INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 103–111, 2003

Say it Like You Mean it: Priming for Structure in Caller Responses to a Spoken Dialog System

TONY SHEEDER AND JENNIFER BALOGH
Nuance Communications, VUI Center, 1380 Willow Road, Menlo Park, CA 94025, USA

Abstract. In this paper we report results of a study undertaken to evaluate the initial prompts of 'open prompt' style call-routing applications. Specifically, we examined how placement and phrasing of examples in the initial query affected caller responses and routing success. We looked at the comparative effectiveness of placing examples before and after the initial query and of phrasing these examples such that they promoted either a succinct structure in the form of a keyword or phrase, or a more complex but natural structure in the form of a question or statement. Findings indicate that examples encouraging a more natural structure, when presented prior to the initial query, result in significantly improved routing performance. We discuss this result in the context of using initial prompts to prime for desired structure in caller responses.

Keywords: call routing, call steering, natural language, prompting, dialog strategy

Introduction

Advances in speech recognition have begun to reap improvements in the accuracy, efficiency, and usability of spoken dialog systems, providing access to a wide range of services over the telephone. Leveraging the ubiquity and convenience of the telephone, and in response to the high costs of maintaining call-centers staffed with live agents, enterprises are seeking to provide a wider variety of more complex automated call-routing services.

Traditionally, these systems have employed relatively rigid hierarchical menus of options from which callers choose using touch-tones or highly constrained spoken commands (e.g., "for your account balance, say 'balance,'" or the more egregious "for your account balance press or say 'one' "). There are disadvantages of the menu structures: the process can be time-consuming, and the menu items themselves can be frustratingly opaque. These disadvantages are clear and well documented (Carpenter and Chu-Carroll, 1998). Because of these limitations, systems have been developed that classify callers' goals, and route their calls, in response to a relatively open-ended initial query like "How can I help you?" (Lee et al., 2000; Gorin et al., 1997).

More sophisticated language modeling offers the promise of conversational, natural language interaction in relatively broad and loosely constrained applications like call-routers. In a recent comparison of a touch-tone system and a speech-enabled call-router, more callers were connected to the correct destination with the speech system, and callers overwhelmingly preferred saying their responses to using touch tones (Suhm et al., 2002). Even with these gains, automated task completion rates with speech still vary widely, contingent upon callers' abilities to articulate responses that form a reasonable match to the corpus of data used to train the router. The short evolution of these systems has resulted in a typical strategy whereby callers are issued a fairly broad, open-ended initial query that is followed, if the caller hesitates, by the presentation of options or examples of what the caller can say. Operating under the assumption that, when problems arise, it is most efficacious to proceed from the more general initial query to increasingly specific instructions, these systems usually offer examples in the form of keywords or key phrases and, in the event of repeated difficulties, often fall back to a speech-enabled version of the traditional hierarchical model, presenting a structured taxonomy of commands (Boyce and Gorin, 1996; Brems et al., 1995). However, because of the difficulty

callers have in mapping their task goals to semantically arbitrary commands, it is not clear that this progression is the best strategy.

For the purposes of this study, 'call-routing' applications are defined as spoken language systems designed primarily to correctly direct or 'route' a caller to one of many destinations on the basis of the caller's utterance. Typically this sort of application is appropriate when the number of routing destinations is so large that it precludes presentation as a list or menu, and/or when the destinations are difficult to group into coherent categories, and/or when the routing destinations themselves defy characterization in well-bounded recognition grammars. Owing to the nature and structure of call-routing applications, the initial prompt is freighted with responsibilities and restrictions not found in other, more linear or hierarchical applications (Glass, 1999). Because it is difficult to simply list all of the possible destinations, the initial prompt must, on the one hand, telegraph the range of possible destinations available and, on the other, imply the boundaries of the application (i.e. what is *not* covered).

