or shaping+no-tutorial (N = 11).

Participants in the tutorial subgroups were provided with a self-guided, nine-slide PowerPoint® presentation. This tutorial informed users about the input structure of canonical Speech Graffiti, about the system's confirmation strategy, and about how to navigate lists and correct errors. It also provided a few hints (*e.g.,* speak in a normal voice; remember to use **`options`** and **`start over`**). The tutorials for both the original and shaping conditions were identical with the exception of the system confirmation strategies depicted. Both tutorials included short audio

examples for participants to listen to as they worked through the tutorial. Tutorial sub-group participants were given five minutes to work on the tutorial. Most users (N = 14) in the tutorial sub-groups finished before the end of the five minute session. When users in the tutorial sub-groups called the system, they heard the following brief introduction: Welcome to the CMU MovieLine. Remember to use Speech Graffiti when you're talking to the system. Users in the no-tutorial sub-group heard the following, more descriptive introduction:

> Welcome to the CMU MovieLine! The Speech Graffiti system only understands very simple English, so speak to it as simply as you can. The system understands only keywords, and not the structure of sentences, so it will understand you best if you speak in the format "something is something." For instance, you might say "movie is Star Wars," or "theater is the Waterfront," and then ask a simple question like, "what are show times?" or "what is the title?" When you hear this sound, {…}, it means there are more items in a list and you can say "more" to hear them. When you hear this sound, {confsig}, it means the system heard you say something but it didn't understand everything. In that case you might need to repeat your input. To clear everything and start from the beginning, you can say, "start over". If you need to erase just the last thing you said, say "scratch that." To find out what you can say at any point, say "options." You should now be ready to start using the system. If you need help at any point, just say "help."

To ensure that all users in the no-tutorial sub-group heard the same information, barge-in was turned off for the introductory prompt. The introduction could not be repeated after the initial playing.

In the original condition, the system provided confirmations that were terse, value-only restatements of user input, as implemented in the ATUE version of Speech Graffiti version. The shaping condition differed from the original by accepting input from the expanded grammar and providing a full, Speech Graffiti-style slot+value restatement as confirmation. Fig. 4.1 shows a few examples contrasting the two confirmation strategies. Note that when non-Speech Graffiti input is provided in the original condition, the system gives the {confsig} error signal. The system also plays a {confsig} in both conditions in response to input that is parsable by neither grammar

sample Original interaction	sample Shaping interaction
Theater Manor, genre is comedy	**Theater Manor, genre is comedy**
Cinemagic Manor Theatre, comedy	Theater is Cinemagic Manor Theatre, genre is comedy
What are movies?	**What are movies?**
3 matches: Friends with Money, Thank You for Smoking, Tsotsi	Requesting movie. 3 matches: Friends with Money, Thank You for Smoking, Tsotsi
Galleria	**Galleria**
{confsig}	Theater is Carmike Galleria 6
Theater is Galleria	**Theater is Galleria**
Carmike Galleria 6	Theater is Carmike Galleria 6
Genre is drama, what's playing?	**Genre is drama, what's playing?**
{confsig}, drama	Genre is drama, requesting movie. Sorry, there are no matches.
Where was I?	**Where was I?**
Theater is Carmike Galleria 6, genre is drama	Theater is Carmike Galleria 6, genre is drama, what is movie?
List	**List**
{confsig}	{confsig}

Figure 4.1. Sample interaction showing system response differences between original and shaping conditions in User Study I.

(*e.g.*, **`list`**). In both systems, the full slot+value versions are provided when the user asks **`where was I?`** Finally, it should be noted that the implicit confirmation in the shaping condition is mostly targeted at specification phrases; expanded grammar query phrases are simply confirmed with `requesting` *slot* before providing query results.

Participants completed the study either in a cubicle in a quiet office at Carnegie Mellon University or in a small conference room in the Biomedical Science Tower at the University of Pittsburgh. In both environments, participants were seated at a desk and interacted with the system over a standard, land-line, office telephone. The audio was recorded over the telephone for all sessions. During the study, the experimenter monitored and recorded the audio portion of the interactions while in the room, but from out of sight of the user.

4.3 Tasks

The domain used in this study was movie information. Participants were asked to complete a series of 15 tasks using the MovieLine system in one of the two basic conditions, original or shaping. The 15 tasks were categorized into four difficulty levels, which were determined by the number of slots that needed to be specified or queried. The four difficulty levels and an example of each are shown in table 4.2.

The order of the tasks presented to users followed a consistent task difficulty order, but the individual tasks at some difficulty levels were varied to reduce order effects. Since the MovieLine is a functional system providing up-to-date movie information and the user sessions were scheduled over a three-week period, some of the movie titles and theaters in the tasks changed over the course of the study. However, theater and title combinations for each task were chosen so that a consistent number of results was always obtained for the same task.

Users were given the complete list of tasks on a sheet of paper and were asked to work through them in order, writing down the answers for each. Tasks were written in a format designed to encourage users to use their own words when speaking to the system rather than simply to repeat what was written on the page. See Appendix C for a representative list of tasks used in the study.

Participants were given forty minutes to work through the set of tasks. Participants were instructed that if they had time at the end of their session, they could go back and work on any tasks that they had abandoned earlier or for which they were not sure they had found the correct information. Twenty users declared that they were finished with the tasks before their forty minutes were up (six in the original condition and 14 in the shaping condition); the remaining nine participants were asked to stop working on the tasks when the forty minutes expired. Only one participant was not able to at least attempt all 15 tasks during the forty minutes. This participant's data was nonetheless included in the study since she had worked

Table 4.2. User study task difficulty levels, as defined by number of specification and query phrases required for each.

Task difficulty level	Number of specification phrases	Number of query phrases	Speech Graffiti example
1	1	1	**Movie is Madagascar, what is theater**
2	2	1	**Theater is Galleria Six, genre is comedy, what are titles**
3	1	2	**Theater is Carmike Ten, what is address, what is phone number**
4	2	2	**Theater is Southside Works, genre is romance, what are titles, what are show times**

earnestly at the tasks for forty minutes (in contrast to the participant whose data was discarded since he abandoned the study early without attempting all of the tasks).

To motivate users to complete tasks successfully, participants were compensated for their time with a flat cash payment for participation ($12.50) plus an additional amount (50 cents) per correctly completed task.

4.4 User survey

Upon declaring that they were finished, or at the end of the forty minute session, users were asked to complete an evaluation questionnaire. This questionnaire was based on Hone and Graham's (2000) Subjective Assessment of Speech System Interfaces (SASSI) project, which grouped a number of subjective user satisfaction statements into related sub-scales via factor analysis. Hone and Graham reported an internal reliability of 0.69-0.90 (Cronbach's alpha) for the resulting six factors. The SASSI procedure has recently been used by other researchers for spoken dialog system evaluation (*e.g.,* González-Ferreras & Cardeñoso-Payo, 2005; Howell, Love, & Turner, 2005; Hakulinen, Turunen, & Räihä, 2006), although it has not been widely adopted. However, as there is in fact no generally accepted gold standard for the subjective assessment of spoken dialog systems, I chose to base my evaluations on the SASSI procedure as it offers some measure of experimental validity.

In addition to the 34 SASSI statements (table 4.3), two statements were added to the questionnaire to assess the quality of the system's text-to-speech output (*I had trouble understanding the system* and *It was easy to understand what the system said.*). Participants scored each item on a 7-point Likert scale, from "strongly disagree (1)" to "strongly agree (7)." The statements marked with asterisks in table 4.3 are reversal items: negative statements whose values are converted to the opposite end of the scale for analysis purposes (*e.g.,* a 7 rating is converted to a 1, a 6 rating is converted to a 2, etc.), so that high scores in all categories are considered positive.

Table 4.3. SASSI subjective user satisfaction factors and their component statements (after Hone & Graham, 2000).

Factor	Component statements
System Response Accuracy	The system is accurate
	The system is dependable
	The system makes few errors
	The interaction with the system is consistent
	The interaction with the system is efficient
	*The system is unreliable
	*The interaction with the system is unpredictable
	*The system didn't always do what I wanted
	*The system didn't always do what I expected
Likeability	The system is useful
	The system is pleasant
	The system is friendly
	I was able to recover easily from errors
	I enjoyed using the system
	It is clear how to speak to the system
	It is easy to learn how to use the system
	I would use this system
	I felt in control of the interaction with the system
Cognitive Demand	I felt confident using the system
	I felt calm using the system
	The system is easy to use
	*I felt tense using the system
	*A high level of concentration is required when using the system
Annoyance	*The interaction with the system is repetitive
	*The interaction with the system is boring
	*The interaction with the system is irritating
	*The interaction with the system is frustrating
	*The system is too inflexible
Habitability	*I sometimes wondered if I was using the right word
	*I was not always sure what the system was doing
	*It is easy to lose track of where you are in an interaction with the system
	I always knew what to say to the system
Speed	The interaction with the system is fast
	*The system responds too slowly

Note. Starred statements are reversal items for analysis purposes.

what they would change about it. Participants were then debriefed on the purpose of the experiment and were given the opportunity to ask questions about their experience. Finally, they were compensated for their time with the appropriate cash payment.

4.5 Analysis

An efficient modality should be effective, fast, satisfying, and easy to learn. For the purposes of this research, a more efficient interaction was operationalized as one in which users completed more tasks, in less time, with increased user satisfaction, and with minimal up-front training time.

Thus, I computed overall task completion and mean time- and turns-to-completion rates. I calculated ASR word- and concept-error rates, and I computed mean scores for each of the six user satisfaction factors and for a combined, overall user satisfaction rating[4]. The presence of the tutorial and no-tutorial sub-groups in the shaping condition allowed for the analysis of the effect of up-front training time.

Throughout this work, statistical analyses were performed with JMP IN™ version 5.1.2 software[5]. An α-level of 0.05 was used for all tests of significance, and all *t*-tests of means reported are two-sided unless otherwise noted.

---

[4] The use of mean scores for the analysis of Likert scale data follows claims in Jaccard and Wan (1996) that this approach does not tend to generate substantial Type I (claiming an effect where there is none) and Type II (failing to reject the null hypothesis when there may be an effect) errors.

[5] Copyright 1989-2004, SAS Institute Inc.

Chapter 5

User Study I Results: Baseline vs. Simple Shaping

The results from Study I suggested a slight trend toward increased interaction efficiency in the shaping condition. The study results also confirmed the effectiveness of the two-pass ASR method and suggested that a pre-use tutorial is not necessary. Finally, the data collected from this study was valuable in suggesting changes to the system, particularly as to what types of shaping confirmation could be more effective and what type of shaping help could be provided.

5.1 Efficiency measures

5.1.1 Task tagging procedures

A task was tagged as correctly completed if the user gave input that triggered the presentation of the correct query result for a given task. In a few cases, due to a bug in the system, the system was unable to find matching results for a particular query. In these cases, tasks were judged to be complete when users gave input that would have produced the correct result if there had been no bug. Tasks were counted as completed even if the user worked on intermediate tasks before retrieving a correct response from the system. Fig. 5.1 presents a slightly modified excerpt of a Study I participant's interaction that is intended to clarify tagging issues. In this example, tasks B and C are tagged as complete because the user has eventually found the correct result. Task A is incomplete.

Time-on-task was calculated in seconds from the start of a user's first utterance directed towards a task to the end of the system's delivery of the correct response for that task, or to the end of the system's response to the user's final utterance for that task, whichever came first. If a user made multiple attempts at a task, the time for all attempts were added together (up to the point of task completion, if appropriate). If the user barged-in on the final system response, end-of-task time was marked at the start of the user's barge-in utterance. Keywords such as **`start over`**, **`scratch that`**, **`help`**, and **`options`** were included at the start of new tasks rather than at the end of prior tasks. While the system was reporting query results, user utterances of **`repeat`** (and the corresponding system reply) were not included in the time calculation, as it was assumed that asking the system to repeat several times could be an artifact of the experimental setup (in which the user was asked to write down the answers). Finally, once the user retrieved the correct information for a task, any further utterances on that task were not counted in the time for that task.

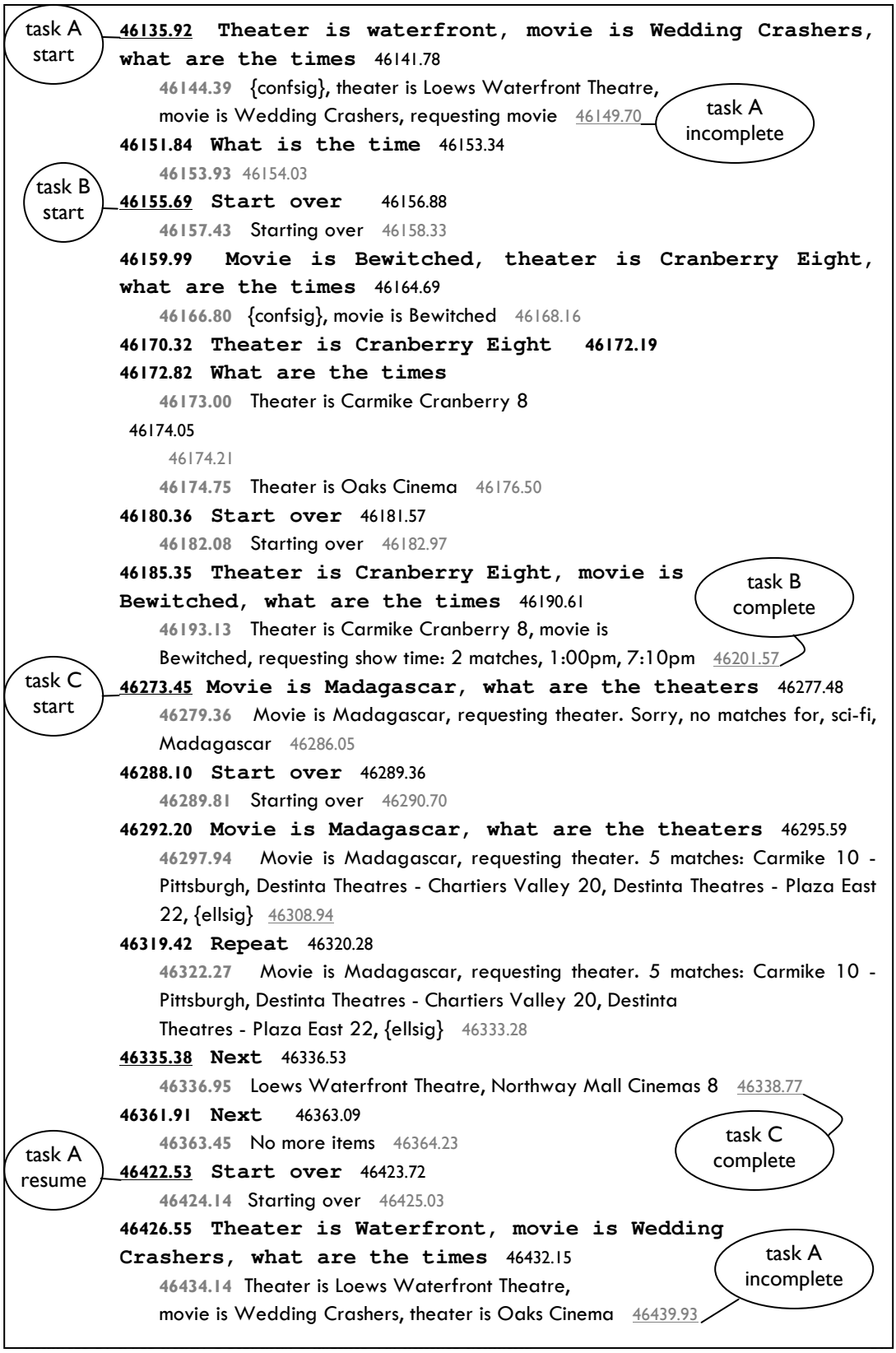

Figure 5.1. Slightly modified user-system interaction intended to clarify task tagging issues. Numbers indicate the system time at each turn. User input timings are in black; system output timings are in gray. Bold numbers indicate the time at the beginning of each user input or system output; non-bold numbers indicate the end times for these. Underlined numbers indicate key times used in the calculation of time-on-task.

Thus, in fig. 5.1, time-on-task for task A is 31.18 seconds (*[46149.7 - 46135.92]* + *[46439.93 - 46422.53]*). The second user utterance, **what is the time**, is not counted in the overall time since the system did not respond to it. For task B, time-on-task is 45.88 seconds (*46439.93 - 46422.53*) and for task C, 38.88 seconds (*[46308.94 - 46273.45]* + *[46338.77 - 46335.38]*). Note that for task C, the time for the user's **repeat** utterance and the corresponding system response are excluded from the overall task time. The user's final **next** utterance for task C is also excluded, as the user has already heard the complete, correct answer by that point.

Turns-on-task were measured as the number of user-utterance-plus-system-response pairs that were made for each task. To be considered as a turn, a user utterance must have contained some content (*i.e.,* utterances containing only noise were excluded) and the system must have responded to that utterance. If a user made multiple attempts at a task, the turns for all attempts were added together (up to the point of task completion, if appropriate).

Thus, in fig. 5.1, turns-on-task for task A is 3. As with time-on-task, the user's utterance **what is the time** is not included in the turn count since the system did not respond to it. Turns-on-task for task B is 6. Although the user began to ask **what are the times** (the fourth utterance of task B) before the system responded to the previous utterance, the two utterances are counted as separate turns since the system did begin to reply to the previous utterance before the end of the user's **what are the times** utterance. Turns-on-task for task C is 4. As in the time-on-task assessment, the user's **repeat** and final **next** utterances are excluded from the turns calculation.

I used the above protocol to calculate mean time-to-completion and mean turns-to-completion for each participant's completed tasks. However, such an analysis ignores the effect of incomplete tasks. If a user does not complete three out of four tasks,

but does complete the fourth task in just two turns, that turns-to-completion rate of two turns surely overestimates her actual interaction performance with the system. In order to take incomplete tasks somewhat into consideration, I also looked at median time- and turns-on-task for each user, setting (for graphing purposes) each incomplete task equal to 105% of the maximum time and turns encountered in the study (*i.e.*, 423.29 seconds and 48.3 turns).

5.1.2 Task completion

Users in the shaping condition completed an average of 10.6 tasks each (S.D. = 3.83), while users in the original condition completed an average of 8.11 tasks each (S.D. = 5.40). A *t*-test of means did not show a significant difference between the two conditions ($t = -1.25$, $p = 0.24$).

5.1.3 Time

Users spent about the same amount of time on completed tasks in both conditions: an average of 93.0 seconds (S.D. = 32.1) in the shaping condition spent and 96.5 seconds (S.D. = 36.0) in the original condition ($t = 0.25$, $p = 0.81$).

The analysis of participants' median time-on-task, which takes incomplete tasks into account, revealed that users in the shaping condition generally spent less time on tasks (mean, 168.1 seconds; S.D. = 138.54) compared to those in the original condition (mean, 254.1; S.D. = 165.08) ($t = 1.36$; $p = 0.20$).

5.1.4 Turns

Users in the shaping condition used an average of 9.6 turns (S.D. = 3.32) on each completed task, while users in the original condition used an average of 10.5 turns (S.D. = 2.93). A *t*-test of means did not show a significant difference between the two conditions ($t = 0.73$, $p = 0.47$). Median turns-on-task were also generally lower

for users in the shaping condition (mean, 18.6; S.D. = 16.1) than in the original condition (mean, 28.4; S.D. = 19.3) ($t = 1.33$; $p = 0.21$).

5.1.5 Effect of tutorial

Within the shaping condition, I analyzed efficiency measures based on whether the participants did or did not receive the pre-use tutorial. Across the board, there were virtually no differences in efficiency measures between the two groups, as shown in table 5.1.

Table 5.1. Summary of efficiency results between shaping sub-groups with and without tutorial in Study I.

	No tutorial		Tutorial			
Efficiency measure	M	SD	M	SD	*t*	*p*
Tasks completed	10.6	3.35	10.5	4.56	0.04	0.97
Time-to-completion	89.3	35.2	97.5	29.3	-0.57	0.58
Turns-to-completion	8.60	2.27	10.9	4.06	-1.51	0.16
Median time-on-task	178.0	132.0	156.0	153.3	0.34	0.74
Median turns-on-task	20.1	15.1	16.7	18.0	0.44	0.67

5.2 User satisfaction

On average, users in the shaping condition gave the system higher ratings in each of the six user satisfaction areas compared to users in the original condition (fig. 5.2), although the scores were not significantly different for the two conditions. For users in both conditions, the text-to-speech and speed factors had the highest mean user satisfaction scores. The habitability factor received the lowest mean scores in both conditions; it was also the factor with the strongest difference between the two groups ($t = -1.41$, $p = 0.17$).

5.2.1 Effect of tutorial

Figure 5.3 shows a comparison of the user satisfaction ratings between the tutorial and no-tutorial sub-groups in the shaping condition. Again, I did not find significant differences between the sub-groups.

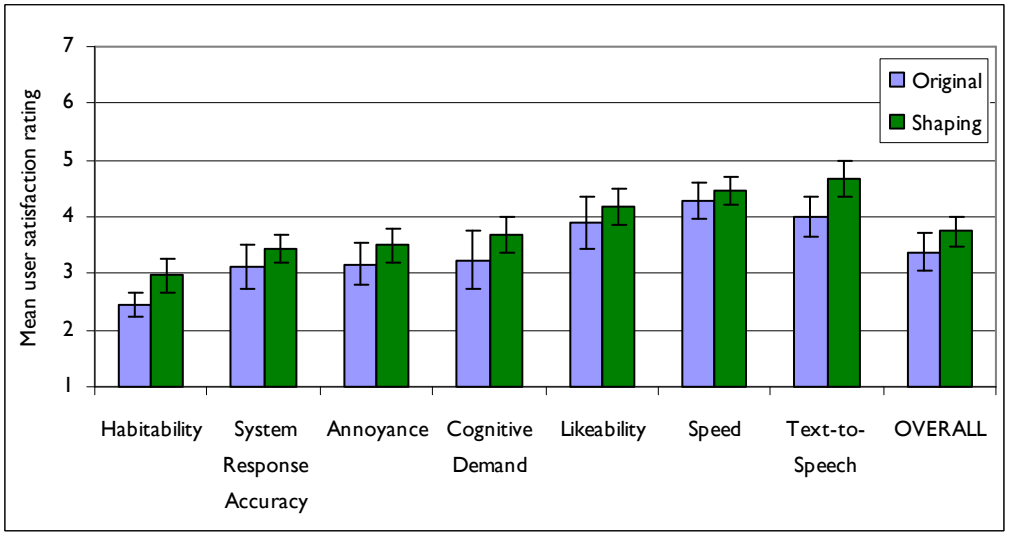

Figure 5.2. Mean user satisfaction ratings from Study I for each of the seven user satisfaction factors, and combined as an overall rating.

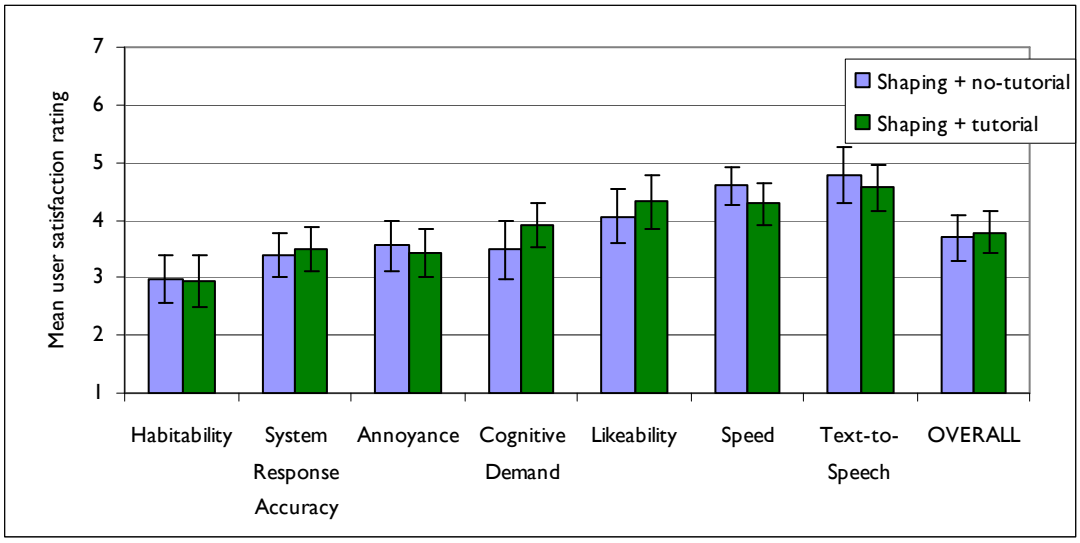

Figure 5.3. Mean user satisfaction ratings for scores from Study I for each of the seven user satisfaction factors, and combined as an overall rating, for users in the no-tutorial and tutorial sub-groups of the shaping condition.

5.3 Grammaticality

One of the key measures in this research is how often users say something that falls within the system's capacity to understand. I refer to this as *grammaticality*, and measure it as the percentage of user utterances that are fully parsable by a given grammar.

Results from the ATUE study suggested that participants who interact with the Speech Graffiti system with at least 80% grammaticality have successful interactions with the system. In this study, only six participants achieved that level: five in the shaping condition and one in the original condition. All six of these participants completed more tasks (at least 11 each) than the whole-study mean (9.8), and all had received the pre-use tutorial.

Grammaticality was virtually identical across conditions. Users in the shaping condition had an average grammaticality of 63.8% (S.D. = 16.3), while users in the original condition had an average grammaticality of 64.6% (S.D. = 16.3; $t = 0.13$, $p = 0.90$).

5.3.1 Effect of tutorial

On average, grammaticality was somewhat higher in the shaping+tutorial condition than in the shaping+no-tutorial condition ($t = 1.83$, $p = 0.09$). Users in the tutorial sub-group had an average grammaticality of 70.9% (S.D. = 17.7), while users in the no-tutorial sub-group had an average grammaticality of 58.0% (S.D. = 13.1).

5.3.2 Intra-session grammaticality

As a key to assessing the effectiveness of shaping, I also analyzed how grammaticality changed over the course of each user's interaction. If users are converging to the Speech Graffiti format, then they should exhibit more grammaticality as interactions

progress. To measure this, I calculated each participant's grammaticality level in the first quarter of their interaction and compared that to his grammaticality level in the final quarter.

As shown in table 5.2, users in both conditions showed significant within-subject increases in grammaticality from the initial quarter to the final quarter. Testing differences across the groups showed that the change in grammaticality was significantly steeper for the original group than for the shaping group ($F = 4.70$, $p = 0.04$).

Table 5.2. Intrasession grammaticality changes for users in Study I.

	Quarter					
Condition	Initial	Final	Mean difference	StdErr	*t*	*p*
Original	51.4	74.9	23.5	5.69	4.13	0.003
Shaping	58.1	67.7	9.58	3.46	2.77	0.01

5.4 System performance

User Study I generated a corpus of 5,508 utterances, 3,731 (68%) of which were from participants in the shaping condition, with the remaining 1,777 (32%) from users in the original condition. Prior to further analysis, the transcriptions of these utterances were cleaned of non-task items such as noise, system feed, and off-task user comments. After the cleaning process, any utterances that had contained only non-task items were retained in the corpus as empty utterances.

5.4.1 What to analyze?

The standard measure of ASR performance is word-error rate (WER), which for a given utterance is calculated as

$$\frac{hypothesis\_insertions + hypothesis\_deletions + hypothesis\_substitutions}{reference\_words}$$

For spoken dialog systems, WER does not always give an accurate representation of the error that a user experiences. For instance, in the Speech Graffiti MovieLine, imagine that a user says **`movie is Crash`**, but the system recognizes this as `movies are Crash`. The WER for this utterance is 67%, but the concept being processed by the system ("movie = Crash") still matches the user's intent. If the system had instead recognized `movie is Bewitched`, the WER would only be 33%, but the new concept ("movie = Bewitched") does not match the user's intent. Thus, as suggested by Boros et al. (1996), concept-error rates may be a more appropriate measure of system performance. However, since WER is such a standard ASR evaluation measure, I report both numbers here for comparisons with other spoken dialog systems.

