

# A comparison of dialog strategies for call routing

*Jason D. Williams*

*Silke M. Witt*

{jason.williams, switt}@edify.com  
Natural Language Solutions Group, Edify Corp  
2840 San Tomas Expressway, Santa Clara, CA 95051 USA

*Abstract:*

Advances in commercially-available ASR technology have enabled the deployment of “How-may-I-help-you?” interactions to automate call routing. While often preferred to *menu-based* or *directed* dialog strategies, there is little quantitative research into the relationships among prompt style, task completion, user preference/satisfaction, and domain. This work applies several dialog strategies to two domains, drawing on both real callers and usability subjects. We find that longer greetings produce high levels of first-utterance routability. Further, we show that a menu-based dialog strategy produces a uniformly higher level of routability at the first utterance in two domains, whereas an open-dialog approach varies significantly with domain. In a domain where users lack an expectation of task structure, users are most successful with a *directed* strategy, for which preference scores are highest, even though it does not result in the shortest dialogs. Callers rarely provide more than one piece of information in their responses to all types of dialog strategies. Finally, a structured dialog repair prompt is most helpful to callers who were greeted with an open prompt, and least helpful to callers who were greeted with a structured prompt.

*Keywords:*

Call routing, call steering, dialogue strategy, prompt design, natural language, earcon

# Table of Contents

<b>1. Introduction and Motivation .....</b>	<b>3</b>
<b>2. Background and approaches .....</b>	<b>3</b>
<b>3. Experimental background .....</b>	<b>5</b>
3.1. Two types of usability tests: open and closed .....	5
3.2. Domains .....	5
3.3. Experimental assumptions and limitations .....	6
3.4. Summary of Experimental Batteries .....	6
<b>4. Inquiries and experiments .....</b>	<b>6</b>
4.1. Experiment Battery A .....	6
4.1.1. Design of battery A .....	6
4.2. Inquiry 1: Routability vs. introduction type .....	8
4.2.1. Summary: Inquiry 1 .....	10
4.3. Inquiry 2: Effects of the dialog strategy on routability .....	10
4.3.1. Summary: Inquiry 2 .....	10
4.4. Experimental Battery B .....	10
4.4.1. Design of Battery B .....	10
4.4.2. Assessing task completion & user preference .....	11
4.5. Inquiry 3: Task completion in a closed environment .....	11
4.5.1. Summary: Inquiry 3 .....	12
4.6. Inquiry 4: Task completion vs. user preference .....	12
4.6.1. Summary: Inquiry 4 .....	14
4.7. Experimental Battery C .....	14
4.7.1. Design of Battery C .....	14
4.8. Inquiry 5: Effects of the domain on routability .....	15
4.8.1. Summary: Inquiry 5 .....	19
4.9. Inquiry 6: effects of a structured repair prompt .....	19
4.9.1. Summary: Inquiry 6 .....	20
4.10. Inquiry 7: Amount of information in first user turn .....	20
4.10.1. Summary: Inquiry 7 .....	21
<b>5. Conclusions .....</b>	<b>21</b>
<b>6. Acknowledgements .....</b>	<b>21</b>
<b>7. Appendix A .....</b>	<b>21</b>
<b>8. Notes .....</b>	<b>23</b>
<b>9. Bibliography.....</b>	<b>24</b>

# 1. Introduction and Motivation

Recent advances in commercially available speech recognition technology have enabled successful deployment of “How may I help you?”-style prompts for call routing. This *open* dialog strategy appeals as more natural than more restricted styles of interaction, such as a *menu* dialog strategy (e.g., “Main Menu – please say ‘check balances, transfer funds or order checks.’”). *Open* dialog strategies are often believed to increase caller satisfaction, decrease task completion times, and increase task completion rates. While these intuitions seem reasonable, there is currently little evidence correlating choice of dialog strategy with user satisfaction, task completion rates, and domain.

This work seeks to explore the inter-relationships among user satisfaction, task completion, domain, and dialog strategy for ASR-based call-routing applications by undertaking the following inquiries:

1. Within one domain, what is the optimal wording and earcon usage for the system introduction/greeting?
2. Within one domain, what are callers’ initial reactions to different dialog strategies and what information do they contain?
3. Within one domain, how does task completion correlate with dialog strategy?
4. Within one domain, how does user preference correlate with dialog strategy and task completion?
5. Across two domains, what variations in response content can be observed for different dialog strategies?
6. Within one domain, what is the effect of a structured system repair prompt following different initial dialog strategies?
7. Within one domain, how much information do users’ utterances contain for different dialog strategies?

This paper is organized as follows. Section 2 explores relevant background material, both motivating the need for this inquiry as well as putting this work in the context of existing findings. Section 3 introduces some background on the 3 batteries of experiments we conduct in this work, labelled Batteries A-C. Section 4 is divided into the 7 inquiries above, introducing each battery of experiments as it is required.

Portions of this work have previously appeared in conference proceedings. In particular, portions of Battery A were presented in (Williams et al., 2003 [1]), portions of Battery B were presented in (Williams et al., 2003 [2]), and portions of Batteries C and A were presented in (Witt and Williams, 2003). The present work seeks to bring together these analyses and expand on them through comparative studies.

## 2. Background and approaches

Call routing is concerned with determining a caller’s task. Before exploring the literature it is useful to define three basic dialog strategies, shown in Table 1. These categories are necessarily broad and not all practical systems will fall distinctly into one category. That said, this taxonomy is generally accepted in practice, is motivated from findings in dialog analysis, and captures useful distinctions between practical systems.

<b>Dialog Strategy</b>	<b>Description</b>
<i>Open strategy</i>	<i>In one question/answer pair, determine the caller's task by asking an open question like "How may I help you?" or "What can I help you with?"</i>
<i>Menu strategy</i>	<i>In one question/answer pair, determine the caller's task by providing a list of choices, such as "account balance, recent transactions, transfer funds, or order checks."</i>
<i>Directed strategy</i>	<i>In multiple question/answer pairs, iteratively determine the caller's task by asking a series of questions – for example, first "Do you currently have an account with us?" / "Yes" / "What's your account number?" / etc.</i>

**Table 1: Description of dialog strategies**

Several studies to date have explored how dialog strategies affect user satisfaction and task completion rates.

(Litman et al., 1998; Swerts et al., 2000; Vanhoucke et al., 2001) give general insights into task completion and caller preference for open vs. directed prompts, but are not applied to the call routing task. Litman et al. (1998) and Swerts et al. (2000) tested a train timetable system of various levels of "open-ness". Litman et al. (1998) found that user satisfaction is shown to derive from user perception of task completion, recognition rates, and amounts of barge-in. "No efficiency measures are significant predictors of performance" for real systems, but "in the presence of perfect ASR, efficiency becomes important." Swerts et al. (2000) found that "strategies that produce longer tasks but fewer misrecognitions and subsequent corrections are preferred by users." Vanhoucke et al. (2001) explored a Yellow-pages search task. Users preferred systems which explicitly listed a few choices using more turns over systems which asked open questions (possibly in fewer turns). "It is interesting to note that even in the case of a search task, where the interaction itself should not matter as much as the information to be retrieved, users did not necessarily prefer the interface that would take them the fastest to the desired result."

Sheeder and Balough (2003) explored reactions to open prompts used for call routing. The authors compared a variety of priming phrases in conjunctions with an open prompt to find an approach that maximized task completion and caller satisfaction in a call routing task. The authors found that examples encouraging a more natural structure presented prior to the initial open question were most successful. The study was conducted in a single domain using usability subjects and explored variations of an open prompt (e.g., "What can I help you with?", "How can I help you?", "How can I help you with your account?") —the study didn't explore a direct comparison with a menu or directed strategy.

McInnes, et al. (1999) studied usability subjects' responses to a banking application's first prompt – a call routing task. Task completion is assessed using a keyword spotting metric. Table 2 summarizes the results.

<b>Dialog strategy</b>	<b>Wording</b>	<b>% Silent</b>	<b>% Containing a keyword in first try</b>
Open	"Main menu – how can I help you?,"	6%	15%
Mid	"Main menu – which service do you require?"	16%	48%
Closed	"Main menu – Please say 'help' or the name of the service you require."	10%	44%

**Table 2: Dialog strategies employed in (McInnes, et al., 1999) and results**

Subjects reported significantly higher rates for "knowing how to select options" (from Likert scores) for the *Closed* style than the others. However, overall, no clear preference between the systems was reported. 69% of subjects asked for "help" in the first turn in the *Closed style* experiment.