Some work has been done exploring the effect of prompting on caller satisfaction, for example on anthropomorphism and wording that indicates the system is automated (Boyce, 1999). In this paper, we are interested in determining whether proceeding from a general, natural language interaction paradigm to a more constrained keyword command model, is the most effective strategy, with regard to both user satisfaction and task completion. It is unclear whether the practice of providing examples following the initial query, only after the caller hesitates (i.e. in the form of delayed help), is the most effective approach since the caller is not primed for how to respond. Furthermore, it is our hypothesis that examples presented as keywords actually *diminish* usability, and could hinder task completion, insofar as they require callers to constrain the structure of their responses and to map them onto contrived semantic category labels.

Competing Strategies

While there are a variety of salient questions around call-routing initial prompts, for this study we narrowed the focus to the effects of example phrasing and placement on the caller's ability to provide an utterance that results in correct routing. We examined four different initial prompt strategies by manipulating both the placement and phrasing of examples of what the caller could say in response to the initial query. Note that references to the 'initial query' in this paper signify the portion of the initial system prompt that directly solicits a response from the caller (e.g., "How can I help you?"). We refer to the caller's response to this initial query—which can take the form of a question, command, or statement of a goal—as a 'request.' For placement, we presented the examples either before the initial query (called *Preceding*) or after the initial query if the caller hesitated longer than 2.5 seconds (referred to as *Following*). For phrasing, we used two different strategies: one we call the *Keyword* strategy, which provides the caller with a representative, though necessarily incomplete, list of words or phrases relating to major routing destinations, and the other, dubbed *Natural* (as in 'natural language'), which provides the caller with one or more examples of the utterance structure encouraged by the system.

The efficacy of the competing strategies was tested using an application based on a wireless telephony carrier call-router with upwards of forty routing destinations. We crafted four versions of the initial prompt—one for each phrasing and placement alternative—while all other elements of the application (recognition grammar, routing logic, etc.) remained consistent. A caller who was exposed to the *Keyword Following* version of the prompt, with keyword examples or options given *after* the initial query, heard the following upon first entering the system:

> *Welcome to Clarion[1] Wireless Customer Service. How can I help you?*

Note that speech recognition in this and all other versions of the system are enabled with barge-in, and the caller could interrupt the prompt at any point following the initial greeting. If the caller paused without speaking for more than 2.5 seconds, the following 'delayed help' prompt would be heard:

> *You can ask me about things like 'minutes used,' 'automatic payments,' and 'calling plans.' So, what can I help you with?*

If the caller should pause for another 2.5 seconds, the prompt would continue as follows:

> *For more assistance, just say 'help.'*

A caller hearing the *Natural Following* version of the prompt, with natural language examples given after the

initial query, would encounter the following upon first entering the system:

Welcome to Clarion Wireless Customer Service. How can I help you?

<2.5 second pause>

You can ask me things like 'how many minutes have I used?' and 'I'd like to set up automatic payments.' So, what can I help you with?

<2.5 second pause>

For more assistance, just say 'help.'

Note that only the second, 'example,' segment of the initial prompt differs from the *Keyword* version. In all other respects, save the wording of the 'help' prompt (see below), the applications are identical. As noted earlier, we also created versions of the *Keyword* and *Natural* strategies that differ in the placement of the examples. A caller to the *Keyword Preceding* version of the system, with keyword examples presented *prior* to the initial query, would hear:

Welcome to Clarion Wireless Customer Service. You can ask me about things like 'minutes used,' 'automatic payments,' and 'calling plans.' So, how can I help you with your account?

<2.5 second pause>

For more assistance, just say 'help.'

A caller using the *Natural Preceding* version of the prompt, with natural language examples preceding the initial query, would hear:

Welcome to Clarion Wireless Customer Service. You can ask me things like 'how many minutes have I used?' and 'I'd like to set up automatic payments.' So, how can I help you with your account?

<2.5 second pause>

For more assistance, just say 'help.'

Because we wanted to avoid presenting callers with competing strategies in the event that help was requested, the *Keyword* and *Natural* versions of the test application use alternate phrasings of the 'help' prompt consistent with the overall prompting strategy in that particular version. So, for example, the *Keyword* help prompt is:

I'm here to handle your customer service needs and help you manage your account. Just tell me what you'd like assistance with. You can ask about things like 'account balance,' 'voicemail password,' and 'activation.' So, what can I do for you?