The Speech Graffiti grammar exploits the highly structured nature of the language by treating common Speech Graffiti phrases as single words (*e.g.,* **`title=is`**) for language modeling and ASR purposes. Since the expanded grammar is designed to be more flexible, similar phrases are generally treated as separate words. However, multi-word members of common value classes like movie titles and theater names are linked as single words in both systems (*e.g.*, **`Good=Night= and=Good=Luck`**, **`AMC=Loews=Waterfront=Twenty=Two`**). Table 5.3 shows a comparison of the sizes of the lexicons from the two grammars with both linked and unlinked words.

Table 5.3. Comparison of lexicon sizes for Speech Graffiti and expanded MovieLine grammars in Study I. In this table, numbers for both grammars include variations for 49 movie titles.

	Size of MovieLine lexicon	
	With linked words	With all words unlinked
Speech Graffiti	488	371
Expanded	437	391

The linked Speech Graffiti lexicon is larger than the expanded lexicon due to the iteration of all of the slot names in linked form (*e.g.*, **`title`**, **`title=is`**, **`titles=are`**, **`the=title`**, **`the=title=is`**, etc.). When reporting WER results throughout this work, I will clarify whether the calculations were based on linked or unlinked word forms.

5.4.2 Word-error rate

Based on unlinked words, users in the original condition had an average WER of 39.9% (S.D. = 12.5), while users in the shaping condition had an average WER of 36.1% (S.D. = 13.1). For the shaping condition, WER was based on the final hypothesis selected in the two-pass procedure. A *t*-test of means did not show a significant difference between the two conditions ($t = 0.73$, $p = 0.47$).

Another way to analyze ASR performance is to measure WER for grammatical utterances only. When assessed this way, there are no out-of-vocabulary items in the input to confuse the recognizer, so this can be regarded as a baseline assessment of ASR performance. Within the shaping condition, I compared the WER of users' Speech Graffiti-grammatical utterances to the WER of their expanded-grammar-grammatical utterances. The mean WER for Speech Graffiti-grammatical utterances was 25.3% (S.D. = 10.5), while the mean WER for expanded-grammar-grammatical input was 46.8% (S.D. = 20.4). The significant difference between the two rates ($t = 6.14$, $p < 0.001$) supports the approach of encouraging users to speak within the Speech Graffiti grammar, since users will likely experience fewer errors when they speak within that grammar.

5.4.3 Two-pass grammar

This study provided the first opportunity for fully evaluating the two-pass ASR processing method. The two-pass approach generates two hypotheses: a Speech Graffiti hypothesis and an expanded hypothesis. Based on the methods described in

Section 3.2.1, one of these is selected and designated as the preferred hypothesis. For a basic assessment of the effectiveness of the two-pass approach, I compared the WER of the preferred two-pass hypotheses to those of three alternative selection processes:

- always choose the Speech Graffiti hypothesis;
- always choose the expanded hypothesis; or
- randomly choose between the two.

A one-way ANOVA confirmed a significant difference between the WERs generated by the four processes ($F = 16.6$, $p < 0.001$). Further comparison of means for each process showed that the two-pass model generated lower WER compared to the other three models, significantly so except in comparison with the always-choose-Speech Graffiti method (see fig. 5.4).

5.4.4 Concept error

As noted earlier, concept error measures the mismatch between the user's intent and the underlying concepts (semantics) in the speech recognition hypotheses. For this

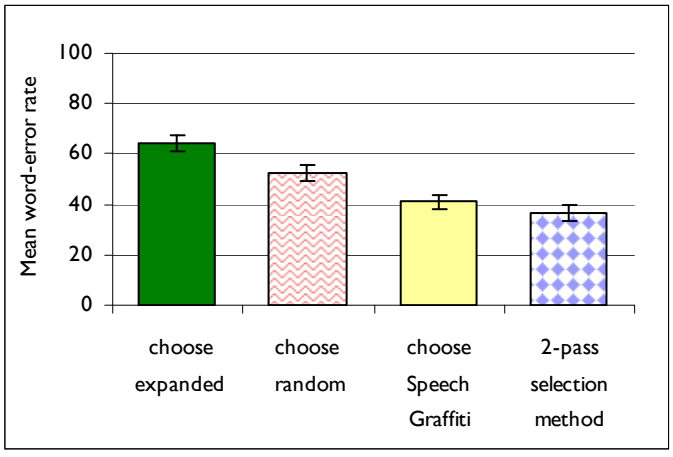

Figure 5.4. Comparison of word-error rates (based on unlinked words) for hypothesis selection options.

analysis, I considered concept error on the basis of the entire user utterance, as a binary decision. To calculate concept error, each user utterance was manually tagged as *concept-correct* or *concept-error*. Since the assessment of user intent is clearly a subjective endeavor, a sample of 5% (275) of the user utterances from Study I, chosen semi-randomly so that roughly the same number of utterances from each participant was included, was tagged by a second evaluator. High interrater agreement scores (Cohen's $\varkappa = 0.95$; Siegel & Castellan's $\varkappa = 0.95$) (Di Eugenio & Glass, 2004) support the validity of the initial tagging.

Concept error was generally lower in the shaping group (mean, 25.6%, S.D. = 10.6), than in the original group (mean, 38.3% S.D. = 16.9; $t = 2.07$, $p = 0.06$).

5.5 Correlations

Finally, I analyzed correlations between the objective and system performance measures and the subjective, user satisfaction scores, as shown in table 5.4. Most measures correlated with user satisfaction at roughly the same strength for both conditions. However, the correlation between Speech Graffiti grammaticality and user satisfaction was much stronger for users in the original condition. Noting that task completion was significantly correlated with user satisfaction with both groups, I also looked at the effect of Speech Graffiti grammaticality on task completion itself.

Table 5.4. Correlations between objective and system performance measures and mean user satisfaction scores in Study I.

	Correlations with overall user satisfaction score			
Objective measure	Original	*p*	Shaping	*p*
Task completion	0.66	0.05	0.61	0.004
Mean turns-to-complete	-0.56	0.12	-0.63	0.003
Mean time-to-complete	-0.65	0.06	-0.68	< 0.001
Speech Graffiti grammaticality	0.79	0.01	0.12	0.62
Word-error rate	-0.65	0.06	-0.50	0.02
Concept-error rate	-0.68	0.04	-0.41	0.07

While there were significant correlations for both of the two groups, the strength of the correlation was much stronger for users in the original group (original: 0.87, R^2 = 0.76, p = 0.002; shaping: 0.51, R^2 = 0.26, p = 0.02)

5.6 Discussion

Although the differences between the groups were generally not statistically significant, the results from this study revealed a slight trend towards increased interaction efficiency for the shaping condition. Mean scores for task completion, median time- and turns-on-task, concept error, and all six user satisfaction factors were better for users in the shaping group. Perhaps most importantly, the lack of significant differences in objective or subjective measures between participants in the shaping+tutorial and shaping+no-tutorial sub-groups suggests that users can interact with the shaping system without a tutorial, which increases the overall interaction efficiency.

Since users were not required to change their input style in order to complete tasks, it would have been possible for users in the shaping group not to exhibit the intrasession grammaticality increases that they did. Thus, I conclude that, as predicted by models of adaptation in human-human conversation, and by some initial results on adaptation in human-computer interaction, there was evidence of convergence in the shaping condition. However, the shaping strategy investigated in Study I did not seem to strongly encourage users to converge to the Speech Graffiti format. Intrasession grammaticality change was actually significantly steeper for users in the original condition, and final quarter grammaticality scores were higher on average (although not significantly) for the original group.

There are a few factors which distinguish User Study I from other research on adaptation, and these differences suggest an explanation for the lack of particularly strong adaptation. Most obviously, this was not a human-human interaction.

Although the system speaks with a very human-like voice, its interaction style is clearly non-natural. It may be violating some conversational maxims by being too terse or "computer-like." Although Pearson et al. (2006) found (lexical) convergence in human-computer interactions (with stronger evidence of convergence with a purportedly more "basic" computer system), it seems that participants in that experiment may have been primed to convergence by the nature of the task, which was a *matching* game. Gustafson, Larsson, Carlson, and Hellman (1997) and Bell (2003) found evidence of syntactic convergence in human-computer interaction, but in directed dialog applications where users responded to system questions. This environment may have more strongly primed them to use the system's style.

Perhaps most critically, in contrast to the participants in the original condition, users in the shaping group were able to complete tasks successfully by using utterances in the expanded grammar. That is, a user utterance such as **`what's playing at the Manor theater?`** generated the same query result information as the Speech Graffiti-grammatical **`Manor theater, what are movies?`**, which may have meant that users had little motivation to adapt their input in what may have already seemed to be an unnatural situation. On the contrary, participants in the original condition actually had *more* motivation to be Speech Graffiti-grammatical, since their input would not be understood and their tasks would not be completed otherwise. The differences in the strength of correlations between Speech Graffiti grammaticality and user satisfaction and task completion for the two groups supports this idea. This result is similar to the findings from the unrestricted condition in Zoltan-Ford's (1991) study, in which users in the unrestricted condition did not model the computer's output as strongly as those in the restricted condition.

One potential motivation for speaking within the Speech Graffiti grammar is the lower word-error rates that can result. In the shaping condition, utterances that were Speech Graffiti-grammatical had significantly lower WER than utterances that were

expanded-language-grammatical. However, although the average per-user difference between the two sets was 21.4 points, it is possible that the differences may not actually have been perceptible to the participants in the study.

Despite the stronger learning effect seen in the original condition, that version of Speech Graffiti is still at a disadvantage in terms of efficiency in that it requires a pre-use tutorial. The results from Study I indicate that although the shaping system could potentially produce more efficient interactions, a simple adaptation-theoretic approach may not be the most effective way to influence users to converge to the target, more efficient, Speech Graffiti format. Thus, more-explicit shaping strategies were investigated in Study II.

5.6.1 Key findings from User Study I

- Trend towards increased efficiency and satisfaction for shaping group.
- Successful interactions in shaping group without tutorial.
- Successful deployment of the two-pass ASR strategy.
- Overall intrasession convergence.

Chapter 6

Changes to the Shaping System

Given the results from Study I, the second user study explored the effectiveness of two more-explicit shaping strategies. The first was an *explicit* strategy, in which I attempted to convey more strongly that the system prefers a certain form of input. Like the shaping strategy used in Study I, the explicit strategy did not *require* the user to speak in the Speech Graffiti format. The second method was a *requiring* strategy, which did require users to rephrase any expanded-grammar input before moving on.

The lack of significant performance differences between the shaping+tutorial and shaping+no-tutorial sub-groups in Study I suggested that a pre-use tutorial is not strictly necessary in the shaping condition. Thus, in Studies II and III, participants did not receive a pre-use tutorial.

6.1 Targeted help

I had initially proposed to provide more explicit feedback via shaping help in "full-confsig" cases (*i.e.*, item C in fig. 3.1). Full-confsig cases occur when no complete phrases are passed from the Phoenix parser to the Speech Graffiti engine for processing, thus generating a {confsig} error beep. However, by implementing the two-pass ASR method I significantly reduced the number of occurrences of this situation in Study I, as shown in fig. 6.1 ($t = 3.24$, $p = 0.01$).

Because of the drastic decrease in the occurrence of full-confsigs, I decided to focus instead on providing help instead situations where problems occurred more frequently in the Study I data. I believe that active help should be useful in many of these situations since I found that users were not all that likely to ask for help on their own. Only nine of the 20 participants in the shaping condition used the **help** command, for a total of 34 **help** instances (comprising only 1% of all utterances in the shaping condition). Furthermore, 15 of the 34 **help** instances were immediately preceded by a system prompt noting the availability of this command, either as part of the system response to an **options** command (...or say 'help' for more information) or as a note about getting additional information after hearing the first

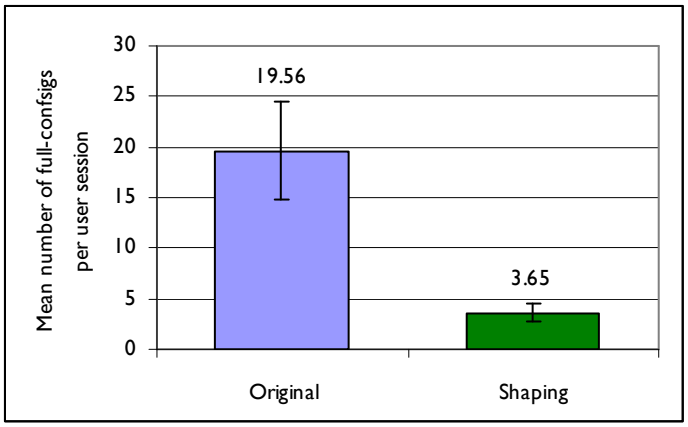

Figure 6.1. Mean number of full-confsigs per user session in Study I.

level of the help prompt (**...to hear more about this, say 'help' again**). I thus conclude that in general, users are not very inclined to ask for help spontaneously, yet satisfaction and performance rates with the system suggest that targeted, active help could improve user interactions with the system.

To assess the potential targets for such active help, I looked more closely at the corpus of interactions from users in the shaping condition. First, I analyzed utterance types and their effect on concept error. In Speech Graffiti, each user utterance can be categorized as one of four types:

- specification-phrase, *e.g.,* **`genre is drama`** (30% of Study I utterances);
- query-phrase, *e.g.,* **`what are theaters`** (21%);
- keyword, *e.g.,* **`start over`** (42%);
- mixed, *e.g.,* **`genre is drama, what are theaters`** (7%).

The contribution of each type of utterance to the overall occurrence of concept errors does not vary greatly, as can be seen by comparing the relative sizes of the bottom portion of each column in fig. 6.2. However, by considering each type of utterance and what percentage of all utterances of that type resulted in concept errors, it appears that mixed utterances are much more likely to generate concept errors than were utterances of the other types.

I also analyzed the shaping condition session logs and made subjective assessments of what seemed to be each user's main impediments to successful, efficient use of the system. These results are summarized in table 6.1, which is ordered by the number of users who experienced each issue. There was no specific frequency threshold for each problem to be included in this list, so some users may have experienced certain issues at higher frequencies than other users. Most users also experienced multiple issues.

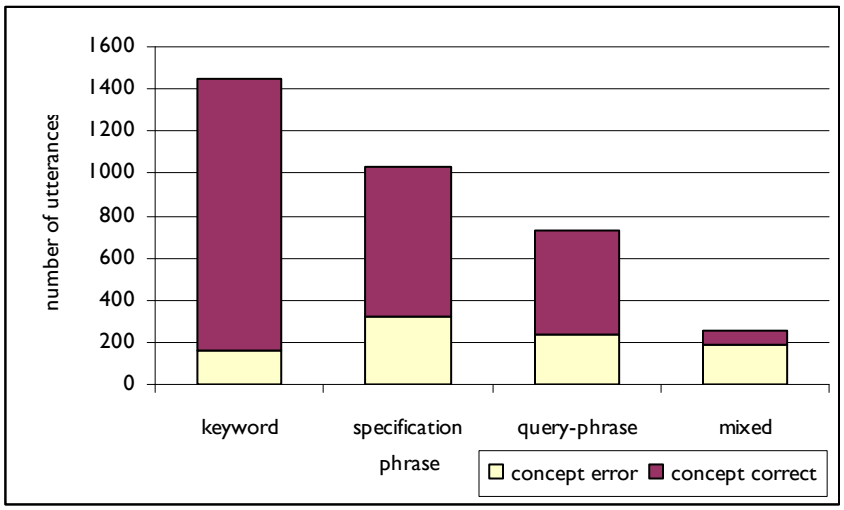

Figure 6.2. Distribution of utterance types in the shaping condition, with each utterance type showing ratio of concept-error to concept-correct utterances.

Table 6.1. Interaction problems experienced by users in the shaping condition of Study I.

Problem	Number of users
1. Using natural language query format	15
2. Not using **`start over`**	5
3. Using long utterances	3
4. Using slot-only query format	3
5. Confusion about semantics of **`location`**	3
6. Using **`next`** instead of **`more`** for list navigation	2
7. Using value-only specification format	2

Based on the results from these analyses, I decided to address issues 2, 3, and 6 from table 6.1 with targeted help prompts. I addressed issues 1 and 4, clearly the most pervasive problems, with a different strategy, discussed below in Section 6.2.

6.1.1 `Start over`

The Speech Graffiti system currently retains query context from turn to turn. Thus, if a user says, **`genre is comedy, theater is Manor, what are movies`**, gets the system result, and then says, **`theater is Waterworks, what are movies`**, the second database query will also contain the constraint "genre = comedy." To clear the query context, the user must issue the **`start over`** command. In the context of the user study, participants often moved from task to task without using **`start over`**, which meant that extraneous, error-causing constraints (in terms of being able to find the "correct" task information) sometimes ended up in their queries. Extraneous constraints sometimes also appeared as a result of recognition errors. In both cases, users often did not seem to be aware of the problem.

Thus, I implemented a help strategy to be triggered in two cases:

- for all queries containing constraints that had been retained from previous tasks, and

- for any queries that returned no results.

Fig. 6.3 shows a sample interaction with examples of the new prompts (see Section 6.2 for a discussion of the **`list`** query format). In the third system prompt in fig. 6.3, since there are no results from the query, the system summarizes the constraints that were used, and then explicitly reminds the user that he can say start over if these were not the intended constraints. In the final system prompt in fig. 6.3, the system summarizes the query before providing the results, since the "genre = action" constraint has been retained from the previous query. In addition to providing targeted help in these cases, I also slightly modified the system's context retention

strategy so that only specification-phrases, and not query-phrases, are retained in the system's context after a user query.

6.1.2 Long utterances

As noted above, mixed utterances were quite likely to result in concept errors. Since by definition this type of utterance contains more than one phrase, these utterances generally tended to be longer than the other types of utterances. Furthermore, I found that 61.4% of ASR hypotheses containing five (linked) words or more resulted in concept errors, while only 25.2% of hypotheses of fewer than five words did so. To provide active help targeted at long utterances, I chose a conservative trigger of seven words or more (69.5% of which generated concept errors in the Study I corpus, compared to 26.6% for hypotheses of less than seven words). The help strategy delivers the following prompt when such a hypothesis is chosen: I'm not sure I got that. It might help to split up your request into shorter phrases, and wait for a confirmation of each part.

Theater is Carmike Ten

Carmike Ten Pittsburgh

Genre is action

Action

List movies

Sorry, no titles to list for Carmike Ten Pittsburgh, action. If these don't sound like the items you wanted to ask about, say "start over" and try again.

Theater is Loews Waterfront

AMC Loews Waterfront 22

List movies

Listing 2 titles for AMC Loews Waterfront 22, action: Mission Impossible III, The Fast and the Furious 3: Tokyo Drift.

Figure 6.3. Sample interaction showing system strategies for alerting users to potential extraneous context information (in third and fifth system turns).

6.1.3 `Next` instead of `more`

The **`next`** keyword returns the next single item from a list, while **`more`** returns the next three items. In two cases from Study I, participants consistently used **`next`** for their list navigation. It may have been the case that they forgot about the **`more`** keyword (or did not realize the difference between it and **`next`**), or they may also simply have preferred the one-item-at-a-time format. If the situation is the former, a suggestion should be offered to the user, but, in the event that the situation is the latter, this suggestion should not be too annoying. Thus I implemented a targeted help message to be delivered on every third consecutive **`next`** prompt. After providing the appropriate query result item, the system gives the following message: Remember that you can also hear three matches at a time instead of just one by saying "more."

6.2 Query format

The most pervasive user difficulty issue from Study I was the use of natural language-style queries. I suspect that this was because the **`what is [slot]`** format is actually not structured enough. That is, although it has a clear structure in terms of its grammar, when the tutorial or help prompts tell users about the format, they may simply interpret this information as "ask the system a question" rather than "say 'what is [slot].'" Thus I decided to introduce more structure into the query-phrase format. This seemed acceptable since users did not seem to have too much difficulty with the idea of structure in the specification-phrase format. I chose to use the format **`list [slot]`**, since this form could be nicely converted into a related template for delivering results (*i.e.*, listing seven titles...). Having a more clearly structured query format also could help users who had the opposite problem: using just slot names for queries (*e.g.*, **`show times`**). The interaction in fig. 6.3 includes examples of the **`list [slot]`** format.

6.3 Grammars and language models

In preparation for the second user study, I also enlarged the expanded grammar to make it a complete superset of the Speech Graffiti grammar (the expanded grammar in Study I did not include all of the Speech Graffiti keywords). This change was made in accord with the philosophy that the expanded grammar should cover "what people say to the system," which includes Speech Graffiti utterances as well as more natural language input. To better align the two grammars, I ensured that all of the linked words from the Speech Graffiti grammar were also linked in the expanded grammar, and that as many additional linkable items as possible were joined as well. I also added a handful of out-of-vocabulary terms that had been used in Study I, such as **`latest`**, **`science fiction`** (as a synonym for the existing genre **`sci-fi`**), and **`when's`**. Finally, per issue 5 from table 6.1, I also removed the term **`location`** (as a synonym for **`area`**) from the grammar, as this term seemed ambiguous and confusing for users. The updated lexicon sizes are shown in table 6.2.

In the interest of trying to reduce the ASR word-error rate—and thus approaching, as far as possible, the ASR performance of a state-of-the-art spoken dialog system—I also made some adjustments to the language models used in the system. I analyzed the Study I corpus and added appropriate probabilities to the language model corpora-generation grammars for keywords and for the number of phrases in an utterance (*i.e.*, single phrase utterances vs. multiple-phrase utterances). I changed the following items to be considered as classes by the language model: movie titles,

Table 6.2. Lexicon sizes for User Study II. In this table, lexicons from both grammars include variations for 40 movie titles.

	Size of MovieLine lexicon	
	With linked words	With all words unlinked
Speech Graffiti	513	363
Expanded	620	409

theater names, minutes, hours, weekdays, movie ratings, genres, areas, months, and ordinal numbers, with equivalent probabilities within each class. Finally, I changed the language model discounting strategy from Good-Turing to absolute. On the full Study I utterance corpus, these changes resulted in a relative WER improvement of 9.8% for Speech Graffiti hypotheses and 4.3% for expanded hypotheses.

Chapter 7

User Study II Design: More-Explicit Shaping

The second user study was designed to compare the effects and effectiveness of three different shaping strategies: implicit, explicit, and requiring. The goal was to determine whether any of these three strategies had significant effects on interaction efficiency and Speech Graffiti convergence.

7.1 Participants

Thirty adults participated in the study. They were recruited from the neighborhood around Carnegie Mellon University and the University of Pittsburgh via small public signs and by postings on Carnegie Mellon's electronic bulletin boards. Twenty-nine were native speakers of American English; the final participant learned French

Canadian as his first language but moved to the United States at an early age and did not have a discernable foreign accent. All participants were between the ages of 21 and 54; 17 were female and 13 were male.

As in Study I, I specifically recruited participants for this study who did not have significant experience with computer programming, and all were new to the Speech Graffiti interface. Again, about two-thirds of the participants reported using telephone-based information services five times a month or less. All of the participants reported having completed at least some undergraduate coursework, and about a third of the participants had received graduate degrees or completed some graduate coursework. Table 7.1 summarizes the participant demographics for this study.

7.2 Conditions

A between-subjects experiment was designed in which participants were randomly assigned to one of three conditions: implicit shaping, explicit shaping, or required shaping. The conditions differed in their wording of prompts in the case where expanded-grammar input was recognized. Unlike in Study I, in all three conditions in Study II users who spoke Speech Graffiti-grammatically received a terse, value-only confirmation of their input. This strategy was implemented so that users who spoke Speech Graffiti would receive some immediate reward, in the form of a shorter system prompt. The full slot+value versions *were* provided when users asked **`where was I?`**. Fig. 7.1 provides examples of the system prompts in all three conditions, and they are each discussed in more detail below.

Table 7.1. Selected demographic characteristics of participants (N = 30) in User Study II.

Characteristic	*N*	%
Age at time of survey (years)		
20-24	5	17.2
25-34	17	58.6
35-44	1	3.4
45-54	6	20.7
Highest education level completed		
Some high school or less	0	0
High school graduate	0	0
Some college	2	6.7
2-year college/technical school	2	6.7
4-year college	15	50.0
Some postgraduate work	5	16.7
Postgraduate degree	6	20.0
Reported frequency of computer programming		
Never	23	76.7
Rarely	6	20.0
Fairly often	1*	3.3
Very frequently	0	0

* participant wrote a note next to this selection: "but here I use 'programming' loosely"

7.2.1 Implicit shaping

The implicit shaping condition replicated the shaping condition from User Study I. When users spoke according to the expanded grammar, they received confirmations in the Speech Graffiti slot+value format. Any input that the system understood from the expanded grammar input was added to the query context as usual.

*sample **Implicit** shaping interaction*

Theater Manor, genre is comedy
Cinemagic Manor Theatre, comedy
What are movies?
Listing 2 titles: For Your Consideration, The Queen
Galleria
Theater is Carmike Galleria 6
Theater is Galleria
Carmike Galleria 6
Genre is drama, what's playing?
Genre is drama, listing 3 titles for drama, Carmike Galleria 6: A Good Year, Babel, The Departed.
Where was I?
Genre is drama, theater is Carmike Galleria 6
List

{confsig}

*sample **Explicit** shaping interaction*

Theater Manor, genre is comedy
Cinemagic Manor Theatre, comedy
What are movies?
I think I heard "what are movies." Next time, it would help to use Speech Graffiti, as in "list title." Listing 2 titles: For Your Consideration, The Queen
Galleria
I think I heard, "Galleria." Next time, it would help to use Speech Graffiti, as in "theater is Carmike Galleria 6."
Theater is Galleria
Carmike Galleria 6
Genre is drama, what's playing?
I think I heard "genre is drama, what's playing." Next time, it would help to use Speech Graffiti, as in "genre is drama, list title." Listing 3 titles for drama, Carmike Galleria 6: A Good Year, Babel, The Departed.
Where was I?
Genre is drama, theater is Carmike Galleria 6
List

{confsig}

*sample **Required** shaping interaction*

Theater Manor, genre is comedy
Cinemagic Manor Theatre, comedy
What are movies?
Please rephrase that using Speech Graffiti. For example, "list title."
Galleria
Please rephrase that using Speech Graffiti. For example, "theater is Carmike galleria 6."
Theater is Galleria
Carmike Galleria 6
Genre is drama, what's playing?
Please rephrase that using Speech Graffiti. For example, "genre is drama, list title."
Where was I?
Genre is comedy, theater is Carmike Galleria 6.
List
{confsig}

Figure 7.1. Sample interaction showing system response differences between the three shaping conditions of User Study II.

7.2.2 Explicit shaping

The following requirements were considered in the design of the explicit shaping prompt:

- it should confirm the user's input;
- it should provide the user with an equivalent Speech Graffiti example;
- it should suggest to the user that he should use the Speech Graffiti format next time;
- it should not imply that the user needs to rephrase his input immediately;
- it should not suggest a **yes**/**no** answer (the current Speech Graffiti implementation does not support **yes**/**no** input, although it could of course be modified to do so); and
- it should leave open the possibility that a recognition error might have occurred.

Given these requirements, the following explicit shaping prompt template was devised:

> I think I heard "[ASR hypothesis]." Next time, it would help to use Speech Graffiti, as in "[equivalent Speech Graffiti input based on ASR hypothesis]."

As in the implicit condition, any input that the system understood from the expanded grammar was added to the query context.

7.2.3 Required shaping

The required shaping prompt had the following design requirements:

- it should confirm the user's input;

- it should provide the user with an equivalent Speech Graffiti example;
- it should request that the user rephrase her input before proceeding;
- it should not suggest a **yes**/**no** answer; and
- it should leave open the possibility that a recognition error might have occurred.

The basic template used for the required prompt was

> Please rephrase that using Speech Graffiti. For example, "[equivalent Speech Graffiti input based on ASR hypothesis]."

In contrast to the other two strategies, expanded grammar input was not added to the context in the required condition. Thus, in the required example in fig. 7.1, when the user asks **where was I?**, the system reports that it has retained the earlier, grammatical **genre is comedy** input rather than replacing it with **genre is drama**, since the latter input was not part of a fully Speech Graffiti-grammatical utterance and the user did not rephrase that input.

7.3 Setup and tasks

Based on the results from Study I, participants in this study were not given a pre-use tutorial. Instead, at the end of the pre-study briefing with the experimenter, users were told the following: "when you first call the system, you'll hear information about Speech Graffiti, which is a special format for talking to a speech recognition system like this. This information is important, so be sure to listen to it." Upon calling the system, participants heard a 68-second introductory recording that was similar to the one presented to the Study I no-tutorial sub-group, except that it explained the **list** format and also included the **where was I** keyword. Barge-in was turned off while the introduction played, to ensure that all users heard the same information, but the prompt could not be replayed.