To our knowledge there are no studies which have made direct comparisons between a menu strategy and an open strategy (at least as defined in this work) in one or more domains.

### 3. Experimental background

This section provides background on the three batteries of experiments conducted in this work, called Batteries A, B, and C. We first describe the two types of usability studies these batteries employ, then describe their respective domains, and finally outline common experimental assumptions.

#### 3.1. Two types of usability tests: open and closed

Two types of usability tests are used for the experiments in this work: *open usability* and *closed usability*, described in Table 3.

Usability type	Description	Advantages	Limitations
<i>Open usability</i>	<ul style="list-style-type: none"> <li>Real callers are diverted from an existing hotline to a trial system.<sup>1</sup></li> <li>There is no interaction between the test facilitator and the caller.</li> </ul>	<ul style="list-style-type: none"> <li>The callers’ linguistic reactions, motivation levels, and distribution of task areas are most genuine</li> </ul>	<ul style="list-style-type: none"> <li>Callers’ satisfaction cannot be measured directly.</li> <li>Callers’ true tasks are not known, making it impossible to truly assess task completion.</li> <li>If the system replaces an existing system, new users can show a “surprise effect.”</li> </ul>
<i>Closed usability</i>	<ul style="list-style-type: none"> <li>Usability subjects are recruited to participate in a test.</li> <li>Subjects are given a specific task to complete, and are subsequently interviewed.</li> </ul>	<ul style="list-style-type: none"> <li>Subjects’ satisfaction and task completion may be directly measured.</li> <li>Subjects may be given multiple systems to trial and asked for a preference.</li> </ul>	<ul style="list-style-type: none"> <li>Subjects’ linguistic reactions and motivation levels may not be genuine</li> <li>The distribution of task areas is manufactured.</li> </ul>

Table 3: Types of usability tests

#### 3.2. Domains

This work explores two domains, called Domain I and Domain II.

Projects in domain I were initiated for an Edify customer in the US consumer electronics industry, here called Client-I. Client-I was seeking to route calls from an existing, customer-facing hotline based on task area. Example task areas in domain I included technical support, product information, requests for new repairs, status of existing repairs, and locations of retail outlets. Most task areas also required capturing the product category (for example, televisions) and/or model number (for example, model XYZ-123) to successfully route the call. In this work, however, we limit ourselves to determining the task area and do not explore the product identification task.

The caller population in domain I consisted of external, infrequent callers. Based on listening to agent/caller interactions, we believe that callers did not have a clear expectation of the task structure when calling – for example, if a product had stopped working, callers didn’t know what kind of technical support was available; if a repair was needed, callers often didn’t know the terms of their warranty or who could perform the repair.

The project in domain II was initiated for an Edify customer in the telecommunications industry, here called Client-II. Client-II was seeking to route calls from an existing, employee-facing hotline using ASR based on task area. Client II employs thousands of people and the employee hotline was primarily concerned with human-resources tasks, including updating direct deposit information, checking vacation balances, verifying health care coverage, and changing W-4 tax forms.

The caller population in domain II consisted of internal callers. Based on listening to calls between agents and callers, we believe callers have some pre-defined expectation of the task structure – for example, if a caller had changed banks and needed to

update direct deposit information, callers usually knew they needed to provide the new bank account number and new routing number. Additionally, the callers in this domain were often familiar with the existing DTMF system.

### 3.3. **Experimental assumptions and limitations**

In the work presented here, we focus on callers’ reactions and not recognition accuracy. We choose “routability” as our metric – i.e., whether there is sufficient information in an utterance to successfully route it. Note that routability gives a top-line or best-case estimate of per-utterance task completion. This choice of metric allows rapid assessment of different dialog strategies without collecting data and tuning a recognition grammars – a time-consuming process.

This choice of metric has several clear limitations: most importantly, “routability” counts can neither be said to over-state or understate actual task completion. While recognition performance will degrade task completion on a per-utterance basis, real systems employ error correction strategies over multiple turns, which tend to increase task completion on a per-attempt basis.

In this work we assume that callers’ first reactions are a suitable measure of a prompt’s success.

### 3.4. **Summary of Experimental Batteries**

In the work presented here, we conduct one battery of open usability experiments in Domain I and another in Domain II. We further explore Domain I with one battery of closed usability experiments. The characteristics of the Batteries are summarized in Table 4.

Battery	Domain	Usability type	Description in this work
<i>Battery A</i>	Domain I	Open	Section 4.1
<i>Battery B</i>	Domain I	Closed	Section 4.4
<i>Battery C</i>	Domain II	Open	Section 4.7

**Table 4: Summary of experimental batteries**

## 4. Inquiries and experiments

### 4.1. **Experiment Battery A**

In the first battery of experiments we strove to understand the effects of the system greeting in conjunction with the open and menu dialog strategies in Domain I.

#### 4.1.1. **Design of battery A**

We identified three possible introductory greetings, shown in Table 5.

Greeting name	Wording
<i>Hello</i>	“Hello, welcome to Acme”
<i>Hello + Earcon</i>	<i>[earcon]</i> “Hello, welcome to Acme”
<i>Hello + Earcon + Persona &amp; recently changed notice</i>	<i>[earcon]</i> “Hello, welcome to Acme. My name is Johnson, your virtual assistant. Please take note, this service has recently changed.”

**Table 5: Greetings (first portion of system prompt) used in Battery A**

We identified two initial prompts to reflect the dialog strategies, shown in Table 6.

Dialog Strategy	Wording
<i>Menu strategy</i>	“Please tell me which of the following options you’d like: tech support, repairs, product information, or store locations.”
<i>Open strategy</i>	“What can I help you with? [2.5 sec pause] You can get tech support, product information, repair information, or store locations. Which would you like?”

**Table 6: Dialog strategies (second part of system prompt) for Battery A**

The greetings and dialog strategies were combined to create 6 initial prompts, shown in Table 7.

Exp	Prompt construction
<i>A1</i>	Hello + Directed
<i>A2</i>	Hello + Open
<i>A3</i>	Hello + Earcon + Directed
<i>A4</i>	Hello + Earcon + Open
<i>A5</i>	Hello + Earcon + Persona & recently changed notice + Directed
<i>A6</i>	Hello + Earcon + Persona & recently changed notice + Open

**Table 7: Experiments comprising Battery A**

Each of the strategies above was implemented using a Wizard-of-Oz (WoZ) based system. An operator (the “wizard”) selected the system’s response from a pre-defined set of options in each interaction (e.g., rejection, selection of next state, transfer to an operator, etc.). The telephony interface supported barge-in throughout all prompts, and the speech/no-speech decision (i.e., activation of the end-pointer indicating user speech) was made by the telephony card (not the wizard). The same voice talent was used for all, and the voice coaching & persona attempted to maintain consistency.

Each interaction consisted of playing the initial prompt, with barge-in, and waiting for a user response. The first utterance (or lack thereof) was collected.

Category	Sub-category label	Sub-category description	
<i>Invalid (excluded from analysis)</i>	I	Invalid utterance – background noise, background voices, end-pointer error, etc. <i>Note that invalid utterances have been excluded from the percentages calculated for each of the categories</i>	
<i>Routable (success)</i>	R	Routable: utterance contained enough information (perhaps over-informative – e.g., included product category) to route the call to one of the task areas	
<i>Failures</i>	Confusion (Conf)	S	Silence – end-pointer didn’t trigger (typically indicating the caller didn’t speak) <sup>2</sup>
		C	Obvious caller confusion – e.g., “Hello?” “Is this Acme?” “Is this a real person?”
	Non-cooperative (non-coop)	H	Hang-up
		D	The caller pressed a DTMF key (note that the prompts did not include any instructions for DTMF)
	A	The caller explicitly requested to speak with an agent/real person	
Content (Cont)	U	Cooperative caller but unroutable for any reason not listed above (e.g., the caller said just a product name, or gave insufficient information to determine a task area, such as “Yes, hello, I have a question.”	

**Table 8: Classification system for utterances in Batteries A & C**

A subset of incoming calls from Client-I’s national toll-free phone number was diverted to a group of agents trained on the WoZ system. After interacting with the WoZ simulation, callers were transferred to agents. Calls were taken during business hours in morning and afternoon shifts.

Given the small fraction of calls taken by the system, and the relatively low number of repeat callers among Client-I’s caller base, we believed it was unlikely that any caller experienced more than one system variant.

The caller’s first response to each interaction was classified into one category, shown in Table 8. (This same classification system was used again in Experiment C, described below.)