The *Natural* help prompt, on the other hand, is:

I'm here to handle your customer service needs and help you manage your account. Just tell me, in your own words, what you'd like assistance with. For example, you can say things like 'I want to activate my phone' and 'I want to change my voicemail password.' So, what can I do for you?

Method

Participants

Seventy-two subjects participated in the study. All participants—who were either recruited through a professional research firm or were friends and family of Nuance employees—were native English speakers between the ages of twenty and sixty-five. All of the subjects had some experience managing a wireless telephone account (e.g., choosing calling plan, paying the bill, etc.) and none had direct experience with the design or development of speech applications or technology.

Design and Materials

The application was based on a wireless telephony carrier call-router with upwards of forty routing destinations. Four versions of the application could be accessed via a phone line. Each version varied only with regard to the initial and help prompts. All versions used the same SayAnything™ recognition grammar, from Nuance 7.04 software, trained on approximately 20,000 tagged call-center utterances. The experiment was a between-subject design, with eighteen people interacting with each version of the system.

We chose a between-subject design over a within-subject design for several reasons. Previous studies on prompt comparisons have shown that preference scores interact with the ordering of conditions (Dialogues Spotlight[2] technical report). Although counterbalancing is supposed to neutralize these effects, callers still might be primed by prompts in previous conditions and respond differently than if they had been exposed

only to one system. We wanted to keep the results clean from such carryover effects. Also, counterbalancing four conditions would result in a much more complex design. Logistically, callers would have had to do at least one task with four different systems and provide subjective ratings after each one. We wanted to keep the callers' sessions simple, with only one subject questionnaire at the end.

Procedure

The experiment was conducted over the telephone, with the experimenter asking each subject to complete the same three tasks associated with maintaining a wireless telephone account. Table 1 lists the assigned tasks, along with the task descriptions presented to subjects. When describing the tasks, care was taken to avoid language that might typically be used by callers so as to prevent the subjects from parroting the exact phrasing.

The order of the tasks was varied across subjects such that all orders were represented in equal numbers for each version of the system. None of the routes required to successfully complete the tasks were explicitly mentioned in the initial prompts, though one task (Activate Service) was referenced in both versions of the help prompt. Once the experimenter described the first task, he called the application and allowed the caller to interact with it without interruption. When the caller had completed the task (successfully or not), the experimenter came back on the line and presented the next task. At the end of the session, the caller was given a verbal questionnaire pertaining to perceived recognition accuracy, ease of use, simplicity or clarity of instructions, efficiency, and helpfulness. Callers were read a series of statements (e.g., "The system understood what I said"), with each concept presented both positively and negatively (e.g., "The system was easy to use" and "The system was difficult to use"), and asked to respond on a seven-point Likert scale ranging from 'strongly agree' to 'strongly disagree.' The entire session was recorded on audiotape, and call logs were generated so that correct routings and time stamps could be tracked.

Table 1. Assigned tasks with task descriptions presented to subjects.

Task	Task description
Activate service	"You recently bought a new cell phone, or received one as a gift, but it hasn't been turned on. You're calling customer service to find out how to get it working."
Bill reprint	"You've accidentally discarded the latest statement on your wireless telephone account, the one you receive every month in the mail. Since you know you'll need to make a payment soon, you're calling the customer service line to get another one."
Change plan	"When you first got your cell phone, you thought you would use it mainly for emergencies, but you've found that you use it more than you anticipated. You're calling customer service to increase the amount of time you can use the phone every month before incurring extra charges."

System performance was evaluated in terms of task completion (i.e., was the caller's query correctly routed?), task efficiency (how many turns did it take to complete a task?), and task mastery (did the caller's performance improve for successive tasks?). We also looked at the duration of callers' utterances (in both seconds and number of words), rates of system errors (rejection, no speech, etc.), and rates of disfluency (hesitations and repetitions).