Participants completed the study in a conference room either at Carnegie Mellon University or in the Biomedical Science Tower (BST) at the University of Pittsburgh. In both environments, participants were seated at a table and interacted with the system over a standard, land-line, office telephone. The audio was recorded over the telephone for all sessions. During the sessions in the BST, the experimenter monitored the session while in the same room as, but out of sight of, the user; at Carnegie Mellon the experimenter monitored the interactions from outside the room.

As in Study I, the domain used in this study was movie information. Participants were asked to complete a series of 15 tasks using the MovieLine system, following the same task ordering framework described in Section 4.3. The only differences in the tasks between Study I and Study II was that the movie names were updated to reflect those in theaters at the time. Boxes were also added next to each task on the user worksheet in which participants were instructed to mark whether they believed they had correctly completed the task (with answer choices *yes - not sure - no*).

Users were given the complete list of tasks on a sheet of paper and asked to work through them in order, writing down the answers for each. Participants were instructed that if they had time at the end of their session, they could go back and work on any tasks they had abandoned or for which they were not sure they had found the correct information.

Participants were given forty minutes to work through the set of tasks. Twenty-two users declared that they were finished with the tasks before their forty minutes were up (seven each in the implicit and required conditions and eight in the explicit condition); the remaining eight participants were asked to stop working on the tasks when the forty minutes expired. Of these eight participants, three were not able to at least attempt all 15 tasks within the forty-minute session (two participants worked on

14 tasks, and the third worked on 11). However, their data was included in the analysis for this study as they had made earnest attempts at as many tasks as possible during the session. To motivate users to complete tasks successfully, participants were compensated for their time with a flat cash payment for participation ($12.50) plus an additional amount (50 cents) per correctly completed task.

7.4 User survey

Upon declaring that they were finished, or at the end of the forty minute session, users were asked to complete an evaluation questionnaire. This questionnaire was identical to the one used in Study I (table 4.3), with the addition of the following four items to evaluate on the 7-point Likert scale:

- *The system gave me useful feedback.*
- *I understand how to use Speech Graffiti.*
- *I understand the difference between Speech Graffiti and "normal" language.*
- *I noticed shorter responses from the system when I used Speech Graffiti.*

After filling out the questionnaire, participants were asked to suggest a tip that they might give to a friend who was about to use the system. They were then debriefed on the purpose of the experiment and were given the opportunity to ask questions about their experience. Finally, they were compensated for their time with the appropriate cash payment.

7.5 Analysis

The basic analyses conducted for Study II were similar to those conducted for Study I: overall task completion, mean time- and turns-to-completion rates, and mean scores for each of the six user satisfaction factors plus a combined, overall user

satisfaction rating. I also assessed grammaticality, the effectiveness of the targeted help prompts, and the effect of each shaping prompt type on local convergence.

Chapter 8

User Study II Results: More-Explicit Shaping

Overall, there were no major efficiency differences among the three shaping conditions of Study II. Significant intrasession grammaticality increases were observed in all conditions, with the required condition generating the strongest local convergence. Interaction effects between condition and initial grammaticality suggested the potential for an adaptive approach to shaping

8.1 Efficiency

8.1.1 Task completion

Across all conditions, participants completed an average of 10.67 tasks (S.D. = 4.02). There were no significant differences in task completion among the three shaping conditions, as shown in fig. 8.1 ($F = 0.54$, $p = 0.59$).

8.1.2 Time

For completed tasks, there were no significant differences in mean time-to-completion among the conditions (fig. 8.2, left) ($F = 0.20$, $p = 0.82$). As in Study I, I also conducted an analysis of medians to take into account all tasks, not just completed ones. All incomplete tasks in Study II were assigned a time of 587.12 seconds (105% of the maximum time observed). Users in the required condition generally spent the most time working on tasks (fig. 8.2, right), but the overall differences were not significant due to the large variances in the data ($F = 0.30$, $p = 0.74$).

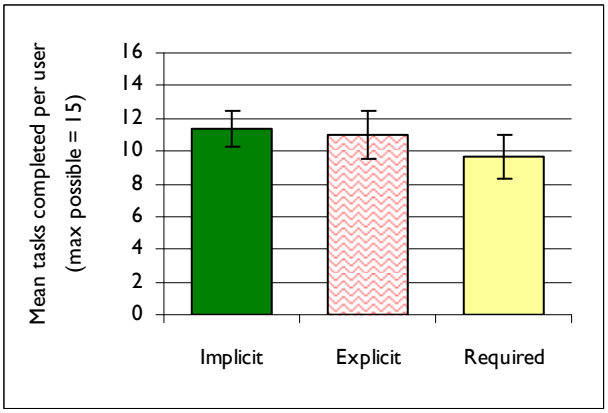

Figure 8.1. Mean number of tasks completed per subject in Study II.

8.1.3 Turns

For completed tasks, there were no significant differences in mean turns-to-completion among the conditions ($F = 0.99$, $p = 0.38$). In the assessment of median turns-on-task, users in the required condition generally spent the most turns on tasks, but again this difference was not significant due to large variances ($F = 0.37$, $p = 0.70$) (fig. 8.3).

8.2 User satisfaction

For most of the seven user satisfaction factors and overall, the required condition generated the lowest mean scores, although there were no significant differences

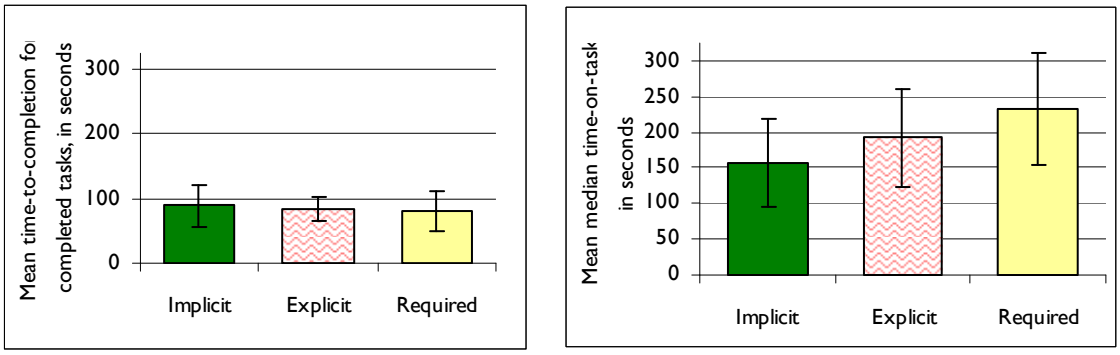

Figure 8.2. Mean per-user time for completed tasks (left) and mean per-user median time spent on all tasks (right) in Study II.

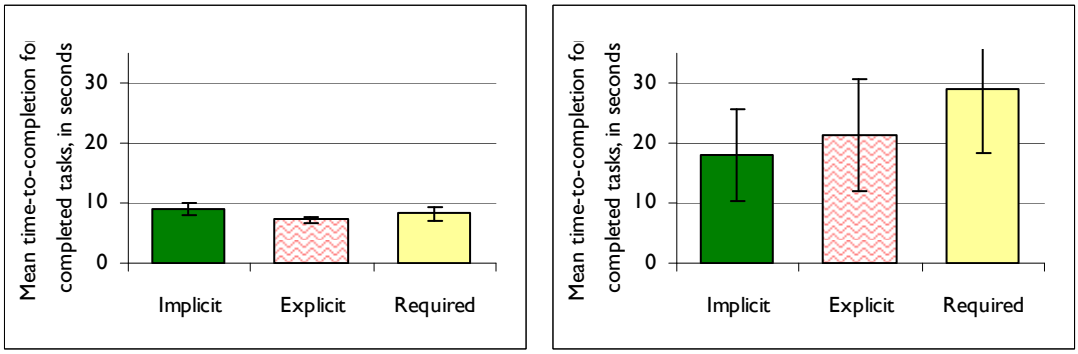

Figure 8.3. Mean per-user turns for completed tasks (left) and mean per-user median turns spent on all tasks (right) in Study II.

among the three conditions (fig. 8.4). The lowest habitability factor score was generated by the implicit condition, and in fact users in this condition gave significantly lower scores to statements in this factor than to those in all other factors. The habitability factor involves knowing what to say to the system (*e.g.*, *I sometimes wondered if I was using the right word*; *I always knew what to say to the system*"), and it appears that this was problematic in the implicit condition, which did not provide explicit examples of how to speak to the system.

8.2.1 Correlations

I assessed correlations between Speech Graffiti grammaticality and user satisfaction and found a relatively strong correlation for the required condition (0.61, $p = 0.06$) but no correlation in the other two conditions (explicit 0.05, $p = 0.89$; implicit 0.01, $p = 0.97$).

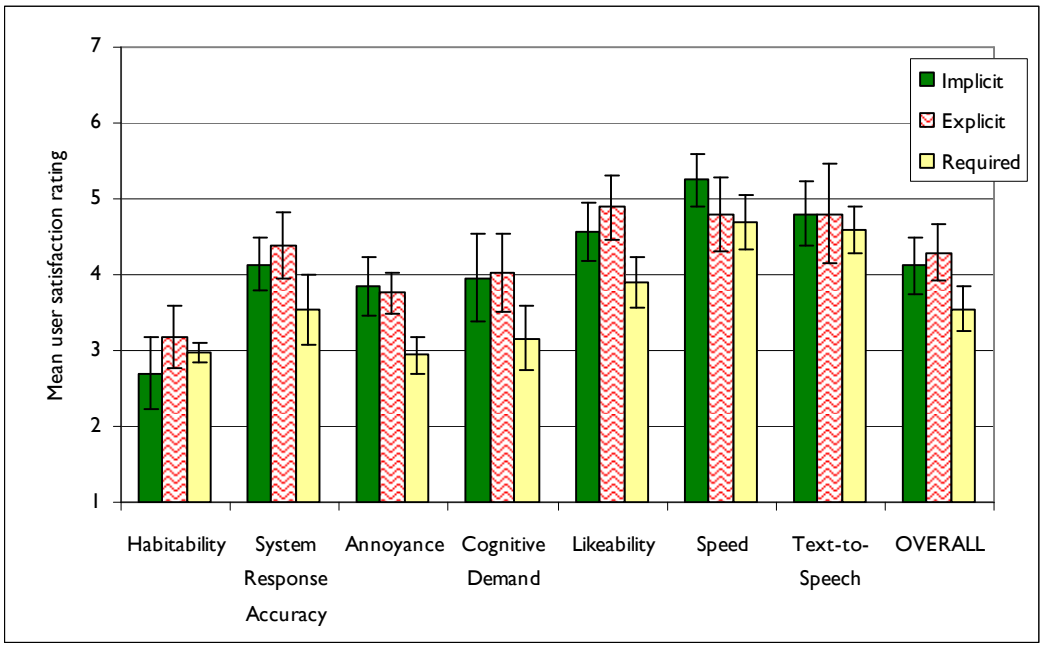

Figure 8.4. Mean user satisfaction ratings from Study II for each of the seven user satisfaction factors, and combined as an overall rating.

8.3 Grammaticality

Speech Graffiti grammaticality was generally quite high in Study II, with a median of 84.7% across all participants (*cf.* the median of Study I, 65.1%). 18 participants were at least 80% grammatical in their sessions (five from the implicit condition, six from the required condition, and seven from the explicit condition). Objective and subjective measures confirmed result from the ATUE study that users who achieve 80% Speech Graffiti grammaticality have much more successful and efficient interactions compared to those who do not achieve that level of grammaticality, as summarized in table 8.1. Across the three conditions, there were no significant differences in overall grammaticality, as shown in fig. 8.5.

Table 8.1. Comparison of objective and subjective measures between highly grammatical and low grammatical participants across all conditions in Study II.

	≥ 80% Speech Graffiti grammatical		< 80% Speech Graffiti grammatical			
Measure	M	SD	M	SD	*t*	*p*
Tasks completed	12.33	3.03	8.17	4.13	-3.00	0.008
Time-to-completion	73.93	21.57	99.95	27.26	2.78	0.01
Turns-to-completion	7.61	2.54	8.84	2.97	1.18	0.25
Overall user satisfaction	4.34	0.98	3.46	1.13	-2.20	0.04

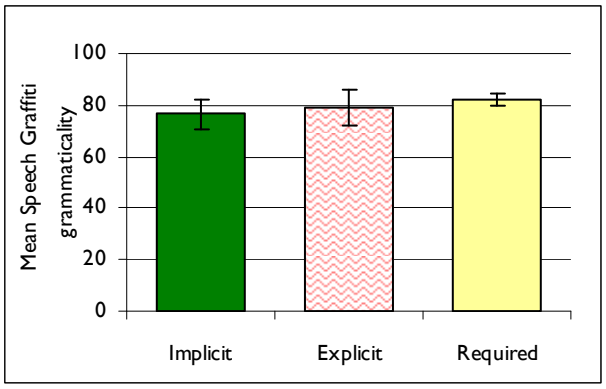

Figure 8.5. Mean grammaticality across conditions in Study II.

8.3.1 Intra-session grammaticality

As in Study I, I assessed intrasession grammaticality change by comparing the first quarter of each user's session with the final quarter of their session. Within each condition, I found significant increases in grammaticality over time, as depicted in fig. 8.6. However, there was no significant difference in the strength of change between the three conditions. ($F = 0.43$, $p = 0.65$).

8.4 Secondary effects

Despite the lack of significant main effects of condition on user satisfaction and objective measures, there were interesting differences when the study participants were divided into two groups: those with high initial grammaticality (operationalized as having target grammaticality of at least 80% in the first quarter of their session) ($N = 12$) and those with low initial grammaticality ($N = 18$). The participants in these

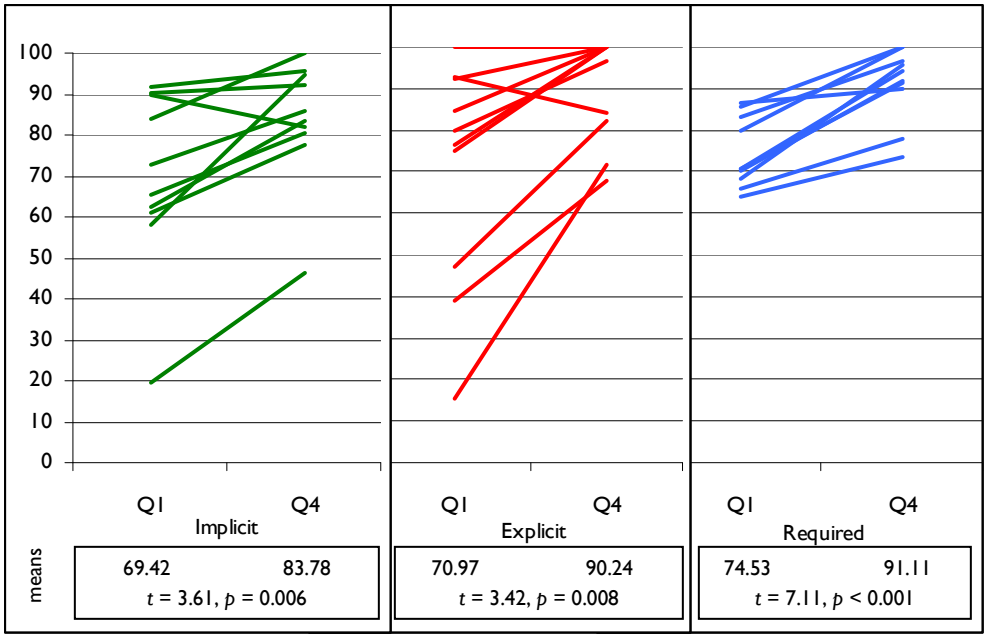

Figure 8.6. Speech Graffiti grammaticality change from first quarter of session to final quarter, by condition. Each line represents one participant.

two groups were coincidentally distributed evenly across all three shaping conditions.

Not surprisingly, initial grammaticality had a significant effect on several measures: overall satisfaction ($t = 3.55$, $p = 0.002$), tasks completed ($t = 4.15$, $p < 0.001$), times-to-completion ($t = -4.75$, $p < 0.001$), and turns-to-completion ($t = -2.92$, $p = 0.007$), with high initial grammaticality resulting in better scores on these measures.

There were significant interaction effects between initial grammaticality and condition on the habitability satisfaction factor (which concerns knowing what to say to the system) ($F = 7.26$, $p = 0.003$), as shown in fig. 8.7. For high initial grammaticality users, habitability scores were high in the implicit and required conditions, and decreased in the required condition. For low initial grammaticality, habitability scores were lowest in the implicit condition, and increased through the explicit and required conditions. This suggests that the extra feedback from the more-explicit prompts helps users who may initially have trouble with the system know what to say. Compared to the other two conditions, implicit shaping provides minimal support in this area.

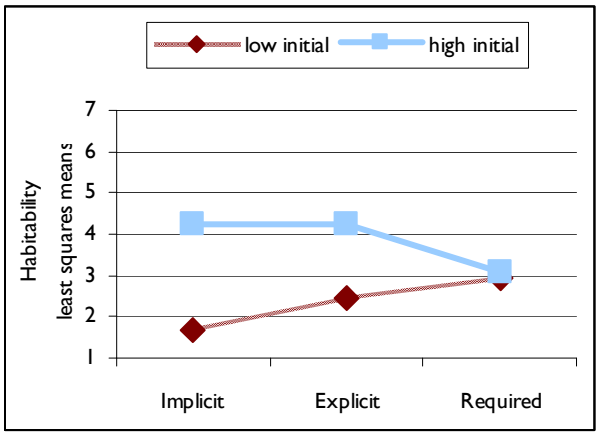

Figure 8.7. Interaction effect of initial grammaticality and condition on habitability ratings in Study II.

On the other hand, high initial grammaticality users could be confused by the prompts in the required condition. Each shaping prompt issued by the system can be categorized with one of the following accuracy tags:

- Correct trigger: the user said something non-Speech Graffiti-grammatical (**Manor**), and the system had the correct ASR hypothesis (Manor) and delivered the correct shaping prompt (theater is Cinemagic Manor Theatre).
- Concept error: the user said something that may or may not have been Speech Graffiti-grammatical (**Manor**; or **theater is Manor**), and the system interpreted it as a different, ungrammatical concept (Oaks; or comedy) and delivered a shaping prompt for that concept (theater is Oaks Cinema; or genre is comedy).
- Mis-trigger: the user said something Speech Graffiti grammatical (**theater is Cinemagic Manor Theatre**), but the system only heard part of the input (Manor), thus generating a shaping prompt telling the user that the proper format is essentially what they just said (theater is Cinemagic Manor Theatre).

Since high initial grammaticality users by definition speak more Speech Graffiti grammatically, they should experience fewer shaping prompts, and the data from Study II supports this (fig. 8.8, left). However, when shaping prompts are triggered for high initial grammaticality users, the distribution of the three accuracy categories is significantly different from the distribution for low initial grammaticality users (likelihood ratio $\chi^2 = 8.34$, $p = 0.02$), with mis-triggers more likely for the former group (fig. 8.8, right). When mis-triggers occur in the implicit or explicit conditions, the user can simply ignore them and move on. In the required condition however, it can be confusing and frustrating when the system asks users to rephrase their input using the same form that they just uttered to the system.

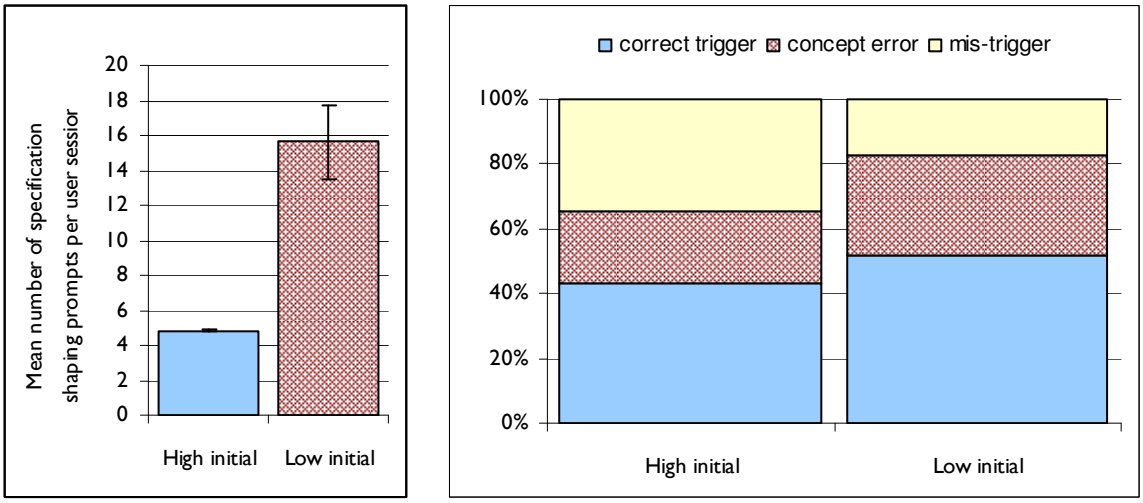

Figure 8.8. Mean number of specification-phrase shaping prompts triggered per user session (left) and distribution of prompt accuracy categories for specification-phrase shaping prompts by initial grammaticality.

There was also an interaction effect between initial grammaticality and condition on turns-to-completion ($F = 3.46$, $p = 0.05$). For turns-to-completion, low initial grammaticality users completed tasks in the fewest turns in the explicit condition, while high initial grammaticality users completed tasks in the *most* turns in that condition (fig. 8.9). For low initial grammaticality users, I suspect that the explicit condition struck a good balance between helping users know what to say and

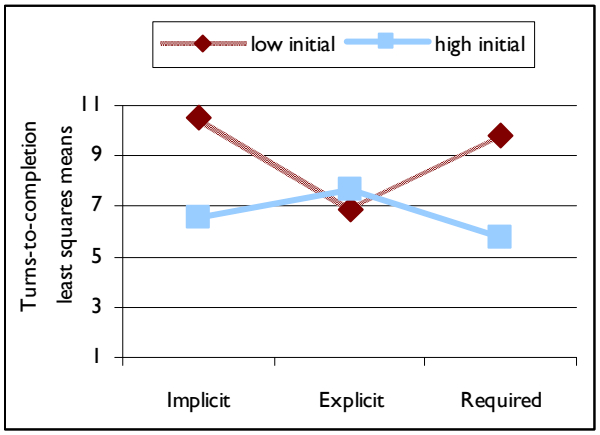

Figure 8.9. Combined effect of initial grammaticality and condition on turns-to-completion in Study II.

increasing task efficiency by not requiring users to rephrase their input immediately.

I also found significant effects of gender on overall user satisfaction (t = -3.07, p = 0.005) and on time-to-completion for completed tasks (t = 2.20, p = 0.04), with male users having better scores on these measures. Female users also had significantly higher word-error rates (as calculated on their target-grammatical input, to remove any effect of out-of-vocabulary items; t = 2.37, p = 0.02), and there were also significant differences in Speech Graffiti grammaticality based on age groups (F = 8.32, p < 0.001), with users in the 45-54 age range having significantly lower grammaticality than the other groups. However, there were no significant interaction effects between gender or age group and the shaping conditions, nor were there interaction effects between gender or age group and word-error rate on user satisfaction or task completion measures.

8.5 Effectiveness of shaping prompts

In Study II, I was concerned with the efficacy of each shaping strategy. As one method of assessing this, I analyzed how influential each strategy was in producing locally converged, Speech Graffiti-grammatical user input. In this case, effectiveness was operationalized as users' Speech Graffiti grammaticality following a shaping prompt. To measure convergence, I considered each shaping prompt instance that met the following two conditions:

1. The prompt included a specification phrase in its shaping content (*e.g.*, Please rephrase that using Speech Graffiti. For example, "title is The Departed."). For the explicit and required conditions, I also considered shaping prompts that included query phrases. Query phrases were not included in the analysis for the implicit condition since their form was embedded in the query result in that condition. For instance, if a user said **what's playing at the**

Manor, the implicit shaping prompt response was theater is Cinemagic Manor Theater, listing 3 movies….

2. The shaping phrase presented in the prompt was appropriate for the user input (*i.e.,* there were no concept errors due to misrecognitions);

To assess the effectiveness of the prompts, I considered whether the next user phrase of that type (specification or query) was Speech Graffiti-grammatical. The specific slot and value in the following utterance did not need to match that given in the prompt, and there could be intermediate user utterances between the shaping prompt and the analyzed user phrase.

For specification phrases, a contingency analysis showed a significant difference between the three conditions (likelihood ratio $\chi^2 = 50.3$, $p < 0.001$), with users much more likely to converge locally to the Speech Graffiti format in the required condition. Local convergence for query phrases was also significantly greater in the required condition compared to the explicit condition (likelihood ratio $\chi^2 = 28.7$, $p < 0.001$) (fig. 8.10).

The strength of local convergence in the required condition is not surprising, since in that condition the system explicitly asks users to rephrase their input. What is interesting however, is that the explicit condition, which suggests that users use the Speech Graffiti format next time, is comparatively weak, with less than 50% local convergence for both types of phrases.

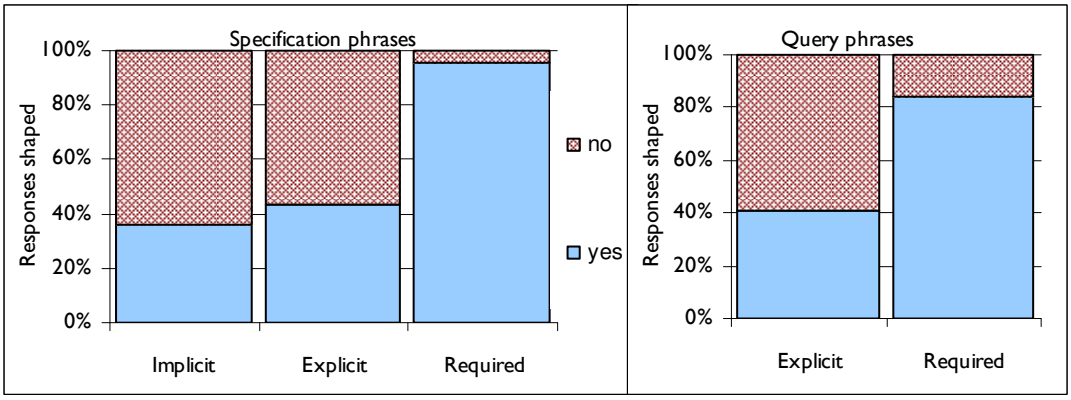

Figure 8.10. Distribution of user utterances displaying convergence after a shaping prompt in Study II. For each shaping prompt, the user utterance considered as convergent or not is the next user utterance of the same type (query or specification) as the shaping prompt. Query prompt effectiveness is excluded for the implicit condition since the shaping query prompts were embedded in the result presentation for that group, unlike in the other conditions.

8.6 Effectiveness of targeted help

As discussed in Section 6.1, three targeted help prompts were implemented in the Study II systems to address situations that caused difficulties in Study I: the use of **`next`** instead of **`more`**, long utterances, and context confusion caused by not using **`start over`**.

The **`next`**-vs.-**`more`** issue never appeared in Study II. I revisited the Study I data to see whether the participants who experienced this issue in Study I had had the tutorial, which did mention **`next`** as a navigation option, while the short introduction that the no-tutorial sub-groups received did not. One participant had been in the tutorial sub-group, and the other had not. The use of **`next`** instead of **`more`** was clearly not a pervasive problem for users, but there was little overhead involved in adding this help prompt to the system.

There were five instances in Study II where the long utterance help prompt (I'm not sure I got that. It might help to split up your request into shorter phrases, and wait for a confirmation of each part.) was triggered, distributed across four users. In each case,

had the system not generated this help prompt, the parse of the long input's ASR hypothesis would have generated a concept error. Users issued single phrase utterances in reply to the help prompt in all five instances, with correct concept recognition in four out of the five cases.

8.7 System performance

User Study II generated a corpus of 4,653 utterances. 1,708 (37%) were from participants in the implicit condition, 1,351 (29%) were from participants in the explicit condition, and 1,594 (34%) were from participants in the required condition. As in Study I, the transcriptions of these utterances were cleaned of non-task items such as noise, system feed, and off-task user comments prior to further analysis. After the cleaning process, any utterances that had contained only non-task items were retained in the corpus as empty utterances.