Results from Battery A are presented as they are discussed in Inquiries 1 and 2.

#### 4.2. *Inquiry 1: Routability vs. introduction type*

First, invalid (I) utterances, which accounted for 0-5% of all utterances, were discarded. Table 9 shows utterance classifications for each experiment in Battery A.

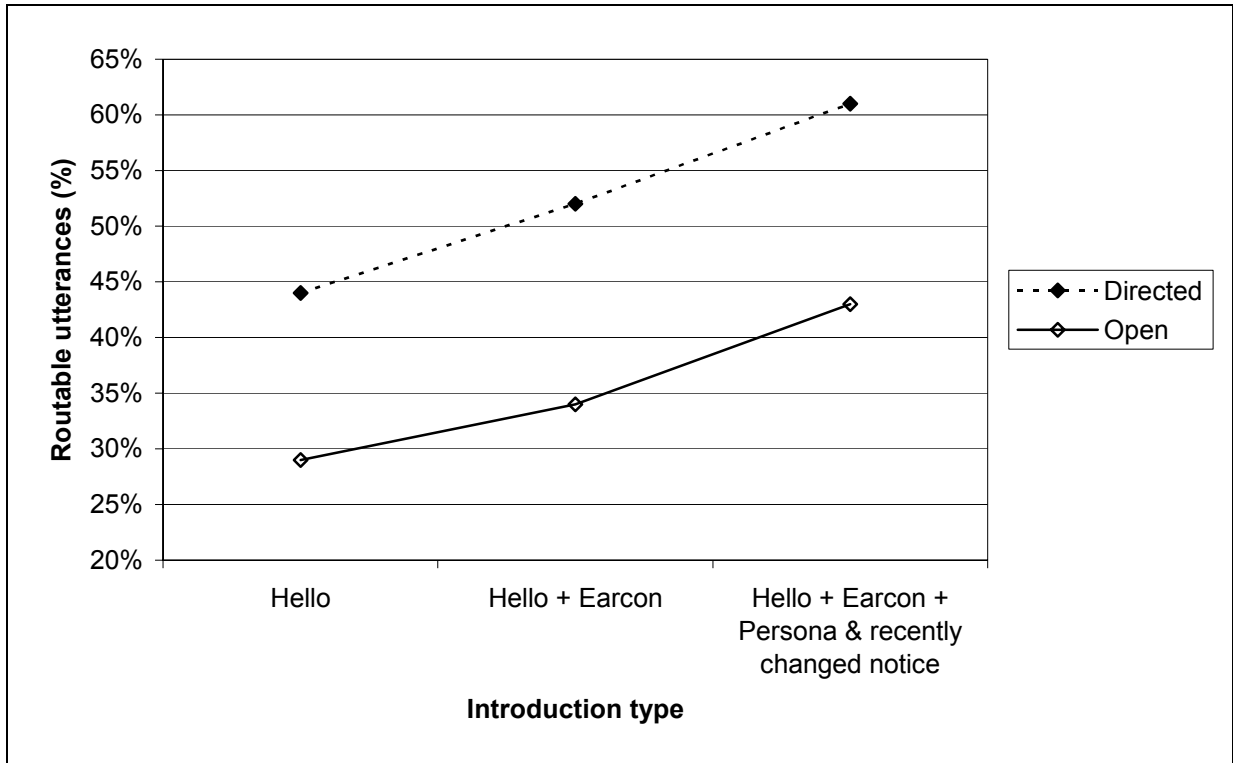
Exp	N (valid)	R	Conf		Non-coop			Cont
			S	C	H	D	A	U
A1	266	44%	42%	3%	6%	0%	0%	4%
A2	63	29%	37%	24%	2%	0%	3%	5%
A3	202	52%	37%	4%	3%	2%	0%	0%
A4	56	34%	13%	34%	5%	0%	0%	14%
A5	144	61%	26%	2%	8%	1%	1%	1%
A6	47	43%	21%	17%	6%	0%	0%	12%

**Table 9: Utterance classifications for Experiment A, first user turn**

The primary reason for *U* (Failure due to utterance content) was callers saying the name of a product in isolation with no indication of task – such as “Television.” In experiment A4, 10% of valid utterances contained a product in isolation (and 4% contained no task or product information, like “I have a question.”). In experiment A6, all 12% of *U* utterances are products in isolation.

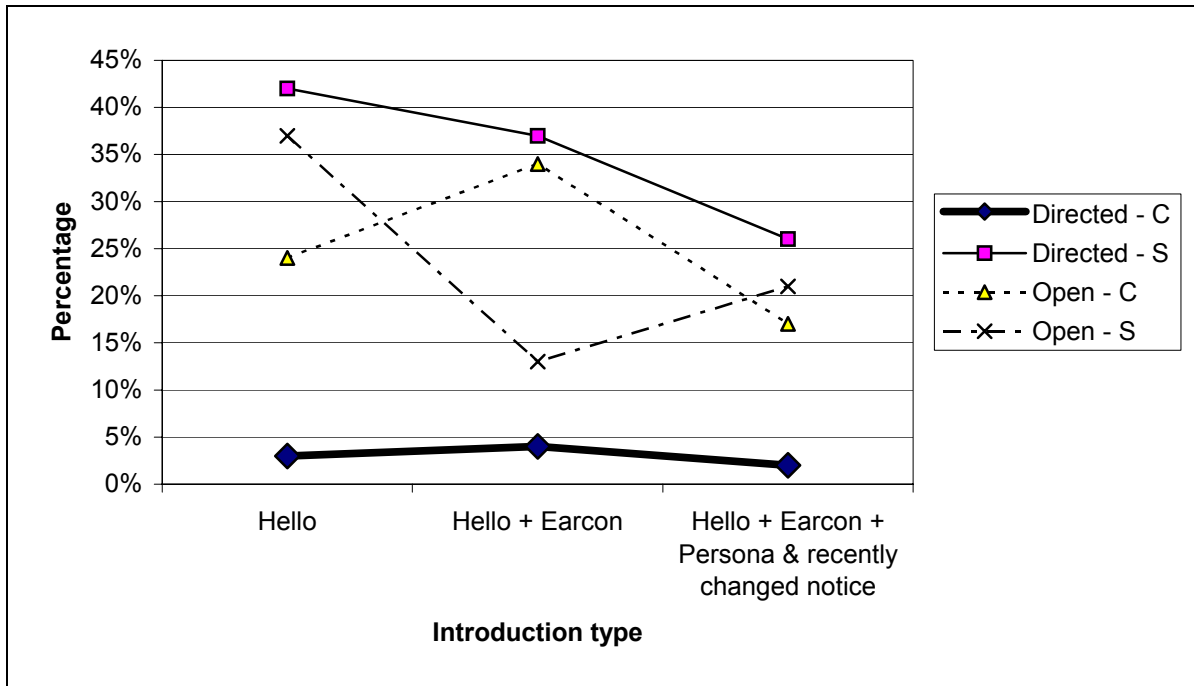
Interestingly, there is a relatively constant (5%-10%) portion of non-cooperative callers who either hang up, use DTMF, or specifically request an agent, regardless of introduction type; this group doesn’t display any clear correlation between non-cooperative behavior and greeting/dialog strategy.

Figure 1 shows the percentage of first utterances which were “routable” for each introduction type.



**Figure 1: R vs. Introduction type for Directed and Open strategies**

Figure 2 shows reasons for confusions (the primary reasons for failures) – *C* (clear user confusion) and *S* (user silence) in each experiment.



**Figure 2: Primary failure reasons in Battery A vs. introduction type**

There is a significant effect of the greeting type on the level of routability. ( $G^2(4) = 13.6, p = 0.009$ ).<sup>3</sup> The “Hello + Earcon + Persona & recently changed notice” greeting resulted in the highest levels of routability. Based on this, we concluded that this variant was optimal, and elected to use this greeting for all subsequent batteries.

One interesting observation is the increase in silence from experiment A4 to A6 (see Figure 2, up-trend of line marked “Open-S”). We noted that this was accompanied by a commensurately larger decrease in verbalized confusion. We hypothesized that adding the named persona and change-notice from experiment A4 to A6 had the effect of trading some confusions for silences, and others for routable utterances. We note that the sum of verbalized confusions and silences decreased from experiment A2 to A4 and again from A4 to A6.

#### **4.2.1. Summary: Inquiry 1**

We concluded that, in Domain I, the greeting comprised of “Hello + Earcon + Persona & recently changed notice” – the longest introduction – resulted in the highest levels of routability at the first user utterance.