Results

Task Completion

To gauge the rates of task completion in the competing systems, the results returned by the recognizer were analyzed for each task performed by each test subject, and were tagged as being in one of four classes—'correct,' 'close,' 'incorrect,' or 'incomplete.' 'Correct' includes those task requests that were routed to the appropriate destination. 'Close' is technically a subset of 'incorrect,' and includes those requests routed to destinations *intermediate* to the 'correct' destination, from which the caller would have a high likelihood of subsequent correct routing. Because the recognition package used in the tests was trained on actual utterances from a deployed system and preserved the routing logic of the parent application, it included a number of destinations that are best characterized as 'disambiguation' states. For example, a caller wishing to receive a copy of her bill might, if the query were ambiguous, be routed to a state where she could choose from amongst several billing-related options (e.g., 'duplicate bill,' 'billing address,' etc.). So, the 'close' class is predicated upon the notion that some 'incorrect' routes are worse than

Table 2. Percent task completion with examples *Following* versus *Preceding* the initial query.

Task completion category (%)	Position of example	
	Following	*Preceding*
Correct	37.04	50.93
Close	18.52	23.15
Incorrect	32.41	22.22
Incomplete	12.04	3.70

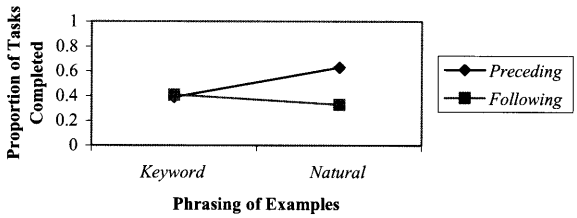

Figure 1. Correct task completion as a function of the phrasing and placement of examples in initial prompts.

others. 'Incorrect' includes those requests that were routed to the wrong destination altogether, and 'incomplete' refers to those that resulted in a transfer to an operator, either at the caller's direct request or because the maximum state error limit was reached.

There is a significant effect of 'correct' routing between *Following* and *Preceding* versions F (1,68) = 6.36, $p = 0.01$, with systems that presented examples *Preceding* the initial query resulting in better routing performance. Also, systems with *Following* examples resulted in more incomplete tasks compared to systems with *Preceding* examples, F (1,68) = 4.32, $p = 0.04$. Table 2 shows the means of average task completion for *Following* versus *Preceding* systems.

Table 3 shows the percentages of task requests that fell into each class for *Keyword* systems versus for *Natural* systems.

There is no statistically significant difference between the *Keyword* and *Natural* systems with regard to correct task completion rates, F (1,68) = 2.29, $p = 0.13$, but there is a significant interaction F (1,68) = 8.17, $p = 0.00$, as shown in Fig. 1.

A post hoc Tukey's HSD test revealed that examples *Preceding* the initial query, phrased in the *Natural* style, resulted in significantly better task completion rates compared to all other conditions, as shown in Table 4.

The accuracy of the call-router reflected the task completion patterns, in that the proportion of utterances correctly understood by the system was higher for *Previous* (0.85) conditions compared to *Following* (0.75). Our focus in this paper was on task completion as opposed to accuracy because some callers, although correctly recognized, did not complete their task because they requested the wrong information. For example, some callers said "minutes used" when they were supposed to increase the minutes for their calling plan. The fact that callers in the *Keyword* conditions did this more often than in the *Natural* conditions (12 percent of all recognized utterances compared to two percent) suggests that some callers repeated back the keyword examples even though the keyword was not a correct representation of the task.

It should be noted that, in our tests, the examples were heard in the *Following* versions of the system less than 13% of the time (see Table 5), on average, because the subjects interrupted the system prior to that portion of the prompt being played, while the examples were heard in 100% of task attempts by subjects using the *Preceding* systems.

In other words, callers to *Following* versions of the system received no instruction with regard to how a request should be phrased in over 85% of their task attempts.

Table 3. Percent task completion with initial prompts using *Keyword* examples versus *Natural* examples.

Task completion category (%)	Phrasing of example	
	Keyword	*Natural*
Correct	39.81	48.15
Close	22.22	19.44
Incorrect	29.63	25.00
Incomplete	8.33	7.41

Table 4. Mean differences of correct task completion of *Natural/Preceding* examples compared to all other conditions and the corresponding p value from Tukey's HSD test for each comparison (with 3 groups, df = 15).