8.7.1 Word-error rate

The mean word-error rate (for unlinked words) across all conditions in Study II was 32.3% (S.D. = 14.8). For Speech Graffiti grammatical input, WER was 20.7% (S.D. = 9.13). There were no real differences in WER between the conditions, as was expected since all three conditions used the same two-pass models. The WER for Speech Graffiti-grammatical input in the two-pass versions in Study II represented an 18% relative decrease from the mean of 25.3% for the shaping condition in Study I, suggesting that the language model adjustments made after Study I were helpful ($t = 1.62$, $p = 0.06$).

8.8 Discussion

There were no real differences in efficiency measures in Study II among the three conditions. Users in the implicit condition generally spent less time- and turns-on-task, but gave the system low habitability ratings. This condition seemed especially

problematic for users with low-initial grammaticality. Users in the required condition tended to have higher time- and turns-on-task, with lower user satisfaction scores. For low-initial grammaticality users, the required condition generated the highest habitability scores, although these scores dropped in this condition for high-initial grammaticality users. The explicit condition seemed to fall in between the implicit and required groups on most measures. For low-initial grammaticality users, the explicit condition was effective in generating fewer turns-to-completion compared to the other two groups.

As measured by significant intrasession grammaticality change, there was general evidence for specific syntactic convergence in Study II. The required condition was most effective at actually getting users to say something locally grammatical, as would be expected since users could not continue along the same path without rephrasing. However, due to the lack of differences between the shaping conditions as assessed by overall Speech Graffiti-grammaticality or intersession grammaticality, it is not clear that there are specific prompts that can enhance the *overall* convergence process.

The strong effect of initial grammaticality on user satisfaction and task success measures indicates that what users initially know about the system is quite important. Thus the introductory message content can have more of an effect on the interaction experience than any particular shaping prompts, so care should be given to the design of this aspect of spoken dialog systems.

The results from this study suggested that an adaptive model of shaping is potentially more useful for promoting convergence and effective interaction. Interaction effects showed that the required condition seemed to be useful for helping low-initial grammaticality participants know what to say, but that this condition had the opposite effect for high-initial grammaticality participants, who perhaps were confused or annoyed by this intrusion into their otherwise reasonably smooth

interaction. On the other hand, the explicit condition was also helpful for low-initial grammaticality users as it generated fewer turns per completed task compared to the other conditions. Thus for Study III I implemented an adaptive approach, which shifts its strategy based on the user's demonstrated grammaticality.

Finally, it is interesting that women generally had more negative experiences with the system, as measured by lower user satisfaction rates. While it often appears to be the case, as it was in this study, that women experience higher word-error rates with speech recognition systems, I did not find interaction effects between gender and word-error rate on user satisfaction. Thus, women rated the system lower than men regardless of how good the ASR performance was for them. This finding is somewhat surprising given that women generally tend to perform better on language-related tasks than men (Halpern, 2000). Age also had effects on Speech Graffiti grammaticality, with older users less grammatical. The fact that internal, user characteristics can have significant effects on user satisfaction and performance further supports the potential benefit of an adaptive approach.

Prior to Study III we made a few changes to the system. First, I replaced the three-beep list continuation signal ({…}) with the actual words and more. It seemed that users often had difficulty retrieving result chunks past the first set, and I wanted to make this a bit more clear. I also retrained the acoustic models used by Sphinx-II to include data from Study I and Study II. Utterances from two male and two female participants from Study II were held out as a test set, and the average per-speaker relative improvement in WER as a result of the retraining was 15.3% on the test set, with larger improvements for the female speakers.

8.8.1 Key findings from User Study II

- Overall intrasession convergence.

- The required shaping condition produced strong local convergence but was not robust to ASR errors and generated somewhat lower user satisfaction scores.

- Lower habitability score in the implicit shaping condition.

Chapter 9

User Study III Design: Adaptive Shaping

The goal of the third study was threefold: assess the effectiveness of an adaptive shaping strategy, investigate the effects of longer-term use of the system, and look at the success of cross-domain transfer of Speech Graffiti skills. Study III was therefore designed as a longitudinal study that took place in six sessions over the course of about seven weeks. During the first four sessions, participants interacted with the Speech Graffiti MovieLine. In the final two sessions, a new system was introduced: the Speech Graffiti DineLine. The basic hypotheses investigated were that adaptive shaping would promote more efficient interactions compared to explicit shaping, and that efficiency would increase in later sessions of the study.

9.1 Participants

Twenty-seven participants were originally enrolled in the study and participated in the first session. After the first session, two participants dropped out. Another dropped out after session three, and another after session five. When the later session recordings for another participant were transcribed, it was discovered that he spent several turns in each session literally screaming his input into the phone, which obviously affected his WER and success in the session. Thus, only data from his first two sessions, in which he was a compliant user, are included in this analysis. When longitudinal or cross-domain comparisons are made in Chapter 10, only data from participants still participating in the necessary sessions are included. For initial-session comparisons, data from all 27 participants is included. All participants were native speakers of American English and were between the ages of 23 and 54; 16 were female and 11 were male. At the end of the study, 14 females and 8 males had completed all six sessions.

As in previous studies, I specifically recruited participants for this study who did not have significant experience with computer programming, and all were new to the Speech Graffiti interface. About 60% of the participants reported using telephone-based information services five times a month or less. All of the participants reported having completed at least some undergraduate coursework, and almost half had received graduate degrees or completed some graduate coursework. Table 9.1 summarizes the participant demographics for Study III. Computer programming experience is not reported in the demographic table for this study as I had screened for it while recruiting subjects and did not include the question on the user survey.

9.2 Conditions

A between-subjects experiment was designed in which participants were randomly assigned to one of two conditions: explicit shaping or adaptive shaping. As in Study

Table 9.1. Selected demographic characteristics of participants in User Study III. Figures on the left represent all participants enrolled in the study (N = 27); figures on the right represent those who completed all six sessions (N = 22).

Characteristic	Initially enrolled		Completed study	
	N	%	*N*	%
Age at time of survey (years)				
20-24	10	37.0	8	36.4
25-34	11	40.7	8	36.4
35-44	2	7.4	2	9.0
45-54	4	14.8	4	18.2
Highest education level completed				
Some high school or less	0	0	0	0
High school graduate	0	0	0	0
Some college	4	14.8	2	9.0
2-year college/technical school	1	3.7	0	0
4-year college	9	33.3	8	36.4
Some postgraduate work	8	29.6	8	36.4
Postgraduate degree	5	18.5	4	18.2

II, users who spoke Speech Graffiti-grammatically received a terse, value-only confirmation of their input in both conditions.

9.2.1 Adaptive shaping

The adaptive shaping design was based on the results from Study II. Since the required condition had helped users with low initial grammaticality, adaptive shaping was designed to start in a state identical to the Study II required condition. Upon recognizing sustained Speech Graffiti grammaticality, the system switched to a state identical to the implicit condition from Study II. This setup was designed to alleviate the issue of shaping prompt mis-triggers for high grammaticality users, in which grammatical input is misinterpreted in the required condition and replied to with a prompt telling users to rephrase their input exactly as they have just said it. When mis-triggers occur in the implicit condition however, users simply hear the full

slot+value confirmation of their input, which is not that much more intrusive than the simple value-only confirmation that they would have heard had no error occurred. Thus, the adaptive condition should be helpful to low initial grammaticality users while being robust to recognition errors for high initial grammaticality users.

Two separate thresholds were set for switching shaping state in the adaptive shaping condition. Once a user uttered five consecutive specification- or query-phrase utterances with Speech Graffiti-grammatical ASR hypotheses (ignoring any intervening keyword-only input), the system switched from required shaping to implicit shaping. The adaptation was bidirectional, but the reverse threshold was lower: if the user later issued three consecutive non-grammatical utterances, the system would switch back to required shaping.

9.2.2 Explicit shaping

The control condition for this study was chosen to be the explicit shaping condition from Study II. The explicit shaping approach can be seen as a middle ground between the two extremes of the adaptive approach, and it performed moderately well in Study II. The only change made to the explicit approach for Study III was to slightly shorten the prompt on most shaping instances: the initial sentence of the prompt, I think I heard "[ASR hypothesis]," was only included every third time the prompt was triggered. Otherwise it began, Next time, it would help to use Speech Graffiti….

9.3 Setup

Participants completed the first session in a conference room at Carnegie Mellon University and the remainder of the sessions remotely. In pre-study briefings immediately before starting the first session, each participant was given a unique user ID to identify them in subsequent sessions' calls. The Speech Graffiti systems were not altered to handle recognition or dialog management of user IDs. Participants

were thus instructed to simply say their number at the start of each call, so that it would be recorded by the system to be picked up in the transcription process, and then to say **`start over`** to flush any side effect of the ID input.

In subsequent sessions, participants called the system on their own time, from a location of their choosing. Approximately once a week, I sent an email to all remaining participants with the deadline for making the next session's call and a link to a webpage containing that session's tasks. After completing each session, users filled in the webpage form and submitted it online. In addition to recording the information that users found for each task, the web form collected information about the time and date of the call, whether it had been made from a cell or land-line phone, whether it had been made from a public or private location, and what the environmental noise level had been during the call. Table 9.2 summarizes the distribution of calls over the five remote-calling sessions of the study.

Overall, there were 148 interactions with the system over the six sessions. Users were generally given about a week from the time each email was sent to make their call

Table 9.2. Summary of call characteristics for Study III sessions two through six.

		%
Type of phone	Land line	61.0
	Cell phone	39.0
Environment	Private	95.0
	Public	5.0
Noise level	1 (quiet)	46.2
	2	43.6
	3	7.7
	4	2.6
	5 (noisy)	0

N = 117 calls

and submit the task answers. There was a full week between the end of the week one sessions and the session two email, so there were at least eight days between each user's first and second interactions. The minimum observed time between session calls was two days (since some users called towards the end of one call period and then called early in the next one) and the maximum observed time was 16 days. The mean time between calls was 7.7 days.

9.4 Tasks

Participants worked on eight tasks during the first, in-office session and six tasks during session five (the first DineLine session). All other sessions had four tasks. The tasks used the same task difficulty scheme as in other sessions, and the actual distribution of task difficulties by session is shown in table 9.3. Appendix D lists the DineLine tasks used in session five.

There were no pre-use tutorials for Study III. Before making their calls in the first session, participants were reminded to listen to the introduction at the start of the call, just as in Study II. The actual introduction recording was the same as in Study II, with the addition of information at the end to inform users that they could say **`introduction`** to replay that at any time. In the first session, barge-in was disabled for the introduction so that all users would hear the entire message. In subsequent

Table 9.3. Task difficulty level orders for each session's tasks in Study III.

Session	System	Task difficulty level order
One	MovieLine	1 - 1 - 1 - 2 - 2 - 3 - 1 - 2
Two	MovieLine	1 - 1 - 2 - 3
Three	MovieLine	1 - 2 - 3 - 1
Four	MovieLine	1 - 2 - 1 - 2
Five	DineLine	1 - 1 - 2 - 3 - 1 - 2
Six	DineLine	1 - 2 - 1 - 2

sessions, barge-in was enabled for the introduction so that users could skip it if they wished to. As in the previous studies, users were given the tasks on a sheet of paper and asked to work through them, writing down the answers for each. The tasks were printed in the same HTML-page format that was used in subsequent sessions so that users would have a chance to familiarize themselves with what the later sessions' web forms would look like. Participants were given 20 minutes to complete the eight tasks. Since in subsequent sessions users called the system on their own, there were no time limits for sessions two through six. The average call length over the five off-site sessions was 8.6 minutes.

To motivate users to complete tasks successfully, participants were compensated for their time with a flat cash payment for participation ($32.00) plus an additional amount (40 cents) per correctly completed task. To encourage users to remain in the study, participants were paid only upon successful completion of calls for all six sessions.

9.5 Speech Graffiti DineLine

The ATUE study included an initial attempt at assessing cross-domain skill transfer with Speech Graffiti. 15 of the 23 ATUE participants also interacted with a Speech Graffiti apartment information system after their interaction with the MovieLine, and grammaticality rates were significantly lower in the ApartmentLine interactions. However, the domains and study tasks did not truly support identical functionality: the ApartmentLine allowed users to query all slots in a single query (e.g. a user could say area is Squirrel Hill, go ahead, and the system would return a query result listing, for each apartment, information for all of the slots that had not been specified, such as address, number of bedrooms, distance from campus, etc.), and this functional difference seemed confusing for users. Thus, in Study III, I wanted to assess cross-

domain transfer in a more comparable system, and also to survey the effects of transfer after several interactions with the initial system.

To assess cross-domain transfer, I created the Speech Graffiti DineLine, which provides information about Pittsburgh restaurants. One argument for using Speech Graffiti is the ease of developing applications in new domains, but application development becomes more complex if expanded grammars need to be created to support the two-pass ASR method and shaping. One solution to this issue would be that in the possible universe of Speech Graffiti applications, there would be a handful of two-pass, shaping applications and the rest would be standard Speech Graffiti systems. The shaping applications could be considered as training applications, and users who had achieved Speech Graffiti proficiency through the shaping process could then transfer their skills to non-shaping Speech Graffiti systems. I decided to explore this possibility in Study III by implementing the DineLine as a non-shaping Speech Graffiti application.

Like the MovieLine, the DineLine system has nine slots: restaurant names, addresses, phone numbers, neighborhoods, cuisines, price ranges, and star ratings, plus the days of the week that each restaurant is open and the meals served on those days. The database contains this information for 150 restaurants in Pittsburgh; the data was scraped mostly from *Pittsburgh* magazine's online restaurant database.[6] Fig. 9.1 shows an excerpt of an actual DineLine interaction from Study III, and Appendix E shows the full Speech Graffiti DineLine grammar. Table 9.4 shows the lexicon sizes of the MovieLine and DineLine grammars used in Study III.

[6] http://www.wqed.org/mag/cityguide/dining_00.shtml

Table 9.4. Lexicon sizes for User Study III. In this table, lexicons from both MovieLine grammars include variations for 50 movie titles.

	Size of lexicon	
	With linked words	With all words unlinked
Speech Graffiti MovieLine	508	365
Expanded MovieLine	614	413
Speech Graffiti DineLine	523	411

9.6 User surveys

Upon finishing their call in session one, participants were asked to fill out the same SASSI-based questionnaire used in Study II. After submitting their tasks via the web form in sessions four and six, users were directed to a second page and asked to fill out the same user satisfaction questionnaire again. Due to a programming oversight, the four additional Speech Graffiti-specific questions (*e.g.*, *I understand how to use Speech Graffiti, etc.*) were not included in the fourth session's survey, but they were included in the other two. The final session's survey also asked users to report what they liked most and least about the system and to indicate their occupation.

9.7 Analysis

The basic analyses conducted for Study III were similar to those conducted in the previous studies: Speech Graffiti grammaticality, overall task completion, mean time- and turns-to-completion rates, median time- and turns-on-task, and mean scores for each of the six user satisfaction factors plus a combined, overall user satisfaction rating. Because each participant in Study III had multiple interactions with the systems, many of these measures were assessed within-subject (*i.e.*, across the study sessions) as well as between-subjects as in the other studies.

Start over.

Starting over

Restaurant is Church Brew Works, list phone number.

Church Brew Works, listing 1 phone number: 412-688-8200.

Start over.

Starting over

Area is South Side.

South Side

American food.

{confsig}, area is South Side.

Food type is American.

{confsig}, cuisine is Indian.

Start over.

Starting over.

Area is South Side.

South Side.

Cuisine is American.

American.

List restaurants.

Listing 7 restaurants: City Grill, Hot Metal Grille, Mario's Southside Saloon, and more

More.

Pittsburgh Steak Co., Primanti Brothers, The Bridge Cafe, The Cheesecake Factory.

Figure 9.1. Excerpt from a Study III DineLine interaction. Since there is no shaping in this system, non-Speech Graffiti input generates a {confsig} (plus a restatement of what the system last entered into its context, if appropriate). The system response to the user input **Food type is American** is due to an ASR error.

Chapter 10

User Study III Results: Adaptive Shaping

Study III investigated the adaptive shaping strategy, longitudinal effects, and cross-domain transfer with the Speech Graffiti DineLine. The adaptive strategy actually generated slower interactions in the first session, but efficiency measures were similar between the two groups in subsequent interactions. Performance with the MovieLine increased significantly in nearly every aspect over the first four sessions. Users in the adaptive condition performed somewhat better in the DineLine, though cross-domain grammaticality increased for users in both groups.

10.1 Efficiency

10.1.1 Task completion

There were 30 tasks total across the six-session study period. Users in the adaptive group completed somewhat more tasks overall than users in the explicit group. Although users in both groups completed about the same number of tasks in the MovieLine sessions, users in the adaptive condition completed significantly more tasks in each of the two DineLine sessions, as shown in Table 10.1.

To assess task completion longitudinally and across domains, I looked at task subsets, since each session did not have the same number of tasks. First, I compared within-subject completion data between sessions one, four, and six. Sessions four and six had the same four-task task difficulty pattern, and to match that, I looked at the subset of tasks #3, 4, 7, and 8 from session 1. Overall, there were significant increases from session one (mean, 3.26) to session four (mean, 3.96; t = 3.43, p = 0.002) with similar increases in both conditions. Task completion dropped from session four to session six (mean, 3.27; t = -3.07, p = 0.006). This drop was more marked for users in the explicit group (F = 4.13, p = 0.06); as noted above users in

Table 10.1. Comparison of mean number of tasks completed in each session of Study III.

	Total tasks in session	Adaptive		Explicit			
		M	S.D.	M	S.D.	*t*	*p*
Session 1	8	6.77	1.17	6.43	1.79	0.59	0.56
Session 2	4	3.64	0.50	3.43	1.28	0.55	0.59
Session 3	4	3.20	0.79	3.21	0.89	-0.04	0.97
Session 4	4	4	0	3.93	0.27	1.00	0.34
Session 5	6	5.67	0.50	4.29	1.86	2.64	0.02
Session 6	4	3.88	0.35	2.93	1.07	3.03	0.008
Overall	30	27.3	2.76	24.6	5.33	1.55	0.14

the adaptive group completed more tasks in session six.

Next, to assess the difference in performance between users' first session with the MovieLine and their first session with the DineLine, I used the subset of the six tasks #1, 2, 5, 6, 7, and 8 from session one to match those of session five. There was no real mean overall change ($F = 3.70$, $p = 0.07$), but the trend for each of the two conditions was in opposite directions: compared to session one, adaptive users completed significantly more tasks in session five than in session one ($t = 2.29$, $p = 0.05$), while explicit users generally completed fewer ($t = -1.04$, $p = 0.32$).

10.1.2 Time

In the first session, users spent significantly more time on completed tasks in the adaptive condition (mean, 87.8 seconds, S.D. = 27.9) than in the explicit condition (mean, 67.6 seconds, S.D. = 19.1) ($t = 2.18$, $p = 0.04$). It is possible that this difference is due to the prevalence of the longer, required-state adaptive prompts while users are initially learning the system. Beyond the first session, time-on-task rates were similar for both groups.

Fig. 10.1 shows mean time-to-completion rates and median time-on-task rates by shaping condition. Mean time-to-completion generally followed the same pattern across sessions for both conditions. The difference in median time-on-task is likely related to the differences in overall task completion between the two conditions: particularly in sessions five and six, explicit condition users spent significant amounts of time trying to figure out tasks but never quite got them correct.

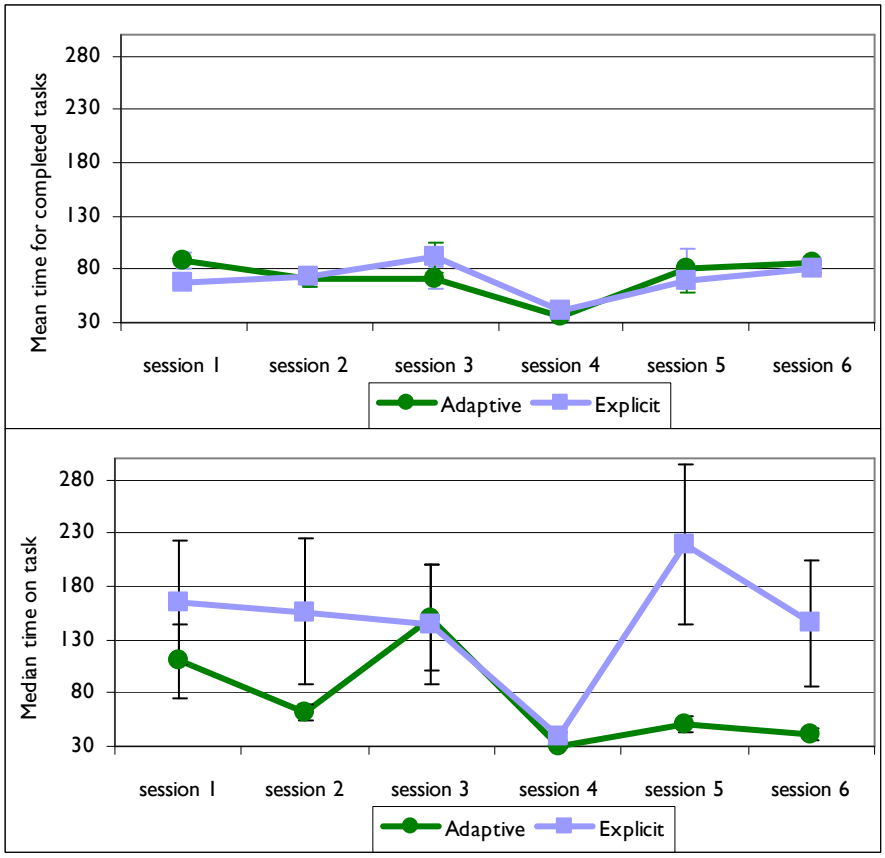

Figure 10.1. Mean time-to-completion (top) and median time-on-task (bottom) rates for each session in Study III, by shaping condition.

In the same within-subject subset analyses from the previous subsection, mean time-on-task decreased across conditions from session one (mean, 58.5 seconds) to session four (mean, 37.0 seconds; t = -2.00, p = 0.06) and increased in session six (mean, 49.5 seconds; t = 2.18, p = 0.04). The increase in time between sessions four and six was somewhat steeper for users in the adaptive condition (F = 3.22, p = 0.09). Comparing the initial sessions with the MovieLine and DineLine systems, overall time-on-task was nearly identical (74.7 seconds vs. 73.8 seconds).

10.1.3 Turns

Following the pattern of time-to-completion rates, users in the adaptive condition took significantly more turns per completed task (mean, 9.23) in session one compared to those in the explicit condition (mean, 6.44; t = 2.87, p = 0.009), with more similar turns-to-completion rates in subsequent sessions. Fig. 10.2 shows the session-by-session patterns of mean turns-to-completion and median turns-on-task by condition, and the overall pattern is similar to that of the time-based rates. In the within-subject subset analyses, turns-on-task decreased somewhat from session one (mean, 6.32) to session four (mean, 4.52; t = -1.91, p = 0.07) and increased in session six (mean, 6.82; t = 3.68, p = 0.02). The rate of increase in turns between sessions

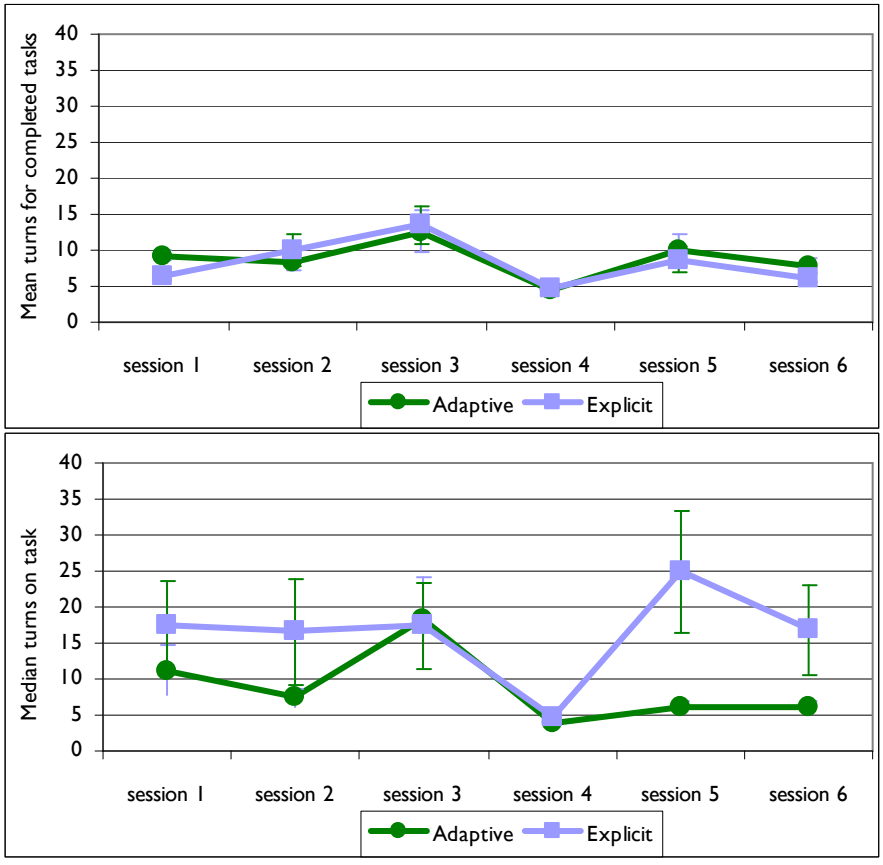

Figure 10.2. Mean turns-to-completion (top) and median turns-on-task (bottom) rates for each session in Study III, by shaping condition.

was similar for both groups. Comparing the initial sessions with both systems, turns-on-task was somewhat higher in session five (7.4 vs. 9.2), but not significantly so.

10.2 User satisfaction

User satisfaction ratings were quite similar between the two conditions, with insignificant differences between the two groups on any of the factors for any of the three survey points (fig. 10.3). In general, overall user satisfaction results generally increased from session one to session four, with strong increases in the habitability and cognitive demand factors (table 10.2). Likeability increased significantly more strongly for users in the adaptive group. Satisfaction then decreased in the session six survey, with significant drops in the system response accuracy, annoyance, and speed factors. Besides likeability, there were no between-groups differences in the intersession user satisfaction ratings change.

The drop in user satisfaction scores after the final DineLine interaction may have been partially due to ASR factors: overall (unlinked) word-error rates were significantly higher for most users in session 6 (mean, 43.4%) compared to session 4 (30.5%; $t = 2.87$, $p = 0.009$). The SASSI questionnaire does not include any factors that explicitly measure user perception of speech recognition quality, but the system response accuracy includes items such as *the system is accurate* and *the system makes few errors*, while the annoyance factor includes items like *the interaction with the system is frustrating*. It seems possible that ratings for both of these factors could have been influenced by lower ASR quality.

10.3 Grammaticality

Speech Grammaticality generally increased over the course of the study. Fig. 10.4 shows mean Speech Graffiti grammaticality for the two conditions for each session.