### **4.3. Inquiry 2: Effects of the dialog strategy on routability**

We next seek to understand what effects the choice of dialog strategy (open vs. menu) has on the routability of the first utterance of the call within Domain I.

Analyzing the data from Table 9, choice of dialog strategy has a significant effect on the level of routability at the first turn ( $G^2(3) = 15.92, p = 0.001$ ). The menu strategy results in more routable utterances at the first turn. We hypothesize that  $U$  (unroutable due to content) utterances that contained only a product but no indication of task could be attributed to the language of the prompt – “what can I help you *with*?” as opposed to “What would you like to do?” or “How can I help you?”. If we count the  $U$  utterances which consist of a product in isolation as members of  $R$  (routable), this diminishes the gap in routability between the Open and Directed strategies, but does not alter the relationship between the two strategies. We conclude that the menu strategy is more likely to evoke a routable first utterance in Domain I.

Next, with respect to  $C$  (both silence and verbalized confusions) there is a significant effect from the greeting ( $G^2(4) = 17.9, p = 0.001$ ), but no significant effect from the prompt itself ( $G^2(3) = 7.04, p = 0.07$ ). Looking at just verbalized confusions, there isn’t a significant effect from the greeting ( $G^2(5) = 5.02, p = 0.29$ ), but there is a significant effect from the prompt strategy ( $G^2(3) = 71.72, p < 0.0001$ ).

#### **4.3.1. Summary: Inquiry 2**

Within Domain I, prompting strategy has a significant effect on routability and level of verbalized confusions. The greeting type has a significant effect on routability and the combined (silence and verbalized) levels of confusions. The menu strategy results in higher levels of routability and lower levels of confusion than the open strategy in Domain I.

### **4.4. Experimental Battery B**

We next seek to explore the effects of dialog strategy on task completion and user preference in a closed usability environment. In this endeavor we constructed Battery B. In addition, since we found that the *menu* strategy was most successful, we further hypothesized that a *directed* strategy might also be appropriate, so we added this to the battery.

#### **4.4.1. Design of Battery B**

We constructed five WoZ-based, closed usability experiments, B1 – B5. All of the systems tested used the “Hello + earcon + persona & recently changed notice” greeting examined in Battery A. The menu strategy (B1) and the open strategy (B2) from Battery A comprised two of the experiments. In addition, we created 3 *directed strategies* which sought to capture user task with multiple questions; by ordering these questions differently we created three different Directed strategies called Direct-1 (B3), Directed-2 (B4), and Directed-3 (B5). Example interactions with all of the strategies are shown in Appendix A.

Each experiment in Battery B included 2 escalating no-input (i.e., silence) and 2 no-match (i.e., invalid/out-of-grammar) prompts at each interaction point. If a subject triggered a total of more than 2 no-inputs or more than 2 no-matches at a given interaction point, an agent transfer was simulated.

Each of the strategies above was again implemented using a Wizard-of-Oz based system. The telephony interface supported barge-in. The same voice talent was used for all experiments, and the voice coaching attempted to maintain a consistent persona throughout. Again here, “Perfect” speech recognition was assumed by the wizard – i.e., all utterances which *could* be classified by the Wizard for a given state were treated as successful recognitions. Thus the system was evaluating utterance content and not utterance quality, which would impact a deployed ASR system.

Subjects were selected from a pool of the Client’s customers with a representative distribution of age, profession, and gender. Subjects were provided with a cash incentive.

Each session consisted of a short welcome and introduction from a standard script. Subjects were told they would be using two different prototypes of a new system, but not told anything about the systems. Each subject was presented with 6 tasks to perform with one system, and then with 6 different tasks to perform with a second system.

Tasks were written and presented one at a time. Example tasks included:

- You’re interested in a 17-inch TV with the picture in picture feature. You are calling to find out how much they cost.
- This camera was given to you as a gift for your birthday last week. You haven’t been able to get the camera to turn on. You are calling to get help trying to turn it on. (The camera was available to the subject in the room).

A total of 16 sessions were conducted (1 subject per session). Each session was video and audio recorded. Because of limited time and subject pool, it was not possible to compare every strategy to every other strategy. Instead, each subject used one of the Directed strategies, and either the Open or Main-menu strategy. We varied which strategy was presented first and second.

#### 4.4.2. Assessing task completion & user preference

A task attempt was graded successful if the caller would have been transferred to the correct “routing bucket” for their task – i.e., both the task and product category/model number had been correctly received by the system. No-inputs (i.e., caller silences) and no-matches (i.e., simulated recognition rejects) did not constitute a failed task attempt unless the no-input or no-match limit was reached in one interaction state.

At the end of the session, subjects were asked which system they preferred. Subjects were then read a series of statements *about the system they preferred* and asked to respond using a 7-point Likert scale, intermixed with several free-response questions. Results are presented as they are discussed in Inquiries 3 and 4.

### 4.5. Inquiry 3: Task completion in a closed environment

We first sought to understand the effect of the dialog strategy on task completion. Table 10 shows the total number of user sessions conducted for each strategy.

Experiment	Strategy	First in session	Second in session	Total experiments
B1	Main-menu	5	4	9
B2	Open	3	4	7
B3	Directed-1	2	2	4
B4	Directed-2	3	3	6
B5	Directed-3	3	3	6

**Table 10: Counts of each experiment conducted, and position (first or second) in the usability session**

Table 11 shows task completion for all task attempts (across both experiments in all sessions).

Experiment	Strategy	Success (task attempts)	Failed (task attempts)	Undecidable task attempts <sup>4</sup>
B1	Main-menu	48 (89%)	6 (11%)	0
B2	Open	31 (79%)	8 (21%)	3
B3	Directed-1	24 (100%)	0	0
B4	Directed-2	31 (86%)	5 (14%)	0
B5	Directed-3	36 (100%)	0	0

**Table 11: Task completion by strategy**

Task completion was highest for the Directed-1 and Directed-3 strategies. The difference in task completion rates between Directed-1 and the open strategy was significant (Fisher Exact Probability test (2-tailed),  $p = 0.02$ ); similarly, the task completion rates between Directed-3 and the open strategy was significant (Fisher Exact Probability Test (2-tailed);  $p = 0.006$ ). No other comparisons were statistically significant.

We looked at why main-menu and open strategy task attempts failed. We found that most of the main-menu failures were due to selecting the wrong item from the main menu. In some cases it was clear that the caller didn't know which option was suited to his/her task, an issue with menu strategies previously noted in (Carpenter and Chu-Carroll, 1998). In addition, most of the failures with the open strategy occurred when the caller waited for the delayed help following the open prompt. The delayed help had the same structure as the menu approach; most failures here were again due to selection of the wrong item. These observations suggest that the prompt wording of the open strategy could be improved, explored in Battery C, below.

From listening to caller/call center agent interactions, we believe that Directed-3 and Directed-1 most closely modeled typical conversation flow; the Directed-2 flow seemed to deviate furthest from normal flow

#### 4.5.1. Summary: Inquiry 3

Directed strategies which most closely model human/human conversations resulted in higher task completion rates than an open strategy by a statistically significant margin.

#### 4.6. Inquiry 4: Task completion vs. user preference

We next examined the results of the users' preferences between the dialog strategies in Battery B; results are summarized in Table 12.

Experiment	Strategy	Preferred (sessions)	Not preferred (sessions)	No preference (sessions)
B1	Main-menu	5 (56%)	4 (44%)	0
B2	Open	3 (43%)	3 (43%)	1 (14%)
B3	Directed-1	1 (15%)	2 (50%)	1 (25%)
B4	Directed-2	1 (17%)	5 (83%)	0
B5	Directed-3	5 (83%)	1 (17%)	0

**Table 12: Preference by strategy**

We considered whether the order a caller experienced the strategies could influence their preference; Table 13 shows that preference was not correlated with experiment position.

First experiment preferred	Second experiment preferred	No preference expressed
8 (50%)	7 (44%)	1 (6%)

**Table 13: Preference for a strategy by position of experiment (first vs. second) in usability session**

Table 14 shows the questions scored on the Likert scale, and Table 15 shows the resulting scores for the strategy that callers preferred.

ID	Statement (for Likert questionnaires)
1	I felt in control while using the speech system to perform my tasks
2	Overall, this speech system gave me an efficient way to perform my tasks
3	Overall I was satisfied with my experience using the speech system to accomplish my tasks
4	Overall, I was comfortable with my experience calling this speech system.
5	When the system didn't understand me, or if I felt hesitant, it was easy to get back on track and continue with the task
6	Overall, the system was easy to use
7	The choices presented to me by the system were what I expected to complete the tasks

**Table 14: Statements used for Likert scale responses**

Strategy	N	1	2	3	4	5	6	7	Av
Main-menu	5	5.4	6.2	6.4	6.2	5.5	6.2	5.2	5.9
Open	3	5.7	5.7	6.0	6.3	7.0	7.0	6.0	6.2
Directed-1	1	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0
Directed-2	1	5.0	6.0	5.0	6.0	3.0	5.0	6.0	5.1
Directed-3	5	6.0	6.6	6.8	7.0	6.6	7.0	6.6	6.7

**Table 15: Likert scale responses by strategy**

User preference rates were highest for the Directed-3 strategy. While Directed-3 was not compared directly to Directed-1 and Directed-2, it seems reasonable to conclude that Directed-3 would be preferred to Directed-2 and Directed-1 given their relative scores vis-à-vis the Open and Menu-based strategies.

In analyzing the subject test scores, we discarded results from Directed-1 and Directed-2 (which were selected only once each). At a confidence interval of  $p \leq 0.05$  (using the Mann-Whitney test as in Lowry (2003)), and taking all Likert scores together as a collective measurement of preference, Directed-3 scored higher than Main-menu ( $p \leq 0.0052$ ). There were no significant comparisons between Directed-3 and the Open strategies; at a confidence interval of  $p \leq 0.07$ :

- Users were more satisfied (question 3) with Directed-3 strategy than with the Open strategy,
- Users were more comfortable (question 4) with Directed-3 than with the Closed strategy, and
- Users thought that the Directed-3 strategy was easier to use than the Closed strategy (question 6).

Anecdotal observations for the Main-menu strategy included:

- Subjects who preferred the Main-menu strategy cited they liked “knowing all their options.”
- Subjects who didn't prefer the Main-menu strategy had trouble determining which menu option was appropriate.

We speculate that presenting a list of choices risked a *false sense of control* as the options were interpreted differently by the subjects.

Anecdotal observations for the Open strategy included:

- Most subjects using this system did not respond before the pause in the initial prompt.

In addition, subjects who did respond to the open question were more likely to respond to subsequent questions with longer utterances. We speculated that, when asked an open question, subjects often lacked an expectation of the system's abilities and usually waited for more guidance.

Anecdotal observations for the directed strategies included:

- Subjects who preferred the Directed Strategy liked that the system was “taking charge of their problem” and leading them through their choices.
- Subjects who didn't prefer the Directed Strategy wanted a sense of all their choices.

One interesting finding, previously shown in (Litman et al., 1998), is that the number of turns (and by extension the elapsed time, as each turn was approximately equal in duration) doesn't seem to influence caller satisfaction: all of the directed

strategies' calls were longer than the open- and menu-strategy calls. As noted by Litman (1998), task completion seems to be a more reliable predictor of preference/satisfaction.

#### **4.6.1. Summary: Inquiry 4**

Although data scarcity limits our conclusions from this experiment, there is evidence that subjects preferred a particular directed strategy over the main menu strategy. While there existed a fundamental trade-off between “*knowing all the choices*” (menu strategy) and “*guidance through the choices*” (directed strategy), on the balance, in this domain, subjects preferred the directed strategy that most closely mirrored human/human interactions.

### **4.7. Experimental Battery C**

For call routing in Domain I, Batteries A and B have given evidence that system-initiative strategies were more preferred and more successful than open strategies. We next explored whether callers would be more successful with an open strategy in a domain in which callers had a stronger mental model of task. We test this hypothesis by adapting Battery A to Domain II, creating Battery C. Battery C augments Battery A in the following respects:

- Battery C is in Domain II; Battery A was in Domain I.
- All prompts in battery C use the “Hello + persona & recently changed notice” greeting.<sup>5</sup>
- Sheeder and Balough (2003) show how variation in utterance content was observed by varying the structure and placement of examples in the initial prompt. Thus we selected a variety of prompts which would both allow direct comparison with Battery A as well as incorporate best practices from (Sheeder and Balough, 2003).
- In addition to an Open and a Main Menu strategy prompt, we also added a prompt which didn't give specific options but which invited a more specific response – in this case, “Please tell me the reason for your call.” We call this an “OpenRequest” strategy.
- We extended the data capture process to look at both the first & (if present) second utterances from callers to test our re-prompting strategy.
- Lastly, we wanted to relax the assumption that only 1 slot could be captured at one time, and explore how often callers provided multiple slots. In Domain II, additional slots were used to specify sub-tasks – for example, for the task “direct deposit”, sub tasks included “set up new direct deposit”, “stop direct deposit”, “change direct deposit details”.

#### **4.7.1. Design of Battery C**

The design of Battery C was identical to Battery A, with the following 4 modifications:

1. Escalating no-match and no-input messages were added. The 2nd system prompt is shown in Table 16.
2. Real speech recognition using a “bootstrapping” grammar was used. The bootstrapping grammar underperformed a tuned grammar, and in particular rejected a number of routable utterances. In assessing routability of the first utterance, we hand-scored the utterances as was done in Battery A (i.e., we did not use the recognition results). We note that the system treated a portion of callers who produced a routable utterance at the initial prompt as though the utterance was not routable; thus a portion of responses to the 2<sup>nd</sup> prompt followed a routable response to the first prompt. (This number is quantified below.)

Prompt Segment	Prompt text
<i>Hello</i>	Hello. Thanks for calling our Human Resources Self Service System. My name is Sarah and this is our new speech recognition system.
<i>Open</i>	What would you like to do?
<i>OpenRequest</i>	Please tell me the reason for your call.
<i>Following examples</i>	For example, you can say: I need to reset my web password or I'd like to talk to someone about savings plans.
<i>Preceding Examples</i>	Here are some examples of what you can say: I need to reset my web password or I'd like to talk to someone about savings plans.
<i>Main Menu</i>	To help me direct your call, please choose from the following options: Web password reset, course enrollment, direct deposit, or benefits. If you didn't hear what you want, say more options.
<i>More Options</i>	Here are the rest of the options. When you hear the one you want, just say it. Non-management jobs, employment verification, w4, health plans, training information, absences, fax-on-demand, vacation balances, career path, instructor reports, employee resources.
<i>No-input Re-prompt</i>	Sorry, I didn't hear anything. Please say Web password reset, course enrollment, direct deposit, or benefits. If you didn't hear what you want, say more options.
<i>No-match Re-prompt</i>	Sorry, I didn't understand. Please say Web password reset, course enrollment, direct deposit, or benefits. If you didn't hear what you want, say more options.

**Table 16: Prompt types and wordings for Battery C**

- The prompt wordings used in Battery C (listed in Table 16) were different than the prompt wordings used in Battery A.
- The logging of the system used in Battery C was more limited. In particular, it was not possible to log hang-ups, nor DTMF usage.

Noting (4) above, each first utterance was classified using the same taxonomy as in Table 8. For experiment C5, “More options” utterances were counted as routable.

Table 16 lists the wording of the prompts used in Battery C; these were then combined to form the experiments listed in Table 17.

Exp	Corresponding experiment A	Prompt Construction
C1	--	Hello + Preceding examples + Open
C2	A6	Hello + Open + [2.0 s pause] + Following examples
C3	--	Hello + Preceding examples + OpenRequest
C4	--	Hello + OpenRequest + [2.0 s pause] + Following examples
C5	A5	Hello + Main Menu

**Table 17: Experiments in Battery C**

#### **4.8. Inquiry 5: Effects of the domain on routability**

The results for the users’ reactions to the first prompt are shown in Table 18. The routability rates for Battery C do not vary greatly. Experiment C3 results in less routability than C5 (Chi-Square/Yates;  $p = 0.047$ ) and Experiment C4 results in less routability than C5 (Chi-Square/Yates;  $p = 0.043$ ), but none of the other relationships are significant. In particular, no significant relationship is found (in first-utterance routability) between prompts which provide examples before or 2.0 s after posing two different questions ( $G^2(2) = 0.02$ ;  $p = 0.99$ ). Thus we have not validated Sheeder and Balough’s (2003) finding that naturally phrased examples which precede a question result in higher task completion than naturally phrased examples which follow. This may be due to differences between our experimental designs – Sheeder and Balough’s work examined actual task completion of a working system and includes the affects of re-tries and ASR errors whereas our work looks at just the content of the first utterance;

also, Sheeder and Balough used usability subjects whereas we use real callers. Lastly, there could be a difference in the effect of the domain.