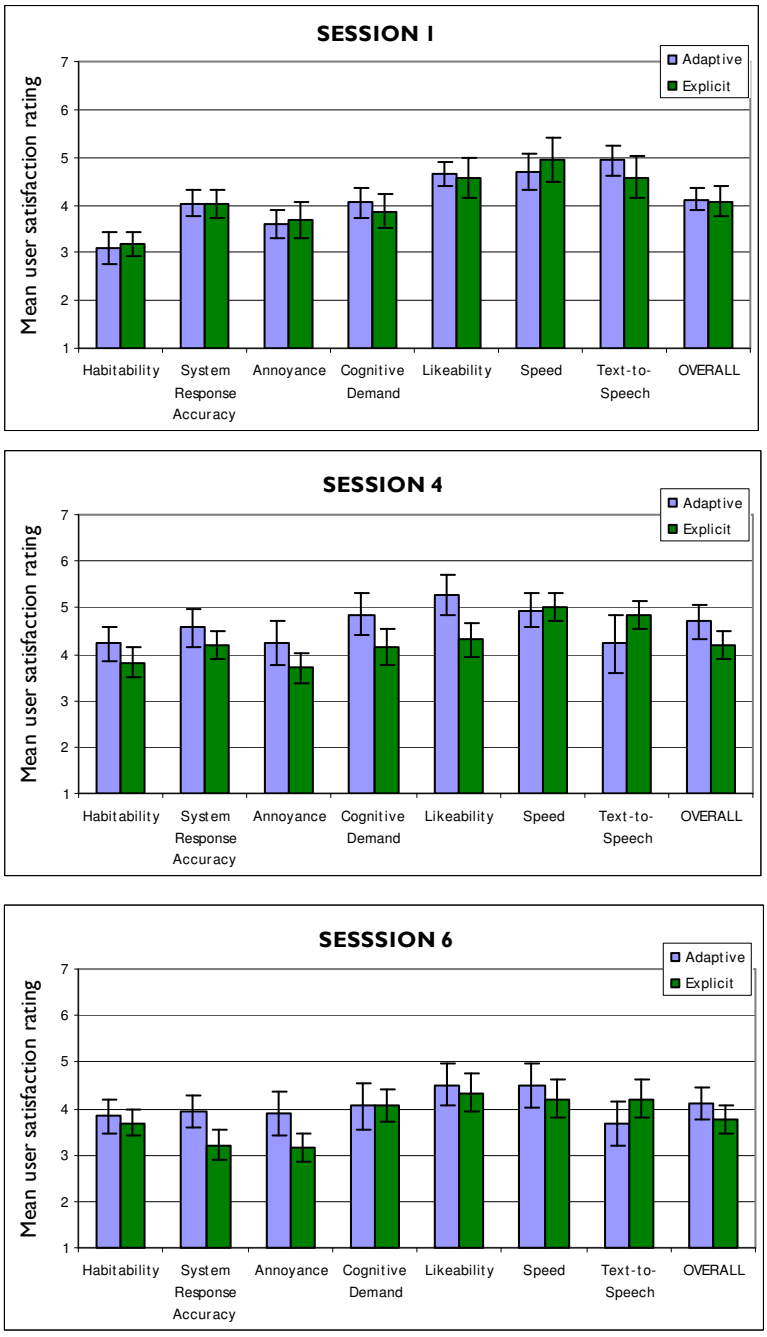

Figure 10.3. User satisfaction ratings for each survey point (weeks 1, 4 & 6) of Study III.

At session two, users in the adaptive condition had slightly higher grammaticality than those in the explicit condition ($t = 1.63$, $p = 0.12$). Grammaticality for adaptive

Table 10.2. Summary of longitudinal user satisfaction changes in Study III for all participants.

		Session 1 to session 4 change			Session 4 to session 6 change		
Factor	Session 1 mean	*t*	*p*	Session 4 mean*	*t*	*p*	Session 6 mean
System response accuracy	4.14	1.05	0.31	4.34 / 4.41	-4.00	< 0.001	3.48
Likeability	4.64	0.20	0.84†	4.68 / 4.67	-1.37	0.19	4.40
Cognitive demand	3.92	2.61	0.02	4.43	-1.55	0.14	4.05
Annoyance	3.77	0.63	0.53	3.91 / 3.90	-2.13	0.05	3.42
Habitability	3.24	3.01	0.007	3.98 / 3.92	-0.67	0.51	3.75
Speed	5.02	-.023	0.82	4.98	-2.97	0.008	4.32
Text-to-speech	4.63	-0.07	0.95	4.61 / 4.55	-1.50	0.15	4.02
Overall	4.16	1.58	0.13	4.39	-2.94	0.008	3.89

* Number after slash is adjusted session 4 mean excluding data from one participant who completed session 4 but not session 6

† Stronger positive change for adaptive condition: $F = 4.44$, $p = 0.05$

users dropped relatively markedly ($F = 3.57$, $p = 0.07$) in session three to about the same level as that of explicit condition users, and it rose significantly again for both groups in session four ($t = 4.57$, $p < 0.001$). The decrease in grammaticality from session two to session three for adaptive users is somewhat puzzling, although several users experienced ASR difficulties with some of the task vocabulary that week. However, this issue was experienced by users in both conditions (although of course they do appear to have converged at the same level of grammaticality for this week).

Over the longer term, regardless of condition, participants exhibited significantly higher grammaticality in sessions four (mean, 92.2) and six (mean, 82.1) compared to session one (mean, 73.0; $t = 9.26$, $p < 0.001$ for the former comparison; mean 72.8; $t = 3.15$, $p = 0.005$ for the latter).

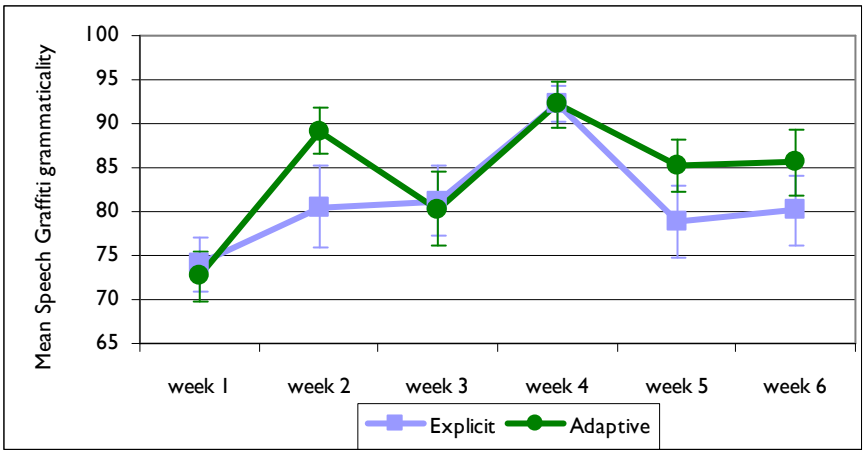

Figure 10.4. Mean Speech Graffiti grammaticality for each session in Study III, by condition.

Session five marked the switch to the DineLine application, and grammaticality rates fell significantly overall compared to session four (from a mean of 92.2% to 81.4%; t = -4.73, $p < 0.001$), somewhat more so for users in the explicit condition ($F = 1.99$, $p = 0.17$). A matched-pairs analysis over all users showed that the initial DineLine grammaticality rates (mean, 81.4) were still significantly higher than the rates in the first MovieLine session (mean, 73.0; $t = 2.94$, $p = 0.008$), suggesting that users have effectively transferred their Speech Graffiti skills to the new domain. This change was somewhat stronger for users in the adaptive condition ($F = 2.47$, $p = 0.13$). In addition to the gross grammaticality change, there was also a marked difference in the number of users passing the 80% threshold between session one and session five. In the first MovieLine session, only six users (22%) spoke at or above the 80% grammaticality level, whereas with the first DineLine session, 16 users (70%) did.

10.3.1 Intra-session grammaticality

As in the previous studies, there was a significant intrasession grammaticality increase from the first (mean, 69.5%) to the last quarter (mean, 79.7%) of the first session ($t = 2.61$, $p = 0.02$) (fig. 10.5). The mean change was greater for the adaptive shaping

group (+14.8 points) than for the explicit group (+6.05 points), but the difference was not significant ($F = 1.25$, $p = 0.27$).

10.4 System performance

Participants in Study III generated 8,842 utterances over the course of the six interactions. 45.7% of these were from users in the adaptive group; the remaining 54.3% were from the explicit group. Fig. 10.6 shows the distribution of utterances collected over the six sessions.

In session one, WER was 25.5% (S.D. = 8.69) overall (using unlinked words) 15.7% (S.D. = 9.67) for Speech Graffiti grammatical input. The WER for Speech Graffiti-grammatical input in Study III represented an 24% relative decrease from the mean of 20.7% for the shaping condition in Study II, suggesting that the acoustic model adjustments made after Study II were effective ($t = 2.26$, $p = 0.03$).

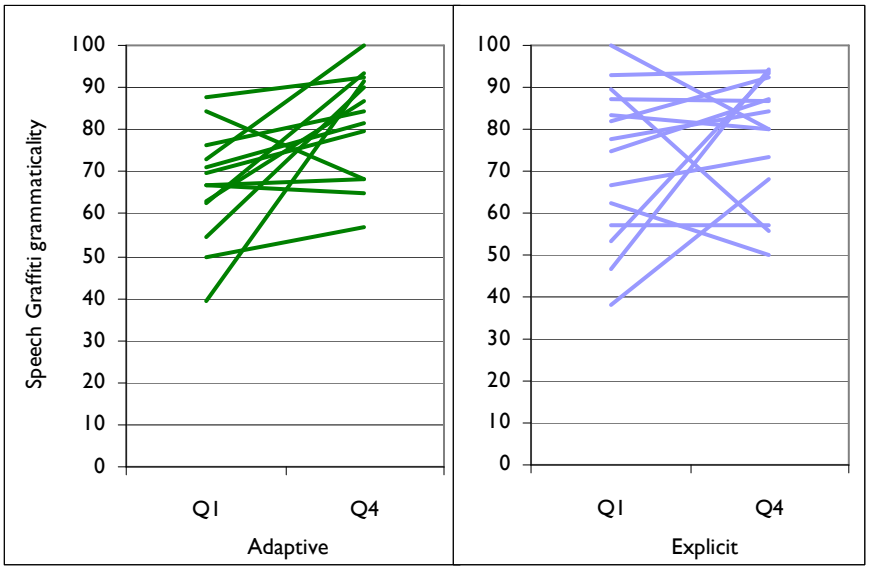

Figure 10.5. Speech Graffiti grammaticality change from first quarter of session 1 to final quarter of session 1, by condition. Each line represents one participant.

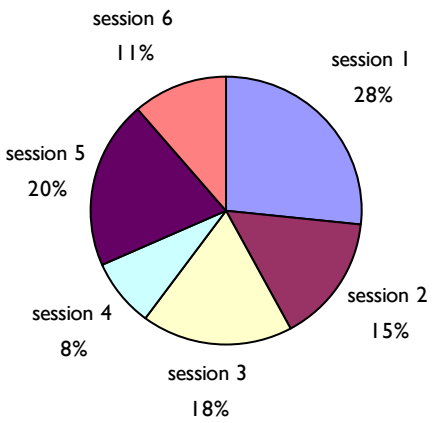

Figure 10.6. Distribution of user utterances from Study III by session.

Word-error rates varied significantly over the course of the study, although they did not vary across shaping conditions. Fig. 10.7 shows WER over the six sessions for all utterances and for grammatical utterances only. At first glance, the pattern of WER over the six sessions looks quite similar to the patterns of time- and turns-to-completion in figs. 10.1 and 10.2, with a peak in session three and increases in sessions five and six. Two things should be noted here however. First, word-error rates increased significantly between sessions one and four, both in terms of only Speech Graffiti grammatical utterances and overall. Despite this increase, there were increases across all participants in user satisfaction and task completion, and

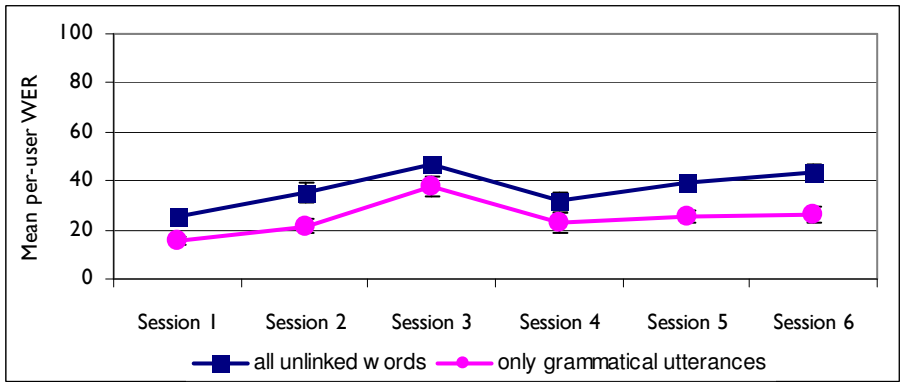

Figure 10.7. Mean per-user word error rates in Study III. The top line shows the overall WER for unlinked words; the lower line shows WER for Speech Graffiti grammatical utterances only.

decreases in time- and turns-to-completion. This suggests that longer-term use of Speech Graffiti supports increased efficiency even when word-error rates increase. Similarly, WER increased significantly between the first session of the MovieLine and the first DineLine session, and users in the adaptive condition were still able to complete significantly more tasks in the DineLine session than in the MovieLine session.

10.5 Discussion

The goal of User Study III was to investigate the following three questions: did the adaptive shaping strategy have a beneficial effect on interaction efficiency? How did interaction efficiency change over four sessions of interaction with the MovieLine system? How did interaction efficiency change across domains, with the introduction of the Speech Graffiti DineLine?

As measured on the first interaction (thus matching the single interaction setup of the previous two studies), the adaptive strategy appears to have had an unfavorable effect on interaction efficiency. Mean time- and turns-to-completion were significantly higher for users in the adaptive condition during the first session, although grammaticality, user satisfaction and task completion rates were similar. The differences in time may simply be related to the occurrence and length of the adaptive prompt. That is, the first session had the lowest grammaticality, thus the required state of the adaptive prompt would have been triggered more often, contributing extra time and turns to the interaction. However, there is a lack of supporting correlation between grammaticality and time or turns in the data from this study.

Significant intrasession grammaticality increases in session one confirm the results from the previous studies that users can learn the system simply by interacting with it. The lack of a strong effect of shaping condition suggests that the specific nature

of the shaping prompt may not be as important as the fact that some sort of shaping support exists.

Beyond the first week, users in both shaping conditions generally performed at about the same rate. Over the four sessions with the MovieLine, task completion rates increased significantly, and both time- and turns-on-task decreased significantly. User satisfaction rates trended upwards, with significant increases in the habitability and cognitive demand factors. These changes occurred despite an increase in word-error rate over the course of the four sessions. These factors suggest that over time, users found the system easier to use and to speak with, regardless of any ASR error issues. Grammaticality also increased significantly from the first to the fourth sessions.

The DineLine interactions in sessions five and six showed that users can successfully transfer skills learned on a shaping application to a non-shaping Speech Graffiti system. The adaptive shaping strategy seemed to provide a cross-domain advantage, as adaptive users completed significantly more tasks in the DineLine sessions than explicit condition users did. Adaptive group users also completed significantly more tasks in the first DineLine session than in the first MovieLine session, despite a higher WER in the DineLine session (users in the explicit condition did not exhibit such strong performance in this situation). Although adaptive users tended to have slightly stronger increases in grammaticality between sessions one and five, skill transfer was generally evident across all users through increased grammaticality, with mean grammaticality on the first DineLine session significantly higher than on the first MovieLine session and with more users attaining the 80% grammaticality threshold.

10.5.1 Key findings from User Study III

- Overall intra- & intersession convergence.

- Evidence of cross-domain transfer to a standard, non-shaping, no-tutorial Speech Graffiti application.

- Users in the adaptive group took more time and turns for completed tasks in initial session, but showed significantly increased cross-domain task completion, significantly increased likeability over time, and a trend towards stronger convergence to Speech Graffiti (as measured by increased grammaticality).

- More efficient interactions for adaptive group users in sessions four and five compared to session one despite increases in word-error rates.

Chapter 11

Conclusion

11.1 Summary of results

User Study I: In Study I there was a trend towards increased efficiency and satisfaction for users in the shaping group. Users in the shaping group performed relatively similarly regardless of whether or not they had had a tutorial, suggesting that a tutorial is not strictly necessary. Intrasession convergence to the Speech Graffiti form was observed across all participants, with stronger convergence for users in the original group. Finally, Study I demonstrated the successful deployment of the two-apps ASR strategy.

User Study II: Significant intrasession convergence was observed across all groups in Study II. Participants in the required group demonstrated significant local convergence, but also somewhat lower user satisfaction scores. The required shaping

condition also proved to be less robust to ASR errors. Users in the implicit shaping condition gave the system lower habitability ratings.

User Study III: Significant intra- and intersession convergence and cross-domain skill transfer was observed across all groups in Study III. Users in the adaptive condition took more time and turns for completed tasks in initial session, but showed significantly increased cross-domain task completion, significantly increased likeability over time, and a trend towards stronger convergence to Speech Graffiti. Users in the adaptive group had more efficient interactions in sessions four and five compared to session one despite increases in word-error rates.

11.2 Contributions

This thesis has presented work on improving interaction efficiency with spoken dialog systems via a process of shaping user input to convergence with a more efficient interaction protocol (*i.e.*, Speech Graffiti), with the following contributions:

- Three user studies were conducted to explore factors that can effectively shape user input and when shaping should occur. Not surprisingly, a strategy of requiring users to rephrase their input had by far the strongest effect on local convergence. However, this strategy was also prone to annoying errors in cases of poor ASR performance. Users generally exhibited intrasession grammaticality increases regardless of the shaping strategy they interacted with, attesting to the power of convergence as a general phenomenon.

- Although users in all of the experimental conditions became more grammatical over time, users in the baseline (non-shaping) condition in Study I did as well, at a comparatively steeper rate. However, any gain these users may have made in on-task efficiency is reduced by the necessity of having to undertake a pre-use tutorial. On the other hand, users in the shaping

conditions were able to skip the tutorial without any corresponding decline in interaction efficiency compared to those who had had a tutorial. The integration of shaping and the two-pass recognition process allows users to complete tasks while using natural language and learning the Speech Graffiti format. Results from Study III showed that over time, users have more efficient interactions with the system even when faced with higher word-error rates.

- Convergence has been observed on various levels in human-human and human-computer communication, but to my knowledge it has not actually been exploited to improve interaction with computer systems. The studies in this thesis have demonstrated a fully-functional, non-directed-dialog system, accessing real-world data, that takes advantage of users' propensity for convergence. Overall, across all of the shaping conditions studied in this work, increases in Speech Graffiti grammaticality correlated significantly with increases in task completion (0.39, $p < 0.001$), decreases in time to complete tasks (-0.31, $p = 0.006$), and decreases in overall word-error rates (-0.53, $p < 0.001$)

11.3 Extensions of the work

The idea for this work came from user experiences in the ATUE study, which generated the question, "how can we help users have more efficient interactions?" This thesis has offered some suggestions, but there is still work to be done to make spoken language interaction with computers as efficient as it is with other humans. A few areas related to this work are discussed here. One of the most interesting extensions of this work would be to make the system more widely available to the public, similar to the Let's Go! system for Pittsburgh bus information (Raux, Bohus,

Langner, Black, & Eskenazi, 2006). This would allow for an even more thorough examination of convergence in functional systems.

In the study sessions that took place in an office or conference room, users were frequently observed making notes about how to speak to the system during the introduction or help prompts. This indicates that for many users, a persistent visual reminder is an effective learning aid. Could this observation be leveraged by integrating Speech Graffiti into a multimodal system? The highly structured format of Speech Graffiti seems like it would fit naturally with a structured visual display format. Would users be able to transfer skills from multimodal to speech-only interactions?

To date, the most extensive evaluations of Speech Graffiti have involved basic information access domains, although work has also been done on simple device control. How would Speech Graffiti scale up to an application containing tens or even hundreds of slots? Given the inherent complexity of such an application, Speech Graffiti seems like a more reasonable approach from a development perspective than, say, a directed dialog or natural language system. How could Speech Graffiti be expanded to handle the user interface issues likely to emerge in such an interface? What other types of interactions might be necessary besides constraint specification and querying, and how could they be made habitable for users? How could shaping interfaces for a smaller applications (like those studied here) help users learn skills for a larger application that may have more functions?

As the shaping strategies discussed in this work essentially encourage users to say things that fall within the system's preferred grammar, they are likely to apply to other varieties of spoken dialog systems besides structured interactions like Speech Graffiti. How might the shaping strategies be applied in a natural language system? For example, perhaps shaping could be used to encourage input that, while not

strictly more grammatical than other input, is more acoustically distinct and would thus be more likely to generate lower ASR error rates.

Appendix A.