Exp	Num	R	Conf		Non-coop			Cont
			S	C	H	D	A	U
C1	312	67%	29%	0%	--	--	3%	2%
C2	285	67%	25%	2%	--	--	3%	3%
C3	393	64%	29%	3%	--	--	2%	3%
C4	371	64%	31%	3%	--	--	0%	2%
C5	390	71%	22%	5%	--	--	0%	3%

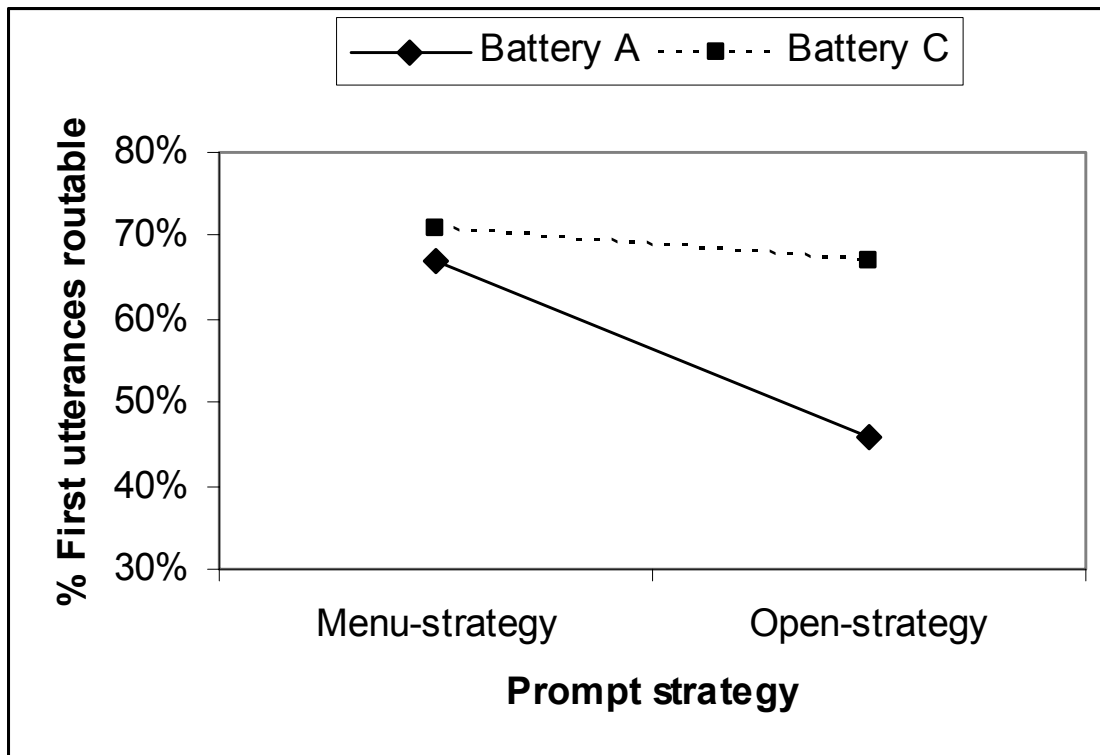
**Table 18: Result Summary for Battery C, first user turn**

Desc	Exp	Num	R	Conf		Non-coop			Cont
				S	C	H	D	A	U
Menu	A5	131	67%	29%	2%	--	--	1%	1%
	C5	390	71%	22%	5%	--	--	0%	3%
Open	A6	44	46%	23%	18%	--	--	0%	13%
	C2	285	67%	25%	2%	--	--	3%	3%

**Table 19: Comparison of corresponding (first user turn) experiments in Battery C and Battery A**

Table 19 shows statistics for corresponding experiments in Battery A and Battery C. The utterance counts and percentages for Battery A have been adjusted to exclude DTMF and hang-ups (as these were not tracked for Battery C) to facilitate a direct comparison of present data. Utterances for “More options” were counted as Routable, and accounted for at most 1.5 % of total utterances.

Both the domain ( $G^2(2) = 8.18; p = 0.017$ ) and the dialog strategy ( $G^2(2) = 7.22; p = 0.027$ ) have significant effects on routability. In figure 3, we compare routability in Batteries A and C directly.



**Figure 3: Comparison of routability in Battery A and Battery C, first user turn**

Figure 3 illustrates the affects of the two strategies in the domains. Recall that in domain I, we believe that callers have little expectation of task structure, and in domain II, we believe that callers have a much stronger notion of task structure. Figure 3 supports this assertion. While the menu-based strategy performs similarly in the two domains, the open-strategy performs much worse in domain I (Battery A) than in domain II (Battery C).

We believe this is further supported by looking at the relative levels of ‘Conf’ utterances in the open strategy in Figure 4. While there is no significant effect due to the domain alone ( $G^2(2) = 4.2; p = 0.117$ ) on Confusion levels, there is a very significant interaction ( $G^2(4) = 22.26; p = 0.0002$ ).

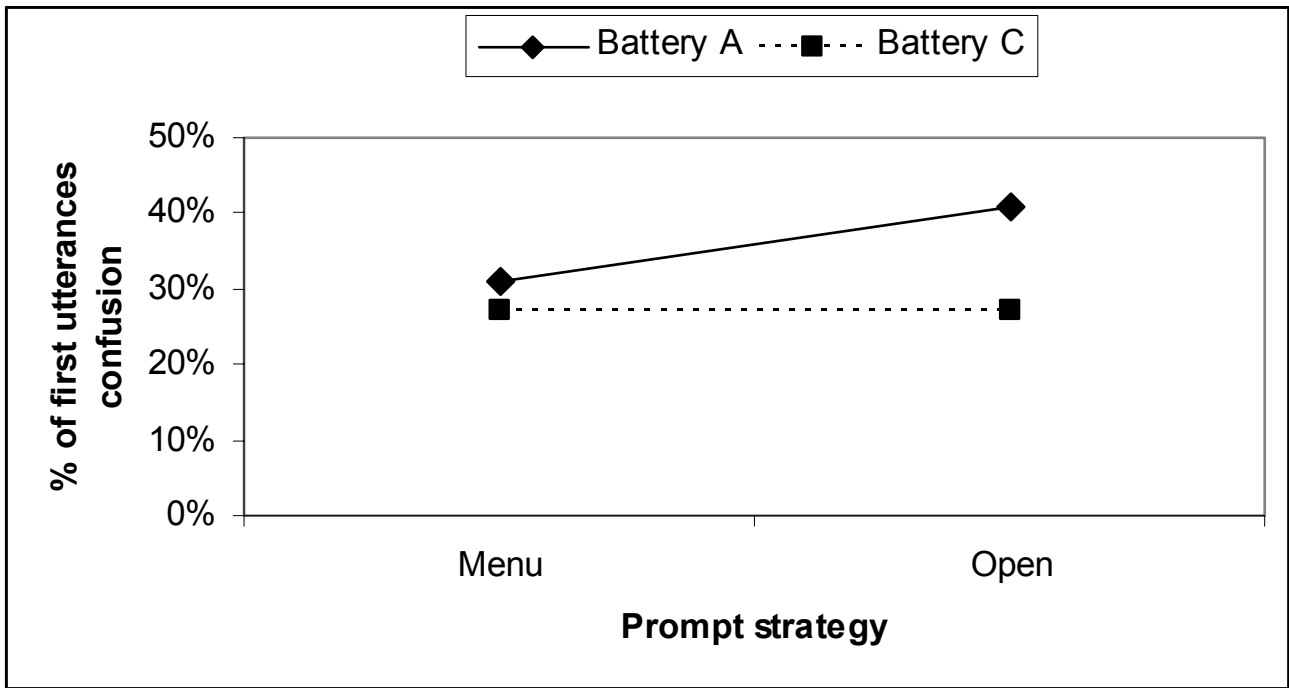


Figure 4: Comparison of confusions expressed in Battery A and Battery C, first utterance

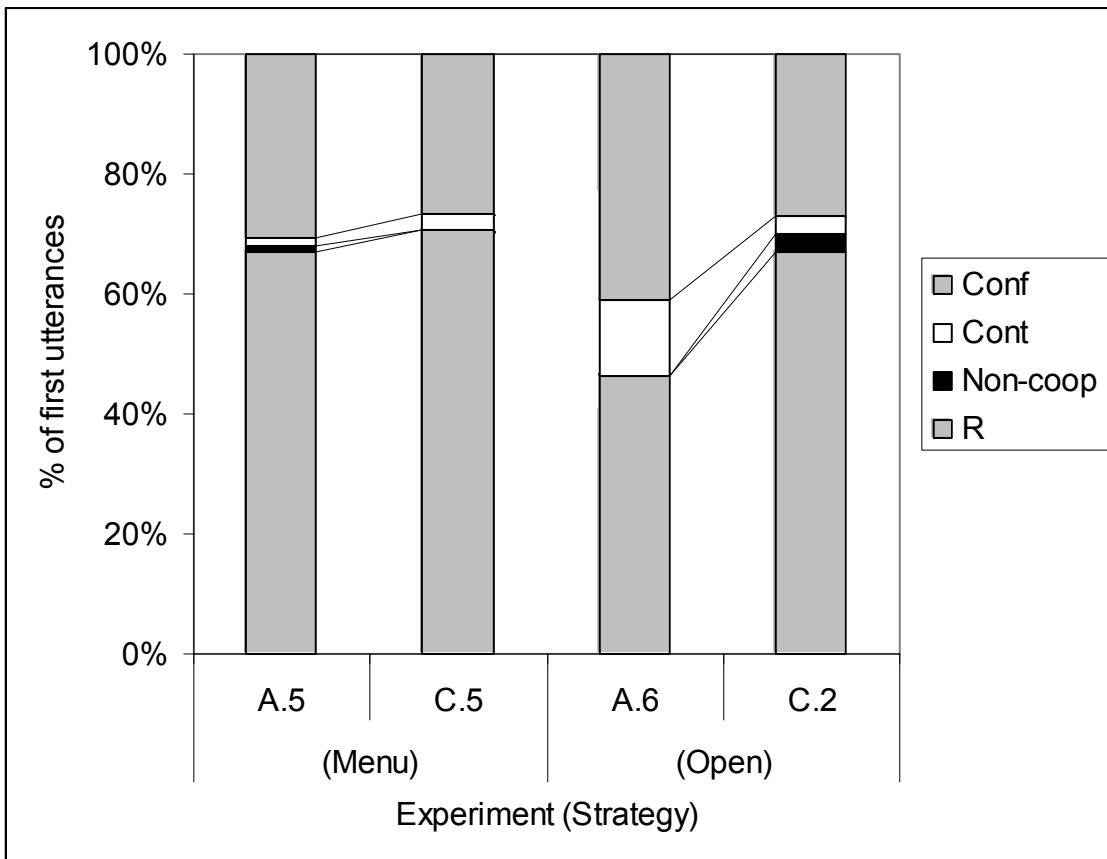


Figure 5: Breakdown of utterance classification for menu- and open-strategies in two domains

Figure 5 summarizes the comparison between Domain I and Domain II, and uses the following nomenclature: *R* = Routable, *Conf* = all confusion types – silence or verbalized confusion, *Non-Coop* = non-cooperative user, and *Cont* = unroutable due to utterance content.

### 4.8.1. Summary: Inquiry 5

In Domain II, a number of strategies produced remarkably similar routability rates.

Comparing Domain I to Domain II, we see that the domain has a significant effect on Routability, and that interaction between the domains and prompt type has a significant affect on confusion levels.

### 4.9. Inquiry 6: effects of a structured repair prompt

We next turned our attention to the effects of the 2<sup>nd</sup> prompt in Battery C. In general, all non-routable utterances at the first prompt resulted in a 2<sup>nd</sup> system prompt; in addition, some routable responses to the first prompt were falsely rejected by the bootstrap grammar. Table 20 summarizes these two categories of callers presented with the 2<sup>nd</sup> prompt.

Exp	First utterance		Second utterance
	False reject of routable first utterance	Correctly rejected unroutable first utterance	Total attempts
C1	66 (39%)	103 (61%)	169
C2	82 (47%)	93 (53%)	175
C3	19 (12%)	142 (88%)	161
C4	41 (24%)	133 (76%)	174
C5	66 (36%)	115 (64%)	181

**Table 20: Break down of cause of caller progressing to 2<sup>nd</sup> system turn**

Utterances spoken in response to the 2<sup>nd</sup> prompt were categorized using the same taxonomy as was used for the first utterance; Table 21 shows the results.

Exp	Num	R	Conf		Non-coop			Cont
			S	C	H	D	A	U
C1	169	73%	27%	0%	--	--	0%	1%
C2	175	74%	23%	2%	--	--	1%	3%
C3	161	66%	25%	4%	--	--	1%	7%
C4	174	63%	32%	3%	--	--	0%	5%
C5	181	64%	33%	3%	--	--	1%	3%

**Table 21: Result Summary for Battery C, 2<sup>nd</sup> utterance**

Figure 6 compares the percentage of routable utterances at the initial and the secondary prompt. The largest increase in the 2<sup>nd</sup> prompt occurs in C2 (open + following examples). The smallest additional gain occurs for the menu-based approach. This suggests that a dialog design that starts with an open prompt and follows with a directed prompt in the case of a recognition failure produces a maximal level of routable utterances. This is intuitively satisfying: an open question followed by explicit choices seems to provide just-in-time help for users who need it. On the other hand, in C5 (menu strategy), the initial and 2<sup>nd</sup> prompts are virtually identical – the re-prompting doesn't provide the caller with additional information. As a possible consequence, C5 scores lowest for routability of the 2<sup>nd</sup> utterance.

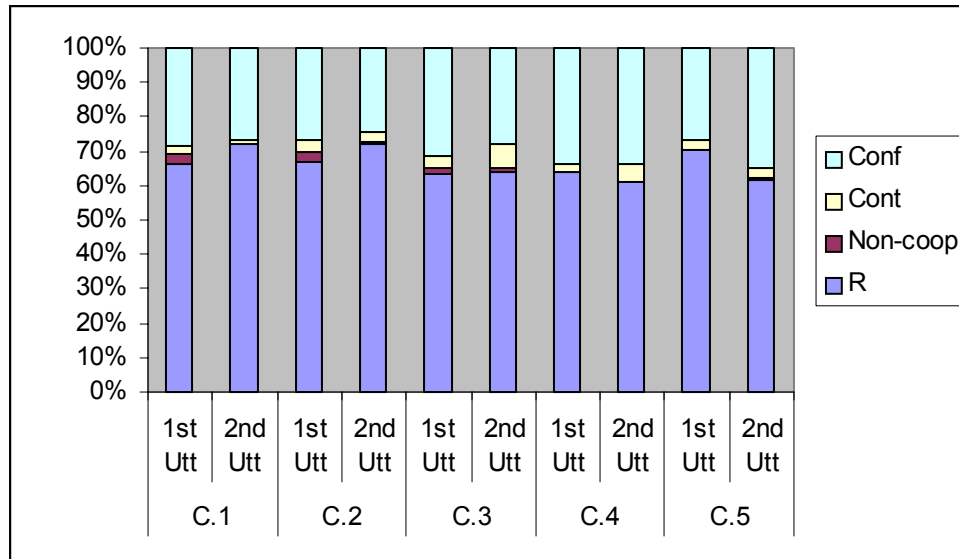


Figure 6: Routability of 1<sup>st</sup> and 2<sup>nd</sup> user turn in Battery C

#### 4.9.1. Summary: Inquiry 6

Using a menu-based re-prompt following an open style initial prompt increases routability more than when it follows a menu-initial prompt.

#### 4.10. Inquiry 7: Amount of information in first user turn

In Battery C, we sub-divided the routable category into 2 sub-types – 1-slot and 2-slot.<sup>6</sup> Table 22 and Table 23 show the distribution of one-slotted and two-slotted utterances on the first turn and 2<sup>nd</sup> turn, respectively.

Exp	%R (ALL)	%R (one-slot)	%R (two-slot)	%R (“more options”)
C1	67%	63%	3%	1%
C2	67%	62%	4%	1%
C3	64%	60%	3%	2%
C4	64%	63%	1%	0%
C5	71%	67%	3%	0%

Table 22: Breakdown of routable utterances; (% of all first user turn utterances)

Exp	%R (ALL)	%R (one-slot)	%R (two-slot)	%R (“more options”)
C1	73%	47%	0%	26%
C2	74%	43%	1%	30%
C3	66%	32%	4%	30%
C4	63%	31%	0%	32%
C5	64%	38%	1%	25%

Table 23: Breakdown of routable utterances (% of all second user turn utterances)

Two-slotted utterances represent a small fraction of caller responses across all dialog strategies. Further, when given the choice of “more options” after a dialog failure, a large portion of callers select it across all strategies. After a rejection, a relatively consistent portion of callers seek out guidance from the system, no matter what initial dialog strategy is employed.