Baseline (non-expanded) Speech Graffiti Phoenix grammar from User Study I.

```
## -------------------- valid utterance ------------------------
[Utt]
 ( +[PHRASES] *[GoPhrase] )
 ( [GoPhrase] )
 ( [KeyPhrase] )
 ( [NavPhrase] )
;
[PHRASES]
 ( [DATE=SLOT] [DATE=VALUE] )
 ( [GENRE=SLOT] [GENRE=VALUE] )
 ( [RATING=SLOT] [RATING=VALUE] )
 ( [AREA=SLOT] [AREA=VALUE] )
 ( [THEATER=SLOT] [THEATER=VALUE] )
 ( [MOVIE=SLOT] [MOVIE=VALUE] )
 ( [SHOWTIME=SLOT] [TIME=VALUE] )
 ( [SHOWTIME=MACRO] )
 ( [WHAT] SLOTS )
 ( [WHAT-EST] [SHOWTIME=SLOT] )
 ( SLOTS [IS=ANYTHING] )
 ( *SLOTS [OPTIONS] )
 ( ERASER )
SLOTS
 ( [DATE=SLOT] )
 ( [AREA=SLOT] )
 ( [ADDRESS=SLOT] )
 ( [RATING=SLOT] )
 ( [GENRE=SLOT] )
 ( [PHONE=SLOT] )
 ( [THEATER=SLOT] )
 ( [SHOWTIME=SLOT] )
 ( [MOVIE=SLOT] )
ERASER
 ( [ClearContext] )
 ( [ClearUtterance] )
;
## -------------------- "what" queries ------------------------
[WHAT]
 ( what )
 ( what=is )
 ( what=are )
 ( requesting )
;
[WHAT-EST]
 ( what LATE-EARLY )
 ( what=is LATE-EARLY )
 ( what=are LATE-EARLY )
 ( requesting LATE-EARLY )
LATE-EARLY
```

```
 ( latest )
 ( earliest )
 ( the=latest )
 ( the=earliest )
;
[SHOWTIME=MACRO]
 ( [WHAT] [SHOWTIME=SLOT] [TIME=VALUE] )
;
## -------------------- keywords ------------------------
[NavPhrase]
 ( [More] )
 ( [Previous] [Hour] )
 ( [Previous] )
 ( [Next] [Hour] )
 ( [Next] )
 ( [First] [Hour] )
 ( [First] )
 ( [Last] [Hour] )
 ( [Last] )
;
[More]
 ( more )
;
[Previous]
 ( previous )
;
[Next]
 ( next )
;
[First]
 ( first )
;
[Last]
 ( last )
;
[ClearContext]
 ( start=over )
;
[ClearUtterance]
 ( scratch=that )
;
[GoPhrase]
 ( [Go] )
;
[KeyPhrase]
 ( [Restate] )
 ( [Repeat] )
 ( [Goodbye] )
 ( [Help] )
;
[Goodbye]
 ( goodbye )
;
[Help]
 ( help )
;
```

```
[Go]
 ( go=ahead )
;
[Restate]
 ( where=was=i )
 ( where=were=we )
 ( where=am=i )
 ( where=are=we )
;
[Repeat]
 ( repeat )
;
[IS=ANYTHING]
 ( anything )
;
[OPTIONS]
 ( options )
;
## ------------------- slots ---------------
[ADDRESS=SLOT]
 ( address=is )
 ( address )
 ( addresses )
 ( the=address )
 ( the=addresses )
;
[DATE=SLOT]
 ( day )
 ( day=is )
 ( date )
 ( date=is )
 ( days )
 ( dates )
 ( the=day )
 ( the=day=is )
 ( the=date )
 ( the=date=is )
 ( the=days )
 ( the=dates )
;
[GENRE=SLOT]
 ( the=genre )
 ( the=genre=is )
 ( the=genres )
 ( genre )
 ( genre=is )
 ( genres )
;
[AREA=SLOT]
 ( the=location )
 ( the=location=is )
 ( the=area )
 ( the=area=is )
 ( the=city )
 ( the=city=is )
 ( the=neighborhood=is )
```

```
 ( the=neighborhood )
 ( the=locations )
 ( the=areas )
 ( the=cities )
 ( the=neighborhoods )
 ( location )
 ( location=is )
 ( area )
 ( area=is )
 ( city )
 ( city=is )
 ( neighborhood=is )
 ( neighborhood )
 ( locations )
 ( areas )
 ( cities )
 ( neighborhoods )
;
[PHONE=SLOT]
 ( phone=number=is )
 ( phone=number )
 ( phone=numbers )
 ( the=phone=number )
 ( the=phone=numbers )
;
[RATING=SLOT]
 ( rating=is )
 ( rating )
 ( ratings )
 ( the=rating )
 ( the=ratings )
;
[SHOWTIME=SLOT]
 ( time=is )
 ( show=time )
 ( time )
 ( show=time=is )
 ( start=time=is )
 ( start=time )
 ( starting=time=is )
 ( starting=time )
 ( show=times )
 ( times )
 ( start=times )
 ( starting=times )
 ( showings )
 ( the=time=is )
 ( the=show=time )
 ( the=time )
 ( the=show=time=is )
 ( the=start=time=is )
 ( the=start=time )
 ( the=starting=time=is )
 ( the=starting=time )
 ( the=show=times )
 ( the=times )
```

```
 ( the=start=times )
 ( the=starting=times )
 ( the=showings )
;
[THEATER=SLOT]
 ( theater=is )
 ( movie=theater )
 ( theater )
 ( movie=theater=is )
 ( theaters )
 ( movie=theaters )
 ( theaters=are )
 ( movie=theaters=are )
 ( the=theater=is )
 ( the=movie=theater )
 ( the=theater )
 ( the=movie=theater=is )
 ( the=theaters )
 ( the=movie=theaters )
 ( the=theaters=are )
 ( the=movie=theaters=are )
;
[MOVIE=SLOT]
 ( movie=is )
 ( movie )
 ( title )
 ( title=is )
 ( movies )
 ( titles )
 ( movies=are )
 ( titles=are )
 ( the=movie=is )
 ( the=movie )
 ( the=title )
 ( the=title=is )
 ( the=movies )
 ( the=titles )
 ( the=movies=are )
 ( the=titles=are )
;
## ------------------- values ---------------
[DATE=VALUE]
 ( [Date=Constraint] )
;
[Date=Constraint]
 ( INTERVAL )
 ( SEMI-INTERVAL [Date] )
 ( *on [Date] )
INTERVAL
 ( between [LoBoundDate] and [HiBoundDate] )
 ( after [LoBoundDate] before [HiBoundDate] )
SEMI-INTERVAL
 ( [Is=Before=Date] )
 ( [Is=After=Date] )
;
[LoBoundDate]
```

```
 ( [Date] )
;
[HiBoundDate]
 ( [Date] )
;
[Is=Before=Date]
 ( before )
 ( earlier=than )
;
[Is=After=Date]
 ( after )
 ( later=than )
;
[Date]
 ( [Relative=Date] )
 ( [Calendar=Date] )
;
[Relative=Date]
 ( [Weekday] )
 ( [rel=date=mod] [Weekday] )
 ( [rel=date] )
;
[rel=date=mod]
 ( last )
 ( next )
 ( this )
;
[Weekday]
 ( sunday )
 ( monday )
 ( tuesday )
 ( wednesday )
 ( thursday )
 ( friday )
 ( saturday )
;
[rel=date]
 ( yesterday )
 ( today )
 ( tomorrow )
;
[Calendar=Date]
 ( [month] [ordinal] )
;
[month]
 ( january )
 ( february )
 ( march )
 ( april )
 ( may )
 ( june )
 ( july )
 ( august )
 ( september )
 ( october )
 ( november )
```

```
 ( december )
;
[ordinal]
 ( first )
 ( twenty=first )
 ( thirty=first )
 ( second )
 ( twenty=second )
 ( third )
 ( twenty=third )
 ( fourth )
 ( twenty=fourth )
 ( fifth )
 ( twenty=fifth )
 ( sixth )
 ( twenty=sixth )
 ( seventh )
 ( twenty=seventh )
 ( eighth )
 ( twenty=eighth )
 ( ninth )
 ( twenty=ninth )
 ( tenth )
 ( eleventh )
 ( twelfth )
 ( thirteenth )
 ( fourteenth )
 ( fifteenth )
 ( sixteenth )
 ( seventeenth )
 ( eighteenth )
 ( nineteenth )
 ( twentieth )
 ( thirtieth )
;
[GENRE=VALUE]
 ( action )
 ( adventure )
 ( animation )
 ( comedy )
 ( crime )
 ( documentary )
 ( drama )
 ( family )
 ( fantasy )
 ( film-noir )
 ( foreign )
 ( horror )
 ( music )
 ( musical )
 ( mystery )
 ( romance )
 ( sci-fi )
 ( short )
 ( sport )
 ( thriller )
```

```
 ( war )
 ( western )
 ( not=available )
;
[AREA=VALUE]
 ( aspinwall )
 ( bellevue )
 ( bridgeville )
 ( century=3 )
 ( cheswick )
 ( cranberry=township )
 ( dormont )
 ( downtown )
 ( east )
 ( homestead )
 ( irwin )
 ( monroeville )
 ( mount=lebanon )
 ( near=c=m=u )
 ( north )
 ( north=hills )
 ( north=side )
 ( north=versailles )
 ( oakland )
 ( oakmont )
 ( penn=hills )
 ( pittsburgh )
 ( pleasant=hills )
 ( regent=square )
 ( robinson )
 ( south )
 ( south=side )
 ( squirrel=hill )
 ( west )
 ( west=mifflin )
 ( edgewood )
 ( moon=township )
;
[RATING=VALUE]
 ( g )
 ( p=g )
 ( p=g=thirteen )
 ( r )
 ( n=c=seventeen )
 ( not=rated )
 ( not=available )
;
[THEATER=VALUE]
 ( *the PTHEATER *THEATER )
THEATER
 ( theater )
 ( cinema )
 ( cinemas )
 ( screens )
PTHEATER
 ( [Carmike=10=-=Pittsburgh] )
```

```
 ( [Carmike=Cranberry=8] )
 ( [Carmike=Galleria=6] )
 ( [Carmike=Maxi=Saver=12] )
 ( [Carmike=Southland=9] )
 ( [Cheswick=Theatres] )
 ( [Cinema=4] )
 ( [Cinemagic=Bellevue=Theater] )
 ( [Cinemagic=Denis=4=Theatres] )
 ( [Cinemagic=Manor=Theatre] )
 ( [Cinemagic=Squirrel=Hill] )
 ( [Dependable=Drive-In] )
 ( [Destinta=Theatres=-=Chartiers=Valley=20] )
 ( [Destinta=Theatres=-=Plaza=East=22] )
 ( [Flagstaff=Hill] )
 ( [Harris=Theatre] )
 ( [Loews=Waterfront=Theatre] )
 ( [Melwood=Screening=Room] )
 ( [Northway=Mall=Cinemas=8] )
 ( [Norwin=Hills=Cinemas] )
 ( [Oaks=Cinema] )
 ( [Omnimax=Theatre=-=Carnegie=Science=Center] )
 ( [Penn=Hills=Cinema] )
 ( [Regent=Square=Theatre] )
 ( [Showcase=Cinemas=Pittsburgh=North] )
 ( [Showcase=Cinemas=Pittsburgh=West] )
 ( [Southside=Works=Cinema] )
 ( [Star=City=Cinemas=-=S.=Fayette=14] )
 ( [University=Center] )
 ( [Waterworks=Cinemas] )
;
[Carmike=10=-=Pittsburgh]
 ( carmike=ten )
 ( carmike=village=ten )
 ( carmike=village=ten=pittsburgh )
 ( carmike=ten=pittsburgh )
;
[Carmike=Cranberry=8]
 ( carmike=eight )
 ( carmike=cranberry=eight )
 ( cranberry=eight )
 ( carmike=cranberry )
;
[Carmike=Galleria=6]
 ( carmike=galleria=six )
 ( galleria )
 ( galleria=six )
 ( carmike=galleria )
;
[Carmike=Maxi=Saver=12]
 ( carmike=maxi=saver )
 ( maxi=saver=twelve )
 ( carmike=maxi=saver=twelve )
 ( maxi=saver )
;
[Carmike=Southland=9]
 ( southland )
```

```
 ( carmike=southland )
 ( southland=nine )
 ( carmike=southland=nine )
;
[Cheswick=Theatres]
 ( cheswick )
 ( cheswick=quad )
 ( cheswick=quads )
;
[Cinema=4]
 ( cinema=four )
;
[Cinemagic=Bellevue=Theater]
 ( bellevue )
 ( cinemagic=bellevue )
;
[Cinemagic=Denis=4=Theatres]
 ( denis )
 ( cinemagic=denis )
 ( cinemagic=denis=four )
 ( denis=four )
;
[Cinemagic=Manor=Theatre]
 ( manor )
 ( cinemagic=manor )
;
[Cinemagic=Squirrel=Hill]
 ( cinemagic=squirrel=hill )
 ( squirrel=hill )
;
[Destinta=Theatres=-=Chartiers=Valley=20]
 ( destinta=bridgeville )
 ( destinta=theatres=chartiers=valley )
 ( destinta=chartiers=twenty )
 ( destinta=chartiers=valley=twenty )
 ( destinta=theatres=chartiers=valley=twenty )
 ( destinta=theatres=chartiers )
 ( destinta=theatres=chartiers=twenty )
 ( destinta=chartiers=valley )
 ( destinta=chartiers )
 ( chartiers=twenty )
 ( chartiers=valley=twenty )
 ( destinta=twenty )
;
[Destinta=Theatres=-=Plaza=East=22]
 ( destinta=north=versailles )
 ( destinta=plaza=east TWENTYTWO )
 ( destinta=plaza=east )
 ( destinta=theatres=plaza=east )
 ( destinta=theatres=plaza=east TWENTYTWO )
 ( plaza=east TWENTYTWO )
 ( plaza=east )
 ( destinta TWENTYTWO )
TWENTYTWO
 ( twenty=two )
;
```

```
[Dependable=Drive-In]
 ( dependable=drive=in )
 ( dependable )
;
[Flagstaff=Hill]
 ( flagstaff=hill )
 ( flagstaff )
 ( schenley )
 ( schenley=park )
;
[Harris=Theatre]
 ( harris )
 ( filmmakers=at=the=harris )
;
[Loews=Waterfront=Theatre]
 ( loews )
 ( loews=waterfront )
 ( waterfront )
 ( homestead=waterfront )
;
[Melwood=Screening=Room]
 ( filmmakers )
 ( melwood=screening=room )
 ( melwood )
;
[Northway=Mall=Cinemas=8]
 ( northway=mall=cinemas=eight )
 ( northway=mall=eight )
 ( northway=mall )
 ( northway=eight )
 ( northway )
;
[Norwin=Hills=Cinemas]
 ( norwin=hills )
;
[Oaks=Cinema]
 ( oaks )
;
[Omnimax=Theatre=-=Carnegie=Science=Center]
 ( CSC )
 ( omnimax=theatre )
 ( omnimax )
 ( omnimax=theatre CSC )
 ( omnimax CSC )
CSC
 ( carnegie=science=center )
 ( science=center )
;
[Penn=Hills=Cinema]
 ( penn=hills )
;
[Regent=Square=Theatre]
 ( regent=square )
;
[Showcase=Cinemas=Pittsburgh=North]
 ( showcase=cinemas=north )
```

```
 ( showcase=pittsburgh=north )
 ( showcase=north )
 ( showcase=cinemas=pittsburgh=north )
;
[Showcase=Cinemas=Pittsburgh=West]
 ( showcase=west )
 ( showcase=cinemas=pittsburgh=west )
 ( showcase=pittsburgh=west )
 ( showcase=cinemas=west )
;
[Southside=Works=Cinema]
 ( southside=works )
 ( southside )
;
[Star=City=Cinemas=-=S.=Fayette=14]
 ( star=city=cinema=south=fayette )
 ( star=city=fourteen )
 ( star=city )
 ( star=city=fayette=fourteen )
 ( star=city=cinema=fayette )
 ( star=city=cinema=fayette=fourteen )
 ( star=city=south=fourteen )
 ( star=city=south=fayette=fourteen )
 ( star=city=cinema=south=fayette=fourteen )
 ( star=city=cinema=south )
 ( star=city=cinema=fourteen )
 ( star=city=south=fayette )
 ( star=city=fayette )
 ( star=city=cinema )
 ( star=city=south )
 ( star=city=cinema=south=fourteen )
 ( south=fayette=fourteen )
 ( fayette=fourteen )
;
[University=Center]
 ( university=center )
 ( mcconomy )
 ( carnegie=mellon=university=center )
 ( u=c )
 ( c.=m.=u.=university=center )
 ( c.=m.=u. )
;
[Waterworks=Cinemas]
 ( waterworks )
;
[TIME=VALUE]
 ( [Time=Constraint] )
;
[Time=Constraint]
 ( INTERVAL )
 ( SEMI-INTERVAL [Time] )
 ( *at [Time] )
INTERVAL
 ( between [LoBoundTime] and [HiBoundTime] )
 ( after [LoBoundTime] before [HiBoundTime] )
SEMI-INTERVAL
```

```
 ( [Is=Before=Time] )
 ( [Is=After=Time] )
;
[LoBoundTime]
 ( [Time] )
;
[HiBoundTime]
 ( [Time] )
;
[Is=Before=Time]
 ( before )
 ( earlier=than )
;
[Is=After=Time]
 ( after )
 ( later=than )
;
[Time]
 ( [Hour] *o'clock *AM-PM )
 ( [Hour] [Minute] *AM-PM )
 ( noon )
 ( midnight )
AM-PM
 ( a=m )
 ( p=m )
;
[Hour]
 ( one )
 ( two )
 ( three )
 ( four )
 ( five )
 ( six )
 ( seven )
 ( eight )
 ( nine )
 ( ten )
 ( eleven )
 ( twelve )
;
[Minute]
 ( oh=five )
 ( ten )
 ( fifteen )
 ( twenty )
 ( twenty=five )
 ( thirty=five )
 ( thirty )
 ( forty=five )
 ( forty )
 ( fifty )
 ( fifty=five )
;
[MOVIE=VALUE]
 ( [the=longest=yard] )
 ( [the=honeymooners] )
```

```
( [cinderella=man] )
( [star=wars=episode=iii=-=revenge=of=the=sith] )
( [batman=begins] )
( [madagascar] )
( [mr.=and=mrs.=smith] )
( [the=sisterhood=of=the=traveling=pants] )
( [the=adventures=of=sharkboy=and=lavagirl=in=3-d] )
( [mad=hot=ballroom] )
( [ladies=in=lavender] )
( [mystery=of=the=nile] )
( [crash] )
( [monster-in-law] )
( [the=amityville=horror] )
( [are=we=there=yet?] )
( [beauty=shop] )
( [boogeyman] )
( [guess=who] )
( [hitch] )
( [miss=congeniality=2=armed=and=fabulous] )
( [the=pacifier] )
( [the=ring=two] )
( [sahara] )
( [lords=of=dogtown] )
( [bunty=aur=babli] )
( [howls=moving=castle] )
( [the=perfect=man] )
( [the=flavor=of=green=tea=over=rice] )
( [the=boys=and=girl=from=county=clare] )
( [enron=the=smartest=guys=in=the=room] )
( [brothers] )
( [rock=school] )
( [a=lot=like=love] )
( [herbie=fully=loaded] )
( [mondovino] )
( [the=interpreter] )
( [bewitched] )
( [george=a.=romeros=land=of=the=dead] )
( [spanglish] )
( [the=forgotten] )
( [scooby-doo=2=monsters=unleashed] )
( [the=aviator] )
( [bride=and=prejudice] )
( [oceans=twelve] )
( [shrek=2] )
( [without=a=paddle] )
( [the=prince=and=me] )
( [the=phantom=of=the=opera] )
( [harry=potter=and=the=prisoner=of=azkaban] )
( [the=terminal] )
( [finding=neverland] )
( [a=cinderella=story] )
( [van=helsing] )
( [the=dust=factory] )
( [napoleon=dynamite] )
( [the=spongebob=squarepants=movie] )
( [save=the=green=planet] )
```

```
 ( [house=of=wax] )
 ( [paheli] )
 ( [the=only=son] )
 ( [war=of=the=worlds] )
 ( [rebound] )
 ( [the=hitchhikers=guide=to=the=galaxy] )
 ( [hostage] )
 ( [sin=city] )
 ( [million=dollar=baby] )
 ( [high=tension] )
 ( [parineeta] )
 ( [walk=on=water] )
 ( [schizo] )
 ( [3-iron] )
 ( [kings=and=queen] )
;
[the=longest=yard]
 ( the=longest=yard )
 ( longest=yard );
[the=honeymooners]
 ( the=honeymooners )
 ( honeymooners );
[cinderella=man]
 ( cinderella=man );
[star=wars=episode=iii=-=revenge=of=the=sith]
 ( star=wars=episode=iii=-=revenge=of=the=sith )
 ( star=wars )
 ( star=wars=episode=iii )
 ( revenge=of=the=sith )
 ( star=wars=revenge=of=the=sith );
[batman=begins]
 ( batman=begins )
 ( batman );
[madagascar]
 ( madagascar );
[mr.=and=mrs.=smith]
 ( mr.=and=mrs.=smith );
[the=sisterhood=of=the=traveling=pants]
 ( the=sisterhood=of=the=traveling=pants )
 ( sisterhood=of=the=traveling=pants );
[the=adventures=of=sharkboy=and=lavagirl=in=3-d]
 ( the=adventures=of=sharkboy=and=lavagirl=in=3-d )
 ( sharkboy=and=lavagirl )
 ( the=adventures=of=sharkboy=and=lavagirl );
[mad=hot=ballroom]
 ( mad=hot=ballroom );
[ladies=in=lavender]
 ( ladies=in=lavender );
[mystery=of=the=nile]
 ( mystery=of=the=nile )
 ( the=mystery=of=the=nile );
[crash]
 ( crash );
[monster-in-law]
 ( monster-in-law );
[the=amityville=horror]
```

```
 ( the=amityville=horror )
 ( amityville=horror );
[are=we=there=yet?]
 ( are=we=there=yet? );
[beauty=shop]
 ( beauty=shop );
[boogeyman]
 ( boogeyman );
[guess=who]
 ( guess=who );
[hitch]
 ( hitch );
[miss=congeniality=2=armed=and=fabulous]
 ( miss=congeniality=2=armed=and=fabulous )
 ( miss=congeniality=2 )
 ( miss=congeniality=armed=and=fabulous );
[the=pacifier]
 ( the=pacifier );
[the=ring=two]
 ( the=ring=two )
 ( ring=two );
[sahara]
 ( sahara );
[lords=of=dogtown]
 ( lords=of=dogtown )
 ( the=lords=of=dogtown );
[bunty=aur=babli]
 ( bunty=aur=babli );
[howls=moving=castle]
 ( howls=moving=castle );
[the=perfect=man]
 ( the=perfect=man );
[the=flavor=of=green=tea=over=rice]
 ( the=flavor=of=green=tea=over=rice )
 ( flavor=of=green=tea )
 ( the=flavor=of=green=tea );
[the=boys=and=girl=from=county=clare]
 ( the=boys=and=girl=from=county=clare )
 ( the=boys=and=girls=from=county=clare );
[enron=the=smartest=guys=in=the=room]
 ( enron=the=smartest=guys=in=the=room )
 ( enron );
[brothers]
 ( brothers );
[rock=school]
 ( rock=school );
[a=lot=like=love]
 ( a=lot=like=love );
[herbie=fully=loaded]
 ( herbie=fully=loaded )
 ( herbie );
[mondovino]
 ( mondovino );
[the=interpreter]
 ( the=interpreter );
[bewitched]
```

```
 ( bewitched );
[george=a.=romeros=land=of=the=dead]
 ( george=a.=romeros=land=of=the=dead )
 ( land=of=the=dead );
[spanglish]
 ( spanglish );
[the=forgotten]
 ( the=forgotten );
[scooby-doo=2=monsters=unleashed]
 ( scooby-doo=2=monsters=unleashed )
 ( scooby=doo=2 );
[the=aviator]
 ( the=aviator );
[bride=and=prejudice]
 ( bride=and=prejudice );
[oceans=twelve]
 ( oceans=twelve );
[shrek=2]
 ( shrek=2 );
[without=a=paddle]
 ( without=a=paddle );
[the=prince=and=me]
 ( the=prince=and=me );
[the=phantom=of=the=opera]
 ( the=phantom=of=the=opera )
 ( phantom=of=the=opera );
[harry=potter=and=the=prisoner=of=azkaban]
 ( harry=potter=and=the=prisoner=of=azkaban )
 ( harry=potter=3 )
 ( harry=potter );
[the=terminal]
 ( the=terminal );
[finding=neverland]
 ( finding=neverland );
[a=cinderella=story]
 ( a=cinderella=story )
 ( cinderella=story );
[van=helsing]
 ( van=helsing );
[the=dust=factory]
 ( the=dust=factory )
 ( dust=factory );
[napoleon=dynamite]
 ( napoleon=dynamite );
[the=spongebob=squarepants=movie]
 ( the=spongebob=squarepants=movie )
 ( spongebob=squarepants )
 ( spongebob=squarepants=movie );
[save=the=green=planet]
 ( save=the=green=planet );
[house=of=wax]
 ( house=of=wax )
 ( the=house=of=wax );
[paheli]
 ( paheli );
[the=only=son]
```

```
 ( the=only=son )
 ( only=son );
[war=of=the=worlds]
 ( war=of=the=worlds )
 ( the=war=of=the=worlds );
[rebound]
 ( rebound );
[the=hitchhikers=guide=to=the=galaxy]
 ( the=hitchhikers=guide=to=the=galaxy )
 ( hitchhikers=guide=to=the=galaxy )
 ( the=hitchhikers=guide )
 ( hitchhikers=guide );
[hostage]
 ( hostage );
[sin=city]
 ( sin=city );
[million=dollar=baby]
 ( million=dollar=baby );
[high=tension]
 ( high=tension );
[parineeta]
 ( parineeta );
[walk=on=water]
 ( walk=on=water );
[schizo]
 ( schizo );
[3-iron]
 ( 3-iron )
 ( bin=jip );
[kings=and=queen]
 ( kings=and=queen );
```

Appendix B.

Expanded grammar from User Study I.

```
## -------------------- valid utterance ------------------------
[Utt]
 ( +[PHRASES] )
;
[PHRASES]
 ( [Q=DATE] )     ## i.e., query the date slot
 ( [Q=AREA] )
 ( [Q=ADDRESS] )
 ( [Q=RATING] )
 ( [Q=GENRE] )
 ( [Q=PHONE] )
 ( [Q=THEATER] )
 ( [Q=SHOWTIME] )
 ( [Q=MOVIE] )
 ( [S=DATE] )     ## i.e., specify a date constraint
 ( [S=AREA] )
 ( [S=RATING] )
 ( [S=GENRE] )
 ( [S=THEATER] )
 ( [S=SHOWTIME] )
 ( [S=MOVIE] )
;
## -------------------- slots ---------------
[Q=ADDRESS]
 ( *WHAT ADDRESS )
ADDRESS
 ( address )
 ( addresses )
WHAT
 ( what *IS-ARE )
 ( what's=the )
 ( requesting )
IS-ARE
 ( is *the )
 ( are *the )
;
[Q=DATE]
 ( *WHAT DATE *IS-ARE)
DATE
 ( date )
 ( dates )
 ( day )
 ( days )
WHAT
 ( what *IS-ARE )
 ( which *IS-ARE )
 ( what's=the )
 ( requesting )
```

```
IS-ARE
 ( is *the )
 ( are *the )
;
[Q=GENRE]
 ( *WHAT GENRE *IS-ARE )
GENRE
 ( genre )
 ( genres )
 ( movie=types )
WHAT
 ( what *IS-ARE )
 ( which *IS-ARE )
 ( what's=the )
 ( requesting )
IS-ARE
 ( is *the )
 ( are *the )
;
[Q=AREA]
 ( *WHAT AREA *IS-ARE )
AREA
 ( area )
 ( areas )
 ( city )
 ( cities )
 ( location )
 ( locations )
 ( neighborhood )
 ( neighborhoods )
WHAT
 ( what *IS-ARE )
 ( which *IS-ARE )
 ( what's=the )
 ( requesting )
IS-ARE
 ( is *the )
 ( are *the )
;
[Q=MOVIE]
 ( *WHAT PLAY-SHOW)
 ( MOVIE *name )
 ( *movie listings *for )
 ( *WHAT *the *NAME-TITLE MOVIE *PLAY-SHOW *there )
 ( WHAT MOVIE SUFFIX )
 ( WHAT MOVIE IS-ARE *available )
 ( WHAT MOVIE IS-ARE *there )
MOVIE
 ( movie )
 ( movies )
 ( film )
 ( films )
 ( title )
 ( titles )
PLAY-SHOW
 ( *IS-ARE playing )
```

```
 ( *IS-ARE showing )
 ( *IS-ARE being=played )
 ( *IS-ARE being=shown )
NAME-TITLE
 ( name=of *the )
 ( names=of *the )
 ( title=of *the )
 ( titles=of *the )
WHAT
 ( what *IS-ARE )
 ( which *IS-ARE )
 ( what's=the )
 ( requesting )
 ( what's )
IS-ARE
 ( is *the )
 ( are *the )
SUFFIX
 ( can=i=see )
 ( could=i=see )
;
[Q=PHONE]
 ( *WHAT PHONE )
PHONE
 ( *phone number )
 ( phone=numbers )
WHAT
 ( what *IS-ARE )
 ( what's=the )
 ( requesting )
IS-ARE
 ( is *the )
 ( are *the )
;
[Q=RATING]
 ( *WHAT RATING *IS-ARE )
RATING
 ( rating )
 ( ratings )
WHAT
 ( what *IS-ARE )
 ( which *IS-ARE )
 ( what's=the )
 ( requesting )
IS-ARE
 ( is *the )
 ( are *the )
;
[Q=THEATER]
 ( *FIND THEATER )
 ( *FIND WHAT THEATER *that *IS-ARE *PLAY-SHOW )
 ( THEATER IS-ARE *there )
 ( THEATER PLAY-SHOW )
 ( THEATER *where )
 ( WHAT the names of THEATER )
 ( WHAT THEATER IS-ARE it PLAY-SHOW at )
```

```
 ( where=is=it PLAY-SHOW )
 ( where=is=that PLAY-SHOW )
THEATER
 ( *the *movie theater )
 ( *the *movie theaters )
FIND
 ( find )
 ( list )
 ( name )
PLAY-SHOW
 ( playing *movies )
 ( showing *movies )
WHAT
 ( what *IS-ARE )
 ( which *IS-ARE )
 ( what's=the )
 ( requesting )
IS-ARE
 ( is *the )
 ( are *the )
;
[Q=SHOWTIME]
 ( *list SHOWTIME )
 ( SHOWTIME there )
 ( WHAT SHOWTIME *IS-ARE *PLAY-SHOW )
 ( when IS-ARE *PLAY-SHOW )
SHOWTIME
 ( show=time )
 ( show=times )
 ( start=time )
 ( *movie time )
 ( *movie times )
 ( showings )
 ( *movie timings )
PLAY-SHOW
 ( it=playing )
 ( it=showing )

WHAT
 ( what *IS-ARE )
 ( which *IS-ARE )
 ( what's=the )
 ( requesting )
 ( when *IS-ARE )
IS-ARE
 ( is *the )
 ( are *the )
;
[S=DATE]
 ( *the DATE *KNOW [DATE=VALUE] )
 ( *the DATE is [DATE=VALUE] )
 ( [DATE=VALUE] )
DATE
 ( day )
 ( days )
 ( date )
```

```
 ( dates )
KNOW
 ( *that i=want=to=know=about *is )
;
[S=GENRE]
 ( *the GENRE *KNOW [GENRE=VALUE] )
 ( *the GENRE is [GENRE=VALUE] )
 ( *a [GENRE=VALUE] *MOVIE *PLAY-SHOW )
 ( *WHAT [GENRE=VALUE] )
GENRE
 ( genre )
 ( genres )
 ( movie=type )
MOVIE
 ( movie )
 ( movies )
PLAY-SHOW
 ( *IS-ARE playing )
 ( *IS-ARE showing )
 ( *IS-ARE being=played )
 ( *IS-ARE being=shown )
WHAT
 ( what *IS-ARE )
 ( which *IS-ARE )
IS-ARE
 ( is *the )
 ( are *the )
;
KNOW
 ( *that i=want=to=know=about *is )
;
[S=AREA]
 ( *the AREA *KNOW [AREA=VALUE] )
 ( *the AREA is [AREA=VALUE] )
 ( *IN [AREA=VALUE] )
IN
 ( *located in )
 ( at )
AREA
 ( location )
 ( area )
 ( city )
 ( neighborhood )
 ( locations )
 ( areas )
 ( cities )
 ( neighborhoods )
KNOW
 ( *that i=want=to=know=about *is )
;
[MOVIE=VALUE]
## same movie values here as in Appendix A grammar
[S=MOVIE]
 ( *the MOVIE *KNOW [MOVIE=VALUE] *PLAY-SHOW )
 ( *the MOVIE is [MOVIE=VALUE] )
 ( *for [MOVIE=VALUE] *PLAY-SHOW )
```

```
 ( of [MOVIE=VALUE] )
 ( PLAY-SHOW *MOVIE [MOVIE=VALUE] )
 ( that [MOVIE=VALUE] )
 ( [MOVIE=VALUE] MOVIE )
MOVIE
 ( movie )
 ( title )
 ( film )
 ( movies )
 ( titles )
 ( films )
PLAY-SHOW
 ( *is playing *at )
 ( *is showing *at )
 ( at )
KNOW
 ( *that i=want=to=know=about *is )
;
[S=RATING]
 ( *the RATING *KNOW [RATING=VALUE] )
 ( *the RATING is [RATING=VALUE] )
 ( [RATING=VALUE] )
RATING
 ( rating )
 ( ratings )
KNOW
 ( *that i=want=to=know=about *is )
;
[S=SHOWTIME]
 ( *the SHOWTIME *KNOW [SHOWTIME=VALUE] )
 ( *the SHOWTIME is [SHOWTIME=VALUE] )
 ( [SHOWTIME=VALUE] *SHOWTIME )
SHOWTIME
 ( show=time )
 ( show=times )
 ( start=time )
 ( *movie time )
 ( *movie times )
 ( showings )
 ( movie=timings )
KNOW
 ( *that i=want=to=know=about *is )
;
[S=THEATER]
 ( THEATER *KNOW [THEATER=VALUE] )
 ( THEATER is [THEATER=VALUE] )
 ( *AT [THEATER=VALUE] )
AT
 ( at )
 ( of )
 ( for )
THEATER
 ( *the *movie theater )
 ( *the *movie theaters )
KNOW
 ( *that i=want=to=know=about *is )
```

```
;
## -------------------- values ---------------
[DATE=VALUE]
 ( [Date=Constraint] )
;
[Date=Constraint]
 ( INTERVAL )
 ( SEMI-INTERVAL [Date] )
 ( [Date] )
INTERVAL
 ( between [LoBoundDate] and [HiBoundDate] )
 ( after [LoBoundDate] before [HiBoundDate] )
SEMI-INTERVAL
 ( [Is=Before=Date] )
 ( [Is=After=Date] )
;
[LoBoundDate]
 ( [Date] )
;
[HiBoundDate]
 ( [Date] )
;
[Is=Before=Date]
 ( before )
 ( earlier=than )
;
[Is=After=Date]
 ( after )
 ( later=than )
;
[Date]
 ( [Relative=Date] )
 ( [Calendar=Date] )
;
[Relative=Date]
 ( [Weekday] )
 ( [rel=date=mod] [Weekday] )
 ( [rel=date] )
;
[rel=date=mod]
 ( last )
 ( next )
 ( this )
;
[Weekday]
 ( sunday )
 ( monday )
 ( tuesday )
 ( wednesday )
 ( thursday )
 ( friday )
 ( saturday )
;
[rel=date]
 ( yesterday )
 ( today )
```

```
 ( tomorrow )
;
[Calendar=Date]
 ( [month] [ordinal] )
;
[month]
 ( january )
 ( february )
 ( march )
 ( april )
 ( may )
 ( june )
 ( july )
 ( august )
 ( september )
 ( october )
 ( november )
 ( december )
;
[ordinal]
 ( first )
 ( twenty=first )
 ( thirty=first )
 ( second )
 ( twenty=second )
 ( third )
 ( twenty=third )
 ( fourth )
 ( twenty=fourth )
 ( fifth )
 ( twenty=fifth )
 ( sixth )
 ( twenty=sixth )
 ( seventh )
 ( twenty=seventh )
 ( eighth )
 ( twenty=eighth )
 ( ninth )
 ( twenty=ninth )
 ( tenth )
 ( eleventh )
 ( twelfth )
 ( thirteenth )
 ( fourteenth )
 ( fifteenth )
 ( sixteenth )
 ( seventeenth )
 ( eighteenth )
 ( nineteenth )
 ( twentieth )
 ( thirtieth )
;
[GENRE=VALUE]
 ( action )
 ( adventure )
 ( animation )
```

```
 ( [comedy] )
 ( crime )
 ( [documentary] )
 ( [drama] )
 ( family )
 ( fantasy )
 ( film-noir )
 ( foreign )
 ( horror )
 ( music )
 ( musical )
 ( [mystery] )
 ( romance )
 ( sci-fi )
 ( short )
 ( [sport] )
 ( [thriller] )
 ( war )
 ( [western] )
;
[comedy]
 ( comedy )
 ( comedies )
;
[documentary]
 ( documentary )
 ( documentaries )
;
[drama]
 ( drama )
 ( dramas )
;
[mystery]
 ( mystery )
 ( mysteries )
;
[sport]
 ( sport )
 ( sports )
;
[thriller]
 ( thriller )
 ( thrillers )
;
[western]
 ( western )
 ( westerns )
;
[AREA=VALUE]
 ( aspinwall )
 ( bellevue )
 ( bridgeville )
 ( century=3 )
 ( cheswick )
 ( cranberry=township )
 ( dormont )
```

```
 ( downtown )
 ( east )
 ( homestead )
 ( irwin )
 ( monroeville )
 ( mount=lebanon )
 ( near=c=m=u )
 ( north )
 ( north=hills )
 ( north=side )
 ( north=versailles )
 ( oakland )
 ( oakmont )
 ( penn=hills )
 ( pittsburgh )
 ( pleasant=hills )
 ( regent=square )
 ( robinson )
 ( south )
 ( south=side )
 ( squirrel=hill )
 ( west )
 ( west=mifflin )
 ( edgewood )
 ( moon=township )
;
[RATING=VALUE]
 ( g )
 ( p=g )
 ( p=g=thirteen )
 ( r )
 ( n=c=seventeen )
 ( not=rated )
;
[THEATER=VALUE]
 ( *the PTHEATER *THEATER )
THEATER
 ( *movie theater )
 ( cinema )
 ( cinemas )
 ( screens )
PTHEATER
 ( [Carmike=10=-=Pittsburgh] )
 ( [Carmike=Cranberry=8] )
 ( [Carmike=Galleria=6] )
 ( [Carmike=Maxi=Saver=12] )
 ( [Carmike=Southland=9] )
 ( [Cheswick=Theatres] )
 ( [Cinema=4] )
 ( [Cinemagic=Bellevue=Theater] )
 ( [Cinemagic=Denis=4=Theatres] )
 ( [Cinemagic=Manor=Theatre] )
 ( [Cinemagic=Squirrel=Hill] )
 ( [Dependable=Drive-In] )
 ( [Destinta=Theatres=-=Chartiers=Valley=20] )
 ( [Destinta=Theatres=-=Plaza=East=22] )
```

```
 ( [Flagstaff=Hill] )
 ( [Harris=Theatre] )
 ( [Loews=Waterfront=Theatre] )
 ( [Melwood=Screening=Room] )
 ( [Northway=Mall=Cinemas=8] )
 ( [Norwin=Hills=Cinemas] )
 ( [Oaks=Cinema] )
 ( [Omnimax=Theatre=-=Carnegie=Science=Center] )
 ( [Penn=Hills=Cinema] )
 ( [Regent=Square=Theatre] )
 ( [Showcase=Cinemas=Pittsburgh=North] )
 ( [Showcase=Cinemas=Pittsburgh=West] )
 ( [Southside=Works=Cinema] )
 ( [Star=City=Cinemas=-=S.=Fayette=14] )
 ( [University=Center] )
 ( [Waterworks=Cinemas] )
;
[Carmike=10=-=Pittsburgh]
 ( carmike=ten )
 ( carmike=village=ten )
 ( carmike=village=ten=pittsburgh )
 ( carmike=ten=pittsburgh )
;
[Carmike=Cranberry=8]
 ( carmike=eight )
 ( carmike=cranberry=eight )
 ( cranberry=eight )
 ( carmike=cranberry )
;
[Carmike=Galleria=6]
 ( carmike=galleria=six )
 ( galleria )
 ( galleria=six )
 ( carmike=galleria )
;
[Carmike=Maxi=Saver=12]
 ( carmike=maxi=saver )
 ( maxi=saver=twelve )
 ( carmike=maxi=saver=twelve )
 ( maxi=saver )
;
[Carmike=Southland=9]
 ( southland )
 ( carmike=southland )
 ( southland=nine )
 ( carmike=southland=nine )
;
[Cheswick=Theatres]
 ( cheswick )
 ( cheswick=quad )
 ( cheswick=quads )
;
[Cinema=4]
 ( cinema=four )
;
[Cinemagic=Bellevue=Theater]
```

```
 ( bellevue )
 ( cinemagic=bellevue )
;
[Cinemagic=Denis=4=Theatres]
 ( denis )
 ( cinemagic=denis )
 ( cinemagic=denis=four )
 ( denis=four )
;
[Cinemagic=Manor=Theatre]
 ( manor )
 ( cinemagic=manor )
;
[Cinemagic=Squirrel=Hill]
 ( cinemagic=squirrel=hill )
 ( squirrel=hill )
;
[Destinta=Theatres=-=Chartiers=Valley=20]
 ( destinta=bridgeville )
 ( destinta=theatres=chartiers=valley )
 ( destinta=chartiers=twenty )
 ( destinta=chartiers=valley=twenty )
 ( destinta=theatres=chartiers=valley=twenty )
 ( destinta=theatres=chartiers )
 ( destinta=theatres=chartiers=twenty )
 ( destinta=chartiers=valley )
 ( destinta=chartiers )
 ( chartiers=twenty )
 ( chartiers=valley=twenty )
 ( destinta=twenty )
;
[Destinta=Theatres=-=Plaza=East=22]
 ( destinta=north=versailles )
 ( destinta=plaza=east TWENTYTWO )
 ( destinta=plaza=east )
 ( destinta=theatres=plaza=east )
 ( destinta=theatres=plaza=east TWENTYTWO )
 ( plaza=east TWENTYTWO )
 ( plaza=east )
 ( destinta TWENTYTWO )
TWENTYTWO
 ( twenty=two )
;
[Dependable=Drive-In]
 ( dependable=drive=in )
 ( dependable )
;
[Flagstaff=Hill]
 ( flagstaff=hill )
 ( flagstaff )
 ( schenley )
 ( schenley=park )
;
[Harris=Theatre]
 ( harris )
 ( filmmakers=at=the=harris )
```

```
;
[Loews=Waterfront=Theatre]
 ( loews )
 ( loews=waterfront )
 ( waterfront )
 ( homestead=waterfront )
;
[Melwood=Screening=Room]
 ( filmmakers )
 ( melwood=screening=room )
 ( melwood )
;
[Northway=Mall=Cinemas=8]
 ( northway=mall=cinemas=eight )
 ( northway=mall=eight )
 ( northway=mall )
 ( northway=eight )
 ( northway )
;
[Norwin=Hills=Cinemas]
 ( norwin=hills )
;
[Oaks=Cinema]
 ( oaks )
;
[Omnimax=Theatre=-=Carnegie=Science=Center]
 ( CSC )
 ( omnimax=theatre )
 ( omnimax )
 ( omnimax=theatre CSC )
 ( omnimax CSC )
CSC
 ( carnegie=science=center )
 ( science=center )
;
[Penn=Hills=Cinema]
 ( penn=hills )
;
[Regent=Square=Theatre]
 ( regent=square )
;
[Showcase=Cinemas=Pittsburgh=North]
 ( showcase=cinemas=north )
 ( showcase=pittsburgh=north )
 ( showcase=north )
 ( showcase=cinemas=pittsburgh=north )
;
[Showcase=Cinemas=Pittsburgh=West]
 ( showcase=west )
 ( showcase=cinemas=pittsburgh=west )
 ( showcase=pittsburgh=west )
 ( showcase=cinemas=west )
;
[Southside=Works=Cinema]
 ( southside=works )
 ( southside )
```

```
;
[Star=City=Cinemas=-=S.=Fayette=14]
 ( star=city=cinema=south=fayette )
 ( star=city=fourteen )
 ( star=city )
 ( star=city=fayette=fourteen )
 ( star=city=cinema=fayette )
 ( star=city=cinema=fayette=fourteen )
 ( star=city=south=fourteen )
 ( star=city=south=fayette=fourteen )
 ( star=city=cinema=south=fayette=fourteen )
 ( star=city=cinema=south )
 ( star=city=cinema=fourteen )
 ( star=city=south=fayette )
 ( star=city=fayette )
 ( star=city=cinema )
 ( star=city=south )
 ( star=city=cinema=south=fourteen )
 ( south=fayette=fourteen )
 ( fayette=fourteen )
;
[University=Center]
 ( university=center )
 ( mcconomy )
 ( carnegie=mellon=university=center )
 ( u=c )
 ( c.=m.=u.=university=center )
 ( c.=m.=u. )
;
[Waterworks=Cinemas]
 ( waterworks )
;
[SHOWTIME=VALUE]
 ( [Time=Constraint] )
;
[Time=Constraint]
 ( INTERVAL )
 ( SEMI-INTERVAL [Time] )
 ( *at [Time] )
INTERVAL
 ( between [LoBoundTime] and [HiBoundTime] )
 ( after [LoBoundTime] before [HiBoundTime] )
SEMI-INTERVAL
 ( [Is=Before=Time] )
 ( [Is=After=Time] )
;
[LoBoundTime]
 ( [Time] )
;
[HiBoundTime]
 ( [Time] )
;
[Is=Before=Time]
 ( before )
 ( earlier=than )
;
```

```
[Is=After=Time]
 ( after )
 ( past )
 ( later=than )
;
[Time]
 ( TIME )
TIME
 ( [Hour] o'clock AM-PM )
 ( [Hour] o'clock )
 ( [Hour] )
 ( [Hour] AM-PM )
 ( [Hour] [Minute] )
 ( [Hour] [Minute] AM-PM )
 ( noon )
 ( midnight )
AM-PM
 ( [AM] )
 ( [PM] )
;
[AM]
 ( a=m )
;
[PM]
 ( p=m )
;
[Hour]
 ( one )
 ( two )
 ( three )
 ( four )
 ( five )
 ( six )
 ( seven )
 ( eight )
 ( nine )
 ( ten )
 ( eleven )
 ( twelve )
;
[Minute]
 ( oh=five )
 ( ten )
 ( fifteen )
 ( twenty )
 ( twenty=five )
 ( thirty=five )
 ( thirty )
 ( forty=five )
 ( forty )
 ( fifty )
 ( fifty=five )
;
```

Appendix C.

Representative tasks from User Study I. Task difficulty levels are in parentheses after each item.

1. You want to see Mad Hot Ballroom. Find out where it's showing. (1)
2. You live in Aspinwall, close to the Waterworks theater. Find out what's playing there. (1)
3. You found a movie you want to see at the Cheswick theater, but you're not sure where it is. Find out the theater's phone number so you can call them later to ask for directions. (1)
4. You want to see Fantastic Four at the Norwin Hills theater. Find out when it's showing there. (2)
5. You just finished shopping in Pittsburgh, near the Squirrel Hill theater, and you're in the mood to see a sci-fi movie. Find out which ones are playing there. (2)
6. You have an appointment south of the city and you want to see a movie afterwards. You want to see Madagascar. Find out where it's showing in that area. (2)
7. You're going to see a movie at the Dependable Drive-In, and you'd like to see a horror movie. Are any showing there? (2)
8. You really want to go see Wedding Crashers. It's playing at the Waterfront, but you're busy most of the day. When's the latest it's showing? (2)
9. You're going to be in the North Hills and you think you might want to see a movie. You know there must be some theaters around there, but you don't know which ones they are or where they're located. Find out this information. (3)

10. You live near the Southside Works theater and you want to see a crime movie. Found out which ones are playing there and when. (4)
11. You want to see Herbie: Fully Loaded. Is it showing at the Destinta Chartiers 20 theater? (1)
12. You want to see Land of the Dead. Find out where it's showing. (1)
13. You live in Robinson, close to the Showcase West Theater. Find out what's playing there. (1)
14. You want to see the Longest Yard at the Plaza East 22 theater. Find out when it's showing there. (2)
15. You've just finished shopping in the North Hills near the Showcase North theater, and you're in the mood to see a drama. Find out which ones are playing there. (2)

Appendix D.

Tasks from session five of User Study III, on the DineLine system. Task difficulty levels are in parentheses after each item.

1. You want to go to the Sonoma Grille. Find out its address. (1)
2. You want to go to the Church Brew Works. Find out its phone number. (1)
3. You're on the South Side and you're hungry for "American" food. Find out which of those restaurants are there. (2)
4. You're thinking about trying out Lidia's Pittsburgh. Find out how the system rates this restaurant and how expensive it is. (3)
5. Find out what kind of food they serve at Max's Allegheny Tavern. (1)
6. You're downtown, and you're looking for a moderately-priced place to eat. Find out where you could go. (2)

Appendix E.

Speech Graffiti DineLine grammar from User Study III.

```
## -------------------- valid utterance ------------------------
[Utt]
 ( [PHRASES] )
 ( [PHRASES] +[PHRASES] )
 ( [KeyPhrase] )
 ( [NavPhrase] )
;
[PHRASES]
 ( [DAY=SLOT] [DAY=VALUE] )
 ( [CUISINE=SLOT] [CUISINE=VALUE] )
 ( [RATING=SLOT] [RATING=VALUE] )
 ( [AREA=SLOT] [AREA=VALUE] )
 ( [MEAL=SLOT] [MEAL=VALUE] )
 ( [REST=SLOT] [REST=VALUE] )
 ( [PRICE=SLOT] [PRICE=VALUE] )
 ( [WHAT] SLOTS )
 ( SLOTS [IS=ANYTHING] )
 ( *SLOTS [OPTIONS] )
 ( ERASER )
VALUES
 ( [DAY=VALUE] )
 ( [CUISINE=VALUE] )
 ( [RATING=VALUE] )
 ( [AREA=VALUE] )
 ( [MEAL=VALUE] )
 ( [REST=VALUE] )
 ( [PRICE=VALUE] )
SLOTS
 ( [DAY=SLOT] )
 ( [AREA=SLOT] )
 ( [ADDRESS=SLOT] )
 ( [RATING=SLOT] )
 ( [CUISINE=SLOT] )
 ( [PHONE=SLOT] )
 ( [MEAL=SLOT] )
 ( [PRICE=SLOT] )
 ( [REST=SLOT] )
ERASER
 ( [ClearContext] )
 ( [ClearUtterance] )
;
## -------------------- query format -----------------------
[WHAT]
 ( list )
;

## -------------------- keywords -----------------------
[NavPhrase]
 ( [More] )
 ( [Previous] [Num] )
```

```
 ( [Previous] )
 ( [Next] [Num] )
 ( [Next] )
 ( [First] [Num] )
 ( [First] )
 ( [Last] [Num] )
 ( [Last] )
;
[More]
 ( more )
;
[Previous]
 ( previous )
;
[Next]
 ( next )
;
[First]
 ( first )
;
[Last]
 ( last )
;
[ClearContext]
 ( start=over )
 ( starting=over )
;
[ClearUtterance]
 ( scratch=that )
;
[KeyPhrase]
 ( [Restate] )
 ( [Repeat] )
 ( [Help] )
 ( [Intro] )
;
[Help]
 ( help )
;
[Restate]
 ( where=was=i )
 ( where=were=we )
 ( where=am=i )
 ( where=are=we )
;
[Repeat]
 ( repeat )
;
[IS=ANYTHING]
 ( anything )
;
[OPTIONS]
 ( options )
;
[Intro]
 ( introduction )
```

```
 ( tutorial )
;
[Num]
 ( one )
 ( two )
 ( three )
 ( four )
 ( five )
 ( six )
 ( seven )
 ( eight )
 ( nine )
 ( ten )
 ( eleven )
 ( twelve )
;
## ------------------- slots ---------------
[ADDRESS=SLOT]
 ( address )
 ( address=is )
 ( addresses )
 ( the=address )
 ( the=address=is )
;
[DAY=SLOT]
 ( day )
 ( day=is )
 ( days )
 ( the=day )
 ( the=day=is )
 ( the=days )
 ( day=of=the=week )
 ( day=of=the=week=is )
 ( days=of=the=week )
 ( the=day=of=the=week )
 ( the=day=of=the=week=is )
 ( the=days=of=the=week )
;
[CUISINE=SLOT]
 ( cuisine )
 ( cuisine=is )
 ( cuisines )
 ( the=cuisine )
 ( the=cuisine=is )
 ( the=cuisines )
 ( type )
 ( type=is )
 ( types )
 ( the=type )
 ( the=type=is )
 ( the=types )
;
[AREA=SLOT]
 ( area )
 ( area=is )
 ( areas )
```

```
 ( neighborhood )
 ( neighborhood=is )
 ( neighborhoods )
 ( the=area )
 ( the=area=is )
 ( the=areas )
 ( the=neighborhood )
 ( the=neighborhood=is )
 ( the=neighborhoods )
;
[PHONE=SLOT]
 ( phone=number )
 ( phone=number=is )
 ( phone=numbers )
 ( the=phone=number )
 ( the=phone=numbers )
 ( telephone=number )
 ( the=telephone=number )
;
[RATING=SLOT]
 ( rating )
 ( rating=is )
 ( ratings )
 ( the=rating )
 ( the=ratings )
 ( star=rating )
 ( star=rating=is )
 ( star=ratings )
 ( the=star=rating )
 ( the=star=ratings )
;
[MEAL=SLOT]
 ( the=meal )
 ( the=meal=is )
 ( the=meals )
 ( meal )
 ( meal=is )
 ( meals )
;
[PRICE=SLOT]
 ( price )
 ( price=is )
 ( prices )
 ( prices=are )
 ( the=price )
 ( the=price=is )
 ( the=prices )
 ( the=prices=are )
 ( the=price=range )
 ( the=price=range=is )
 ( the=price=ranges )
 ( the=price=ranges=are )
 ( price=range )
 ( price=range=is )
 ( price=ranges )
 ( price=ranges=are )
```

```
;
[REST=SLOT]
 ( restaurant )
 ( restaurant=is )
 ( restaurants )
 ( restaurants=are )
 ( name )
 ( name=is )
 ( names )
 ( names=are )
 ( the=name )
 ( the=name=is  )
 ( the=names )
 ( the=names=are )
 ( the=restaurant )
 ( the=restaurant=is )
 ( the=restaurants )
 ( the=restaurants=are )
 ( restaurant=name )
 ( restaurant=name=is )
 ( restaurant=names )
 ( the=restaurant=name )
 ( the=restaurant=name=is )
 ( the=restaurant=names )
;
## ------------------- values ---------------
[DAY=VALUE]
 ( sunday )
 ( monday )
 ( tuesday )
 ( wednesday )
 ( thursday )
 ( friday )
 ( saturday )
;
[CUISINE=VALUE]
 ( african )
 ( american )
 ( asian )
 ( bakery )
 ( barbecue )
 ( belgian )
 ( cambodian )
 ( caribbean )
 ( chinese )
 ( coffee=house )
 ( contemporary )
 ( continental )
 ( deli )
 ( desserts )
 ( diner )
 ( eastern=european )
 ( eclectic )
 ( ethiopian )
 ( european )
 ( french )
```

```
 ( german )
 ( greek )
 ( indian )
 ( irish )
 ( italian )
 ( japanese )
 ( mediterranean )
 ( mexican )
 ( middle=eastern )
 ( peruvian )
 ( pizza )
 ( portuguese )
 ( seafood )
 ( spanish )
 ( steakhouse )
 ( sushi )
 ( thai )
 ( vegetarian )
 ( vietnamese )
;

[AREA=VALUE]
 ( bloomfield )
 ( downtown )
 ( east=liberty )
 ( garfield )
 ( highland=park )
 ( homestead )
 ( lawrenceville )
 ( mount=washington )
 ( north=side )
 ( oakland )
 ( point=breeze )
 ( regent=square )
 ( shadyside )
 ( south=side )
 ( squirrel=hill )
 ( station=square )
 ( strip=district )
;
[MEAL=VALUE]
 ( breakfast )
 ( lunch )
 ( dinner )
;
[RATING=VALUE]
 ( [1] )
 ( [2] )
 ( [3] )
 ( [4] )
 ( [5] )
;
[1]
 ( one=star )
;
[2]
```

```
 ( two=stars )
;
[3]
 ( three=stars )
;
[4]
 ( four=stars )
;
[5]
 ( five=stars )
;
[PRICE=VALUE]
 ( [cheap] )
 ( [moderate] )
 ( [expensive] )
 ( [very=expensive] )
 ( [Amount=Constraint] *dollars )
;
[cheap]
 ( inexpensive )
 ( cheap )
;
[moderate]
 ( moderate )
;
[expensive]
 ( expensive )
;
[very=expensive]
 ( very=expensive )
;
[Amount=Constraint]
 ( SEMI-INTERVAL [Num-100] )
 ( *AROUND [Num-100] )
AROUND
 ( about )
 ( around )
SEMI-INTERVAL
 ( [MoreThan] )
 ( [LessThan] )
;
[Num-100]
 ( one )
 ( two )
 ( three )
 ( four )
 ( five )
 ( six )
 ( seven )
 ( eight )
 ( nine )
 ( ten )
 ( eleven )
 ( twelve )
 ( thirteen )
 ( fourteen )
```

```
 ( fifteen )
 ( sixteen )
 ( seventeen )
 ( eighteen )
 ( nineteen )
 ( twenty )
 ( twenty=five )
 ( thirty )
 ( thirty=five )
 ( forty )
 ( forty=five )
 ( fifty )
 ( fifty=five )
 ( sixty )
 ( sixty=five )
 ( seventy )
 ( seventy=five )
 ( eighty )
 ( eighty=five )
 ( ninety )
 ( ninety=five )
 ( a=hundred )
 ( one=hundred )
;
[MoreThan]
 (more=than)
 (over)
;
[LessThan]
 (less=than)
 (under)
;
[AtLeast]
 (at=least)
;
[AtMost]
 (at=most)
;
[REST=VALUE]
 ( [ABAY=ETHIOPIAN=CUISINE] )
 ( [ABRUZZIS] )
 ( [ALADDINS=EATERY] )
 ( [ALEXANDERS=PASTA=EXPRESS] )
 ( [ALI=BABA] )
 ( [ASIAGO=EURO-CUISINE] )
 ( [AUSSOME=AUSSIE=BOOMERANG=BBQ] )
 ( [BANGKOK=BALCONY] )
 ( [BRAVO=FRANCO] )
 ( [BRUSCHETTAS] )
 ( [BUCA=DI=BEPPO] )
 ( [BUFFALO=BLUES] )
 ( [CAFE=ALLEGRO] )
 ( [CAFE=ASIA] )
 ( [CAFE=DU=JOUR] )
 ( [CAFE=EURO] )
 ( [CAFE=SAM] )
```

```
( [CAFE=ZAO] )
( [CAFE=ZINHO] )
( [CAFFE=AMANTE] )
( [CAPPYS=CAFE] )
( [CASBAH] )
( [CHAYA=JAPANESE=CUISINE] )
( [CHINA=PALACE] )
( [CHRISTOS] )
( [CHURCH=BREW=WORKS] )
( [CIAO=BABY=RISTORANTE] )
( [CITY=GRILL] )
( [CLADDAGH=IRISH=PUB] )
( [COMMON=PLEA=RESTAURANT] )
( [COZUMEL=RESTAURANTE=MEXICANO] )
( [DEJAVU=LOUNGE] )
( [DELS] )
( [DELUCAS] )
( [DISH=OSTERIA=AND=BAR] )
( [DOWES=ON=9TH] )
( [EAST=END=CO-OP=CAFE] )
( [EATUNIQUE] )
( [ELBOW=ROOM] )
( [ELEVEN] )
( [ENO] )
( [ENRICOS=RISTORANTE] )
( [ENRICOS=TAZZA=DORO=CAFE=AND=ESPRESSO=BAR] )
( [GEORGETOWNE=INN] )
( [GIRASOLE] )
( [GRAND=CONCOURSE] )
( [GRANDVIEW=SALOON] )
( [GULLIFTYS] )
( [HOT=METAL=GRILLE] )
( [INDIA=GARDEN] )
( [INDICA] )
( [ISABELA=ON=GRANDVIEW] )
( [JOE=MAMAS=ITALIAN=DELUXE] )
( [JOJOS] )
( [KASSABS] )
( [KAYA] )
( [KAZANSKYS] )
( [KIKU] )
( [LA=CUCINA=FLEGREA] )
( [LA=FERIA] )
( [LA=FIESTA] )
( [LAFORET] )
( [LE=POMMIER] )
( [LEGENDS=OF=THE=NORTH=SHORE] )
( [LEMONT] )
( [LIDIAS=PITTSBURGH] )
( [LUCCA] )
( [LULUS] )
( [MAD=MEX] )
( [MALLORCA] )
( [MARIANIS=PLEASURE=BAR] )
( [MARIOS=SOUTHSIDE=SALOON-BLUE=LOUS] )
( [MARKS=GRILLE=AND=CATERING] )
```

```
( [MAXS=ALLEGHENY=TAVERN] )
( [MCCORMICK=AND=SCHMICKS=SEAFOOD=RESTAURANT] )
( [MELTING=POT] )
( [MINEOS] )
( [MITCHELLS=FISH=MARKET] )
( [MONTEREY=BAY=FISH=GROTTO] )
( [MORTONS=THE=STEAKHOUSE] )
( [MULLANEYS=HARP=AND=FIDDLE] )
( [MY=THAI] )
( [NAKAMA=JAPANESE=STEAKHOUSE=AND=SUSHI=BAR] )
( [NICOS=RECOVERY=ROOM] )
( [OLD=EUROPE] )
( [OPUS] )
( [ORIENT=KITCHEN] )
( [ORIGINAL=OYSTER=HOUSE] )
( [P.F.=CHANGS=CHINA=BISTRO] )
( [PALAZZO=RISTORANTE] )
( [PALOMINO] )
( [PAMELAS] )
( [PENN=BREWERY] )
( [PER=MIE=FIGLIA=RESTAURANT] )
( [PHNOM=PENH] )
( [PHO=MINH] )
( [PICCOLO=FORNO] )
( [PICCOLO=PICCOLO=RISTORANTE] )
( [PINOS=MERCATO] )
( [PIPERS=PUB] )
( [PITTSBURGH=RARE] )
( [PITTSBURGH=STEAK=CO] )
( [POINT=BRUGGE=CAFE] )
( [PRELUDE=WINE=BAR] )
( [PRIMANTI=BROTHERS] )
( [PRINCE=OF=INDIA] )
( [RED=ROOM=CAFE=AND=LOUNGE] )
( [RITTERS] )
( [ROLANDS=IRON=LANDING] )
( [RUTHS=CHRIS=STEAK=HOUSE] )
( [SESAME=INN] )
( [SHARP=EDGE=BAR=AND=RESTAURANT] )
( [SHILOH=INN] )
( [SITAR=OF=PITTSBURGH] )
( [SIX=PENN=KITCHEN=RESTAURANT] )
( [SMALLMAN=ST.=DELI] )
( [SOBA=LOUNGE] )
( [SONOMA=GRILLE] )
( [SPICE=ISLAND=TEA=HOUSE] )
( [SQUARE=CAFE] )
( [STAR=OF=INDIA] )
( [SUNNYLEDGE=OUTDOOR=CAFE=MARTINI=BAR] )
( [SUSHI=KIM] )
( [SUSHI=TOO] )
( [SUSHI=TWO] )
( [TAMBELLINI=RESTAURANT] )
( [TESSAROS] )
( [THAI=CUISINE=RESTAURANT] )
( [THAI=PLACE=RESTAURANT] )
```

```
 ( [THE=BRIDGE=CAFE] )
 ( [THE=CAFE=AT=THE=FRICK] )
 ( [THE=CARLTON] )
 ( [THE=CHEESECAKE=FACTORY] )
 ( [THE=ORIGINAL=FISH=MARKET] )
 ( [TIN=ANGEL] )
 ( [TONIC=BAR=AND=GRILL] )
 ( [TRAMS=KITCHEN] )
 ( [TRILOGY] )
 ( [TYPHOON] )
 ( [UMI=JAPANESE=RESTAURANT] )
 ( [UNION=GRILL] )
 ( [WALNUT=GRILL] )
 ( [ZARRAS] )
 ( [ZENITH] )
;
[ABAY=ETHIOPIAN=CUISINE]
 ( abay=ethiopian=cuisine )
 ( abay );
[ABRUZZIS]
 ( abruzzis  );
[ALADDINS=EATERY]
 ( aladdins=eatery )
 ( aladdins  );
[ALEXANDERS=PASTA=EXPRESS]
 ( alexanders=pasta=express )
 ( alexanders );
[ALI=BABA]
 ( ali=baba );
[ASIAGO=EURO-CUISINE]
 ( asiago=euro-cuisine )
 ( asiago );
[AUSSOME=AUSSIE=BOOMERANG=BBQ]
 ( aussome=aussie=boomerang=barbecue )
 ( aussome=aussie )
 ( boomerang=barbecue );
[BANGKOK=BALCONY]
 ( bangkok=balcony );
[BRAVO=FRANCO]
 ( bravo=franco );
[BRUSCHETTAS]
 ( bruschettas )
 ( bruschetta );
[BUCA=DI=BEPPO]
 ( buca=di=beppo )
 ( buca );
[BUFFALO=BLUES]
 ( buffalo=blues );
[CAFE=ALLEGRO]
 ( cafe=allegro );
[CAFE=ASIA]
 ( cafe=asia );
[CAFE=DU=JOUR]
 ( cafe=du=jour );
[CAFE=EURO]
 ( cafe=euro );
```

```
[CAFE=SAM]
 ( cafe=sam );
[CAFE=ZAO]
 ( cafe=zao );
[CAFE=ZINHO]
 ( cafe=zinho );
[CAFFE=AMANTE]
 ( caffe=amante );
[CAPPYS=CAFE]
 ( cappys=cafe )
 ( cappys );
[CASBAH]
 ( casbah );
[CHAYA=JAPANESE=CUISINE]
 ( chaya=japanese=cuisine )
 ( chaya );
[CHINA=PALACE]
 ( china=palace )
 ( the=china=palace );
[CHRISTOS]
 ( christos );
[CHURCH=BREW=WORKS]
 ( church=brew=works )
 ( the=church=brew=works );
[CIAO=BABY=RISTORANTE]
 ( ciao=baby=ristorante )
 ( ciao=baby );
[CITY=GRILL]
 ( city=grill )
 ( the=city=grill );
[CLADDAGH=IRISH=PUB]
 ( claddagh=irish=pub )
 ( claddagh )
 ( the=claddagh );
[COMMON=PLEA=RESTAURANT]
 ( common=plea=restaurant )
 ( the=common=plea=restaurant )
 ( common=plea )
 ( the=common=plea );
[COZUMEL=RESTAURANTE=MEXICANO]
 ( cozumel=restaurante=mexicano )
 ( cozumel );
[DEJAVU=LOUNGE]
 ( dejavu=lounge )
 ( the=dejavu=lounge )
 ( dejavu );
[DELS]
 ( dels );
[DELUCAS]
 ( delucas );
[DISH=OSTERIA=AND=BAR]
 ( dish=osteria=and=bar )
 ( dish );
[DOWES=ON=9TH]
 ( dowes=on=9th )
 ( dowes );
```

```
[EAST=END=CO-OP=CAFE]
 ( east=end=co-op=cafe )
 ( the=east=end=co-op=cafe )
 ( east=end=co-op )
 ( the=east=end=co-op );
[EATUNIQUE]
 ( eatunique )
 ( craig=street=coffee );
[ELBOW=ROOM]
 ( elbow=room )
 ( the=elbow=room );
[ELEVEN]
 ( eleven );
[ENO]
 ( eno );
[ENRICOS=RISTORANTE]
 ( enricos=ristorante )
 ( enricos )
 ( enricos=shadyside )
[ENRICOS=TAZZA=DORO=CAFE=AND=ESPRESSO=BAR]
 ( enricos=tazza=doro=cafe=and=espresso=bar )
 ( enricos=tazza=doro )
 ( enricos=tazza=doro=cafe )
 ( enricos=highland=park );
[GEORGETOWNE=INN]
 ( georgetowne=inn )
 ( the=georgetowne=inn );
[GIRASOLE]
 ( girasole );
[GRAND=CONCOURSE]
 ( grand=concourse )
 ( the=grand=concourse );
[GRANDVIEW=SALOON]
 ( grandview=saloon )
 ( the=grandview=saloon );
[GULLIFTYS]
 ( gulliftys );
[HOT=METAL=GRILLE]
 ( hot=metal=grille )
 ( the=hot=metal=grille );
[INDIA=GARDEN]
 ( india=garden );
[INDICA]
 ( indica );
[ISABELA=ON=GRANDVIEW]
 ( isabela=on=grandview )
 ( isabela );
[JOE=MAMAS=ITALIAN=DELUXE]
 ( joe=mamas=italian=deluxe )
 ( joe=mamas=italian )
 ( joe=mamas );
[JOJOS]
 ( jojos );
[KASSABS]
 ( kassabs )
 ( kassab );
```

```
[KAYA]
 ( kaya );
[KAZANSKYS]
 ( kazanskys )
 ( kazanskys=deli );
[KIKU]
 ( kiku );
[LA=CUCINA=FLEGREA]
 ( la=cucina=flegrea );
[LA=FERIA]
 ( la=feria );
[LA=FIESTA]
 ( la=fiesta );
[LAFORET]
 ( laforet );
[LE=POMMIER]
 ( le=pommier );
[LEGENDS=OF=THE=NORTH=SHORE]
 ( legends=of=the=north=shore )
 ( legends );
[LEMONT]
 ( lemont );
[LIDIAS=PITTSBURGH]
 ( lidias=pittsburgh )
 ( lidias );
[LUCCA]
 ( lucca );
[LULUS]
 ( lulus );
[MAD=MEX]
 ( mad=mex );
[MALLORCA]
 ( mallorca );
[MARIANIS=PLEASURE=BAR]
 ( marianis=pleasure=bar )
 ( marianis )
 ( the=pleasure=bar )
 ( pleasure=bar );
[MARIOS=SOUTHSIDE=SALOON-BLUE=LOUS]
 ( marios=southside=saloon-blue=lous )
 ( marios=southside=saloon )
 ( marios )
 ( blue=lous );
[MARKS=GRILLE=AND=CATERING]
 ( marks=grille=and=catering )
 ( marks=grille );
[MAXS=ALLEGHENY=TAVERN]
 ( maxs=allegheny=tavern )
 ( allegheny=tavern )
 ( maxs );
[MCCORMICK=AND=SCHMICKS=SEAFOOD=RESTAURANT]
 ( mccormick=and=schmicks=seafood=restaurant )
 ( mccormick=and=schmicks );
[MELTING=POT]
 ( melting=pot )
 ( the=melting=pot );
```

```
[MINEOS]
 ( mineos );
[MITCHELLS=FISH=MARKET]
 ( mitchells=fish=market )
 ( mitchells );
[MONTEREY=BAY=FISH=GROTTO]
 ( monterey=bay=fish=grotto )
 ( monterey=bay );
[MORTONS=THE=STEAKHOUSE]
 ( mortons=the=steakhouse )
 ( mortons=steakhouse )
 ( mortons );
[MULLANEYS=HARP=AND=FIDDLE]
 ( mullaneys=harp=and=fiddle )
 ( mullaneys );
[MY=THAI]
 ( my=thai );
[NAKAMA=JAPANESE=STEAKHOUSE=AND=SUSHI=BAR]
 ( nakama=japanese=steakhouse=and=sushi=bar )
 ( nakama=japanese=steakhouse )
 ( nakama );
[NICOS=RECOVERY=ROOM]
 ( nicos=recovery=room )
 ( nicos )
 ( the=recovery=room );
[OLD=EUROPE]
 ( old=europe );
[OPUS]
 ( opus );
[ORIENT=KITCHEN]
 ( orient=kitchen )
 ( the=orient=kitchen );
[ORIGINAL=OYSTER=HOUSE]
 ( original=oyster=house )
 ( the=original=oyster=house );
[P.F.=CHANGS=CHINA=BISTRO]
 ( p.f.=changs=china=bistro )
 ( p.f.=changs );
[PALAZZO=RISTORANTE]
 ( palazzo=ristorante )
 ( palazzo );
[PALOMINO]
 ( palomino );
[PAMELAS]
 ( pamelas );
[PENN=BREWERY]
 ( penn=brewery )
 ( the=penn=brewery );
[PER=MIE=FIGLIA=RESTAURANT]
 ( per=mie=figlia=restaurant )
 ( per=mie=figlia );
[PHNOM=PENH]
 ( phnom=penh );
[PHO=MINH]
 ( pho=minh );
[PICCOLO=FORNO]
```

```
 ( piccolo=forno );
[PICCOLO=PICCOLO=RISTORANTE]
 ( piccolo=piccolo=ristorante )
 ( piccolo=piccolo );
[PINOS=MERCATO]
 ( pinos=mercato )
 ( pinos );
[PIPERS=PUB]
 ( pipers=pub )
 ( pipers );
[PITTSBURGH=RARE]
 ( pittsburgh=rare );
[PITTSBURGH=STEAK=CO]
 ( pittsburgh=steak=company )
 ( the=pittsburgh=steak=company );
[POINT=BRUGGE=CAFE]
 ( point=brugge=cafe )
 ( the=point=brugge=cafe )
 ( point=brugge );
[PRELUDE=WINE=BAR]
 ( prelude=wine=bar )
 ( prelude );
[PRIMANTI=BROTHERS]
 ( primanti=brothers )
 ( primantis );
[PRINCE=OF=INDIA]
 ( prince=of=india )
 ( the=prince=of=india );
[RED=ROOM=CAFE=AND=LOUNGE]
 ( red=room=cafe=and=lounge )
 ( red=room=cafe )
 ( red=room )
 ( the=red=room=cafe=and=lounge )
 ( the=red=room=cafe )
 ( the=red=room );
[RITTERS]
 ( ritters )
 ( ritters=diner );
[ROLANDS=IRON=LANDING]
 ( rolands=iron=landing )
 ( rolands )
 ( the=iron=landing );
[RUTHS=CHRIS=STEAK=HOUSE]
 ( ruths=chris=steak=house )
 ( ruths=chris );
[SESAME=INN]
 ( sesame=inn )
 ( the=sesame=inn );
[SHARP=EDGE=BAR=AND=RESTAURANT]
 ( sharp=edge=bar=and=restaurant )
 ( sharp=edge=restaurant )
 ( sharp=edge )
 ( the=sharp=edge=bar=and=restaurant )
 ( the=sharp=edge=restaurant )
 ( the=sharp=edge );
[SHILOH=INN]
```

```
 ( shiloh=inn )
 ( the=shiloh=inn );
[SITAR=OF=PITTSBURGH]
 ( sitar=of=pittsburgh )
 ( sitar )
 ( the=sitar );
[SIX=PENN=KITCHEN=RESTAURANT]
 ( six=penn=kitchen=restaurant )
 ( six=penn=kitchen )
 ( six=penn );
[SMALLMAN=ST.=DELI]
 ( smallman=street=deli );
[SOBA=LOUNGE]
 ( soba=lounge )
 ( the=soba=lounge )
 ( soba );
[SONOMA=GRILLE]
 ( sonoma=grille )
 ( the=sonoma=grille );
[SPICE=ISLAND=TEA=HOUSE]
 ( spice=island=tea=house )
 ( spice=island )
 ( the=spice=island=tea=house );
[SQUARE=CAFE]
 ( square=cafe )
 ( the=square=cafe );
[STAR=OF=INDIA]
 ( star=of=india )
 ( the=star=of=india );
[SUNNYLEDGE=OUTDOOR=CAFE=MARTINI=BAR]
 ( sunnyledge=outdoor=cafe=martini=bar )
 ( sunnyledge=outdoor=cafe )
 ( sunnyledge=cafe )
 ( sunnyledge )
 ( sunnyledge=cafe=and=martini=bar )
 ( the=sunnyledge=outdoor=cafe=martini=bar )
 ( the=sunnyledge=outdoor=cafe )
 ( the=sunnyledge=cafe )
 ( the=sunnyledge )
 ( the=sunnyledge=cafe=and=martini=bar );
[SUSHI=KIM]
 ( sushi=kim );
[SUSHI=TOO]
 ( sushi=too )
 ( sushi=too=shadyside )
 ( sushi=too=on=walnut )
 ( sushi=too=walnut=street );
[SUSHI=TWO]
 ( sushi=two )
 ( sushi=two=south=side )
 ( sushi=two=on=carson )
 ( sushi=two=carson=street );
[TAMBELLINI=RESTAURANT]
 ( tambellini=restaurant )
 ( tambellini )
 ( tambellinis );
```

```
[TESSAROS]
 ( tessaros );
[THAI=CUISINE=RESTAURANT]
 ( thai=cuisine=restaurant );
[THAI=PLACE=RESTAURANT]
 ( thai=place=restaurant )
 ( thai=place );
[THE=BRIDGE=CAFE]
 ( the=bridge=cafe )
 ( bridge=cafe );
[THE=CAFE=AT=THE=FRICK]
 ( the=cafe=at=the=frick )
 ( cafe=at=the=frick )
 ( the=frick=museum=cafe );
[THE=CARLTON]
 ( the=carlton );
[THE=CHEESECAKE=FACTORY]
 ( the=cheesecake=factory )
 ( cheesecake=factory );
[THE=ORIGINAL=FISH=MARKET]
 ( the=original=fish=market )
  ( original=fish=market );
[TIN=ANGEL]
 ( tin=angel )
 ( the=tin=angel );
[TONIC=BAR=AND=GRILL]
 ( tonic=bar=and=grill )
 ( tonic )
 ( tonic=grill );
[TRAMS=KITCHEN]
 ( trams=kitchen )
 ( trams );
[TRILOGY]
 ( trilogy );
[TYPHOON]
 ( typhoon );
[UMI=JAPANESE=RESTAURANT]
 ( umi=japanese=restaurant )
 ( umi );
[UNION=GRILL]
 ( union=grill )
 ( the=union=grill );
[WALNUT=GRILL]
 ( walnut=grill )
 ( the=walnut=grill );
[ZARRAS]
 ( zarras );
[ZENITH]
 ( zenith );
```