Finally, the proportion of two-slotted utterances is usually higher at the 1<sup>st</sup> utterance, indicating that callers who provide more information initially but aren’t understood remove content from their speech.

### 4.10.1. Summary: Inquiry 7

Frequency of two-slotted utterances appears to be uniformly low across all dialog strategies. When a menu-style 2<sup>nd</sup> prompt is used after a system rejection, a large percentage of callers select “more-options” regardless of the initial prompt.

## 5. Conclusions

In applying ASR to call routing, we have explored the relationships among domain, prompt/greeting wording, task completion/utterance routability, and user preference.

First, we found that a longer greeting including an earcon, named persona, and recently changed notice increased the routability of the first utterance for both menu- and open-strategy prompts. We believe the key factors in this finding are the average caller frequency and expectation for ASR-based system. Given a generally low-level of the latter at this time, we believe that this finding is extendable from its testbed domain to any domain in which the caller base is relatively infrequent.<sup>7</sup>

Through usability subjects, we found that caller preferences roughly correlated with task completion rates, agreeing with previous studies (Litman et al., 1998). In addition, a directed strategy was found to have higher task completion scores than an open strategy.

By applying the same prompt design to another domain, we have shown that for domains in which callers have a more structured expectation of task, more initial responses (to either a menu- or open-strategy prompt) are routable than for domains in which callers have a less-structured sense of task. Further, measuring performance by routability of the first utterance, we have shown that a menu strategy outperforms an open-strategy in the two domains investigated here, but by a much smaller margin in a domain where callers have a stronger sense of task structure. In addition, callers rarely provide more than 1 slot in response to a range of dialog strategies. Lastly, after a dialog failure, a structured response helps callers who first heard an open prompt most; in addition, after a dialog failure which offers more information, a high portion of users take up that offer.

## 6. Acknowledgements

The authors would like to thank Judee Humberg for providing valuable input on the of the usability test procedure conducted in Battery B, and for her input on the interview questions.

## 7. Appendix A

The following figures show sample dialogs with the three experimental systems from Battery B.

**System:** *Hello, welcome to <Client I>. My name is Johnson, your virtual assistant. Please take note: this service has recently changed. Now, please tell me which of the following options you'd like: tech support, repairs, product information, or store locations.*

**User:** Ah, product information

**System:** *Ok, product information. What product are you calling about?*

**User:** It's a television – seventeen inch television

**System:** *What's the model number? It's ok to say, I don't know.*

**User:** Ah, I don't know.

**System:** *No problem. I'm transferring you to an agent who can help you with your question.*

**Figure 7: Main-menu strategy (sample dialog)**

**System:** *Hello, welcome to <Client I>. My name is Johnson, your virtual assistant. Please take note: this service has recently changed. Now, what can I help you with? [2.5 sec pause] You can get tech support, product information...*

**User:** *Ah, yes, I need... I'm thinking about buying a TV*

**System:** *Ok, product information. What product are you calling about?*

**User:** *It's a television – seventeen inch television*

**System:** *What's the model number? It's ok to say I don't know.*

**User:** *Ah, I don't know.*

**System:** *No problem. I'm transferring you to an agent who can help you with your question*

**Figure 8: Open strategy (sample dialog)**

**System:** *Hello, welcome to <Client I>. My name is Johnson, your virtual assistant. Please take note: this service has recently changed. To get started, if you've got an event number or work order number, tell me it now. Otherwise, say NEW QUESTION and we'll start from the beginning.*

**User:** *New Question*

**System:** *Ok, a new question What product are you calling about? Or, if it's an accessory you're after, say ACCESSORIES.*

**User:** *It's a Television*

**System:** *What's the model number? It's ok to say, I DON'T KNOW*

**User:** *Ah, I don't know*

**System:** *No problem. Do you own this product now?*

**User:** *No*

**System:** *Great. I'm transferring you to an agent who can help you.*

**Figure 9: Directed-1 strategy (sample dialog)**

**System:** *Hello, welcome to <Client I>. My name is Johnson, your virtual assistant. Please take note: this service has recently changed. To get started, do you need help with a product you own now?*

**User:** *No*

**System:** *Oh, ok, what product are you calling about?*

**User:** *It's a Television*

**System:** *What's the model number? It's ok to say, I DON'T KNOW*

**User:** *Ah, I don't know*

**System:** *No problem. I'm transferring you to an agent who can help you.*

**Figure 10: Directed-2 Strategy (sample dialog)**

**System:** *Hello, welcome to <Client I>. My name is Johnson, your virtual assistant. Please take note: this service has recently changed. To get started, if you've got an event number or work order number, tell me it now. Otherwise, say NEW QUESTION and we'll start from the beginning.*

**User:** *New Question*

**System:** *Ok, a new question. To get started, are you calling about a product you own now?*

**User:** *No*

**System:** *Oh, ok. What product are you calling about? Or, if it's an accessory you're after, say ACCESSORIES.*

**User:** *It's a Television*

**System:** *What's the model number? It's ok to say, I DON'T KNOW*

**User:** *Ah, I don't know*

**System:** *No problem. I'm transferring you to an agent who can help you.*

**Figure 11: Directed-3 strategy (sample dialog)**

## 8. Notes

<sup>1</sup> The “trial system” may be either a real system or a Wizard-of-Oz system.

<sup>2</sup> Here we classify silence as a type of confusion – i.e., the caller didn’t know what to say, or didn’t know that they should speak. “It has been observed that if the caller experiences problems with the service or becomes confused about what response is expected, then they will tend to remain silent while they work out what to do.” [5].

<sup>3</sup>  $G^2(n)$  indicates out test statistic -- the log-linear approximation to the Chi-Square distribution with  $n$  degrees of freedom. The null hypothesis in all cases is that the dimension being explored has no effect. For more information see (Lowry, 2003).

<sup>4</sup> One subject exhibited extensive confusion with the Open strategy, and their results are excluded form the percentages reported.

<sup>5</sup> The earcon was omitted from this greeting due to the requirements of Client II; given the explicit explanation of the nature of the system, we believe this still allows a direct comparison between Batteries A and C.

<sup>6</sup> See section 4.7 for a description of “1-slot” and “2-slot” utterances.

<sup>7</sup> We believe the significant characteristic of the greeting is that it was longer and provided more explanation about the system – the fact that the system’s persona was *named* we don’t believe was significant.

## 9. Bibliography

- Carpenter, B. and Chu-Carroll, J. (1998). Natural Language Call Routing: A Robust Self-Organizing Approach. Proceedings of ICSLP 1998. Sydney, Australia: ICSLP.
- Litman, D. J. et al. (1998). Evaluating Response Strategies in a Web-Based Spoken Dialogue Agent. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL'98). Montreal, Canada: ACL, , pp. 780-786.
- Lowry, R. (2003). Concepts and Applications of Inferential Statistics. Vassar College. On-line textbook. <http://faculty.vassar.edu/lowry/webtext.html>.
- McInnes, F. et al. (1999). Effects of prompt style on user responses to an automated banking service using word spotting. BT Technical Journal, No 1, Vol 7, pp 160-171.
- Sheeder, T. and Balough, J (2003). Say it Like You Mean It: Priming for Structure in Caller Responses to a Spoken Dialog System. International Journal of Speech Technology, No. 6, pp 103-111.
- Swerts, M. et al. (2000). Corrections in spoken dialogue systems. Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000), Volume II. Beijing, China: ICSLP, pp. 615-618.
- Vanhoucke, W. V. et al. (2001). Effects of Prompt Style when Navigating through Structured Data (paper and presentation). Proceedings of INTERACT 2001, Eighth IFIP TC13 Conference on Human Computer Interaction. Tokyo, Japan: IOS Press, pp.530-536.
- Williams, J. et al. (2003 [1]). Evaluating real callers' reactions to Open and Directed Strategy prompts. AVIOS 2003 Proceedings. San Jose, CA: AVIOS.
- Williams, J. et al. (2003 [2]). Perception, and Task Completion of Open, Menu-based, and Directed Prompts for Call Routing: a Case Study. Eurospeech 2003 Proceedings. Geneva: Eurospeech.
- Witt, S., and Williams, J. (2003). Two Studies of Open vs. Directed Dialog Strategies in Spoken Dialog Systems. Eurospeech 2003 Proceedings. Geneva: Eurospeech.