References

Allen, J.F., Byron, D.K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Towards conversational human-computer interaction. *AI Magazine, 22*(4), 27-37.

Bell, L. (2003). *Linguistic adaptations in spoken human-computer dialogues – Empirical studies of user behavior.* (PhD thesis, KTH, Stockholm).

Black, A., & Lenzo, K. (2000). Limited domain synthesis. In *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP), 2*, 411-414.

Black, A., Taylor, P. & Caley, R. (1998). The Festival Speech Synthesis System. http://www.cstr.ed.ac.uk/projects/festival.html

Black, J.B. & Moran, T.P. (1982). Learning and remembering command names. In *Proceedings of the Conference on Human Factors in Computing Systems*, 8-11.

Blickenstorfer, C.H. (1995, January). Graffiti: wow!!!! *Pen Computing Magazine*, 30-31.

Bohus, Dan. (2004). *Error awareness and recovery in task-oriented spoken dialogue systems.* (Ph.D. thesis proposal, Carnegie Mellon University). http://www.cs.cmu.edu/~dbohus/docs/proposal.pdf

Boros, M., Eckert, W., Gallwitz, F., Görz, G., Hanrider, G., & Neimann, H. (1996). Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP), Addendum,* 1009-1012.

Bousquet-Vernhettes, C., Privat, R., & Vigouroux, N. (2003). Error handling in spoken dialogue systems: Toward corrective dialogue. In *Proceedings of the ISCA Workshop on Error Handling in Spoken Dialogue Systems*, 41-45.

Branigan, H.P., Pickering, M.J., & Cleland, A.A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*, B13-25.

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., & Nass, C.I. (2003). Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, 186-191.

Brennan, S.E. (1991). Conversation with and through computers. *User Modeling and User-Adapted Interaction*, *1*, 67-86.

Brennan, S.E. (1996). Lexical entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialogue*, 41-44.

Brennan, S.E. (1998). The grounding problem in conversations with and through computers. In S.R. Fussell & R.J. Kreuz (Eds.), *Social and Cognitive Psychological Approaches to Interpersonal Communication* (pp. 201-225). Hillsdale, NJ: Lawrence Erlbaum.

Burgoon, J.K., Stern, L.A., & Dillman, L. (1995). *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge: Cambridge University Press.

Carroll, J.M. & McKendree, J. (1987). Interface design issues for advice-giving expert systems. *Communications of the ACM*, *30*(1), 14-31.

Chin, D. (1984). Analysis of scripts generated in writing between users and computer consultants. In *Proceedings of the National Computer Conference, 53*, 637-642.

Clark, H.H. (1994). Managing problems in speaking. *Speech Communication, 15*, 243-250.

Clarkson, P. & Rosenfeld, R. (1997). Statistical language modeling using the CMU-Cambridge Toolkit. In *Proceedings of Eurospeech*, 2707-2710.

Cohen, P.R. & Oviatt, S.L. (1995). The role of voice input in human-machine communication. In *Proceedings of the National Academy of Sciences USA, 95*, 9921-9927.

Coulston, R., Oviatt, S., & Darves, C. (2002). Amplitude convergence in children's conversational speech with animated persons. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 2689-2692.

Darves, C. & Oviatt, S. (2002). Adaptation of users' spoken dialogue patterns in a conversational interface. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 561-564.

Denecke, M. (2002). Rapid prototyping for spoken dialogue systems. In *Proceedings of COLING*.

Di Eugenio, B. & Glass, M. (2004). The Kappa statistic: A second look. *Computational Linguistics*, *30*(1), 95-101.

Domjan, M. (2005). *The essentials of conditioning and learning*. Belmont, CA: Thomson Wadsworth.

Eskenazi, M., Rudnicky, A., Gregory, K., Constantinides, P., Brennan, R., Bennett, C., et al. (1999). Data collection and processing in the Carnegie Mellon Communicator. In *Proceedings of Eurospeech*, 2695-2698.

Fischer, G., Lemke, A., & Schwab, T. (1985). Knowledge-based help systems. In *Proceedings of CHI*, 161-167.

Giles, H., Mulac, A., Bradac, J.J., & Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. In M.L. McLaughlin (Ed.), *Communication yearbook 10* (pp. 13-48). Newbury Park, CA: Sage.

Glass, J. (1999). Challenges for spoken dialogue systems. In *Proceedings of IEEE ASRU Workshop*.

Glass, J. & Weinstein, E. (2001). Speechbuilder: Facilitating spoken dialogue system development. In *Proceedings of Eurospeech*, 1335-1338.

Goldberg, J., Ostendorf, M., & Kirchoff, K. (2003). The impact of response wording in error correction subdialogs. In *Proceedings of the ISCA Workshop on Error Handling in Spoken Dialogue Systems*, 101-106.

González-Ferreras,C. & Cardeñoso-Payo, V. (2005). Development and evaluation of a spoken dialog system to access a newspaper web site. In *Proceedings of Interspeech*, 857-860.

Gorrell, G., Lewin, I., & Rayner, M. (2002). Adding intelligent help to mixed initiative spoken dialogue systems. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 2065-2068.

Grice, H. (1975). Logic and conversation. *Syntax and semantics* (Vol. 3: Speech Acts, pp. 41-58). New York: Academic Press.

Guindon, R., Shuldberg, K., & Conner, J. (1987). Grammatical and ungrammatical structures in user-adviser dialogues: Evidence for the sufficiency of restricted languages in natural language interfaces to advisory systems. In *Proceedings of the 25th Annual Meeting of the ACL*, 41-44.

Gustafson, J., Larsson, A., Carlson, R., & Hellman, K. How do system questions influence lexical choices in user answers? In *Proceedings of Eurospeech*, 2275-2278.

Hakulinen, J., Turunen, M. & Räihä, K-J. (2006). *Tutoring in a spoken language dialogue system.* (Tech report A-2006-3, University of Tampere Department of Computer Sciences). http://www.cs.uta.fi/reports/pdf/A-2006-3.pdf

Halpern, D.F. (2000). *Sex differences in cognitive abilities.* [LOC?]: Lawrence Erlbaum.

Helander, M. (1997). Systems design for automatic speech recognition. In M. Helander, T.K. Landauer, & P. Prabhu (Eds.), *Handbook of human-computer interaction* (pp. 301-319). Amsterdam: Elsevier Science BV.

Hendler, J.A. & Michaelis, P.R. (1983). The effects of limited grammar on interactive natural language. In *Proceedings of CHI*, 190-192.

Hockey, B.A., Lemon, O., Campana, E., Hiatt, L., Aist, G., Hieronymus, J., et al. (2003). Targeted help for spoken dialogue systems: Intelligent feedback improves

naïve users' performance. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL).*

Hone, K. & Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering, 6*(3/4), 287-305.

Howell, M., Love, S., & Turner, M. (2005). Spatial metaphors for a speech-based mobile city guide service. *Personal and Ubiquitous Computing, 9*(1), 32-45.

Huang, D., Alleva, F., Hon, H.W., Hwang, M.Y., Lee, K.F., & Rosenfeld, R. (1993). The Sphinx-II speech recognition system: An overview. *Computer, Speech and Language, 7*(2), 137-148.

Jaccard, J. & Wan, C.K. (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage.

Jackson, M.D. (1983). Constrained languages need not constrain person/computer interaction. *SIGCHI Bulletin*, *15*(2-3), 18-22.

Kamm, C., Walker, M., & Rabiner, L. (1997). The role of speech processing in human-computer intelligent communication. *Speech Communication 23*, 263-278.

Kelly, M. (1977). Limited vocabulary natural language dialogue. *International Journal of Man-Machine Studies, 9*, 479-501.

Komatani, K., Ueno, S., Kawahara, T., & Okuno, H.G. (2003). User modeling in spoken dialogue systems for flexible guidance generation. In *Proceedings of Eurospeech*, 745-748.

Litman, D., Hirschberg, J., & Swerts, M. (2001). Predicting user reactions to system error. In *Proceedings of the 39th Meeting of the Association for Computational Linguistics (ACL'01)* , 362-369.

Matarazzo, J.D., Weitman, M., Saslow, G., & Weins, A.N. (1963). Interviewer influence on duration of interviewee speech. *Journal of Verbal Learning and Verbal Behavior, 1*, 451-458.

Meng, H., Lee, S., & Wai, C. (2000). CU FOREX: A bilingual spoken dialogue system for foreign exchange inquiries. In *Proceedings of the ICASSP*, 1229-1232.

Moreno, P.J. & Stern, R.M. (1994). Sources of degradation of speech recognition in the telephone network. In *Proceedings of the ICASSP*, 109-112.

Nakano, M., Miyazaki, N., Yasuda, N., Sugiyama, A., Hirasawa, J., Dohsaka, K. & Aikawa, K. (2000). WIT: A toolkit for building robust and real-time spoken dialogue systems. In *Proceedings of SIGdial Workshop*,150-159.

Niederhoffer, K.G. & Pennebaker, J.W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, *21*(4), 337-360.

Ogden, W.C. & Bernick, P. (1997). Using natural language interfaces. In M. Helander, T.K. Landauer, & P. Prabhu (Eds). *Handbook of human-computer interaction* (pp. 137-161). Amsterdam: Elsevier Science BV.

O'Hara, K.P. & Payne, S.J. (1999). Planning and the user interface: The effects of lockout time and error recovery cost. *International Journal of Human-Computer Studies, 50*, 41-59.

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction, *Computer Speech and Language, 9*, 19-35.

Oviatt, S.L., Cohen, P.R. & Wang, M. (1994). Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity. *Speech Communication, 15*, 283-300.

Oviatt, S., Cohen, P., Wu, L., Vergo, J., Duncan, L., Suhm, B. et al. (2000). Designing the user interface for multimodal speech and pen-based gesture applications: State of the art systems and future research directions. *Human-Computer Interaction, 15*, 263-322.

Pearson, J., Hu, J., Branigan, H.P., Pickering, M.J. & Nass, C.I. (2006). Adaptive language behavior in HCI: How expectations and beliefs about a system affect users' word choice. In *Proceedings of CHI*, 1177-1180.

Perlman, G. (1984). Natural artificial languages: Low level processes. *International Journal of Man-Machine Studies, 20*, 373-419.

Pickering, M.J. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169-226.

Porzel, R. & Baudis, M. (2004). The Tao of CHI: Towards effective human-computer interaction. In *Proceedings of HLT/NAACL*, 209-216.

Raux, A., Bohus, D., Langner, B., Black, A.W., & Eskenazi, M. (2006). Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In Proceedings of *Interspeech*, 65-68.

Rich, E. (1989). Stereotypes and user modeling. In A. Kobsa & W. Wahlster (Eds.), *User Models in Dialog Systems* (pp. 35-51). Berlin: Springer-Verlag.

Ringle, M.D. & Halstead-Nussloch, R. (1989). Shaping user input: A strategy for natural language design. *Interacting with Computers, 1*(3), 227-244.

SALT, Speech Application Language Tags, http://www.saltforum.org/.

Searle, J.R. (1970). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Shin, J., Narayanan, S., Gerber, L., Kazemzadeh, A. & Byrd, D. (2002). Analysis of user behavior under error conditions in spoken dialogs. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*.

Shneiderman, B. (1980a). Natural vs. precise concise languages for human operation of computers: Research issues and experimental approaches. In *Proceedings of the 18th Meeting of the Association for Computational Linguistics (ACL)*, 139-141.

Shneiderman, B. (1980b). *Software psychology: Human factors in computer and information systems*. Cambridge: Winthrop Inc.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer-interaction*. Reading, MA: Addison-Wesley.

Shriberg, E., Wilder, E. & Price, P. (1992). Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In *Proceedings of the DARPA Speech and NL Workshop*, 49-54.

Shriver, S. & Rosenfeld, R. (2002). Keywords for a universal speech interface. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 726-727.

Shriver, S., Rosenfeld, R., Zhu, X., Toth, A., Rudnicky, A. & Flueckiger, M. (2001). Universalizing speech: Notes from the USI project. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, 1563-1566.

Sidner, C. & Forlines, C. (2002). Subset languages for conversing with collaborative interface agents. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 281-284.

Suzuki, N. & Katagiri, Y. (2003). Prosodic synchrony for error management in human computer interaction. In *Proceedings of the ISCA Workshop on Error Handling in Spoken Dialogue Systems*, 107-111.

Swartz, L. (2003). *Why people hate the paperclip: Labels, appearance, behavior, and social responses to user interface agents*. (Honor's Thesis, Symbolic Systems program, Stanford University). http://xenon.stanford.edu/~lswartz/paperclip/paperclip.pdf

TellMe Networks Inc., 1-800-555-TELL™, http://www.1-800-555-tell.com/

Tomko, S. (2003). *Speech Graffiti: Assessing the user experience*. (Master's Thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University).

Tomko, S. & Rosenfeld, R. (2004a). Shaping spoken input in user-initiative systems. In Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP).

Tomko, S. & Rosenfeld, R. (2004b). Speech Graffiti vs. natural language: Assessing the user experience. In *Proceedings of HLT/NAACL*, companion volume, 73-76.

Tomko, S. & Rosenfeld, R. (2006). Shaping user input in Speech Graffiti: A first pass. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*.

Toth, A., Harris, T., Sanders, J., Shriver, S. & Rosenfeld, R. (2002). Towards every-citizen's speech interface: An application generator for speech interfaces to databases. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, 1497-1500.

van den Bosch, A., Krahmer, E. & Swerts, M. (2001). Detecting problematic turns in human-machine interactions: Rule-induction versus memory-based learning approaches. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, 499-506.

van Nimwegen, C., Burgos, D., van Oostendorp, H. & Schijf, H. (2006). The paradox of the assisted user: Guidance can be counterproductive. In *Proceedings of ACM CHI*, 917-926.

VoiceXML, http://www.w3.org/TR/voicexml20/

Wahlster, W. & Kobsa, A. (1989). User models in dialog systems. In W. Wahlster & A. Kobsa (Eds.), *User models in dialog systems* (pp. 4-34). Berlin: Springer-Verlag.

Walker, M., Langkilde, I., Wright, J., Gorin, A., & Litman, D. (2000). Learning to predict problematic situations in a spoken dialogue system: Experiments with How May I Help You? In *Proceedings of NAACL*, 210-217.

Ward, W. (1990). The CMU Air Travel Information Service: Understanding spontaneous speech. In *Proceedings of the DARPA Speech and Language Workshop*, 127-129.

Weintraub, M., Taussig, K., Hunicke-Smith, K., & Snodgrass, A. (1996). Effect of speaking style on LVCSR performance. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP)*, addendum, 16-19.

Williams, J.D. & Witt, S.M. (2004). A comparison of dialog strategies for call routing. *International Journal of Speech Technology, 7*(1), 9-24.

Yankelovich, N. (1996, November-December). How do users know what to say? *Interactions*, 32-43.

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies, 34*, 527-547.

Credits