Compared conditions	Mean difference	p value
Natural/Preceding vs. *Keyword/Preceding*	22.22	0.03
Natural/Preceding vs. *Keyword/Following*	24.07	0.01
Natural/Preceding vs. *Natural/Following*	29.63	0.00

Table 5. Percent of tasks in which examples were heard for each type of initial prompt.

	Phrasing and position of examples			
	Keyword/ Following	*Keyword/ Preceding*	*Natural/ Following*	*Natural/ Preceding*
Examples heard (% tasks)	16.67	100.00	9.26	100.00

Table 6. Average turns per task for each type of initial prompt.

	Phrasing and position of examples			
Average turns per task type	*Keyword/ Following*	*Keyword/ Preceding*	*Natural/ Following*	*Natural/ Preceding*
Correct	1.35	1.31	1.41	1.25
Incorrect	1.91	1.67	1.52	1.10
All tasks	1.59	1.43	1.46	1.22

Task Efficiency

Task efficiency was judged by measuring the number of "turns" (where a "turn" is one interchange between the system and the user) required to initiate a routing decision. Table 6 shows the average number of task turns for each system, along with the average number of turns for a 'correct' or an 'incorrect' routing decision.

The data reveal no appreciable benefit, in terms of efficiency, from one prompt version over any other.

Task Mastery

'Task mastery'—whether a particular subject was able to increase task efficiency from one task to the next-was evaluated by measuring the relative decrease in the number of turns required to effect 'correct' routing decisions for successive tasks. This measurement relies solely on 'correct' designations insofar as reducing the number of turns to achieve an inappropriate result is a Pyrrhic success, at best. Table 7 shows the average number of turns required to achieve 'correct' routing, broken down by system and by task number (i.e. whether a given task was the subject's first, second, or third assigned task).

Intuitively, the average number of task turns needed to arrive at a 'correct' destination could be predicted to decrease for successive tasks as individual subjects become more familiar with, and adept at using, a given system (the progression seen for the *Keyword Preceding* system in Table 7). In fact, we found no significant increase in task efficiency (even in the case of the *Keyword Preceding* system).

Table 7. Average turns per task, for each type of initial prompt (*Keyword/Natural, Preceding/Following*), by task number.

	Phrasing and position of examples			
Average turns per task number	*Keyword/ Following*	*Keyword/ Preceding*	*Natural/ Following*	*Natural/ Preceding*
1st task	1.25	1.45	1.50	1.27
2nd task	1.30	1.31	1.27	1.08
3rd task	1.56	1.17	1.50	1.40

Utterance Characteristics

In addition to results focusing on task completion and efficiency, we looked at characteristics of callers' responses to the systems. Specifically, we examined utterance length (both duration in seconds and number of words), the relative rates at which errors occurred, and at the relative rates at which utterances included disfluencies. Utterance length, in terms of clock time, revealed no significant effects, nor did utterance length measured in number of words. Table 8 shows the average scores of utterance length for all four systems.

As noted earlier, subjects using the *Following* versions of the system received no prompting whatsoever, with regards to formulating their requests, in over 85% of task attempts. It should be pointed out that there is no significant correlation between utterance length and successful task completion. It is also interesting to note that utterance length does not reflect the number of words in the examples presented by the competing strategies. In the system prompt, the average length of the *Keyword* examples is 2 words, while the *Natural*

Table 8. Average utterance length in response to initial prompts with different phrasing and position of examples.

	Phrasing and position of examples			
	Keyword/ Following	*Keyword/ Preceding*	*Natural/ Following*	*Natural/ Preceding*
Average duration (s)	3.53	3.27	3.67	3.51
Average no. of words	8.33	7.74	9.13	9.57

examples average 6.5 words. So, while short examples seem to encourage circumspection, and longer example phrasings give license for longer responses, the affective capacity of example phrasing to impact utterance length is sharply limited.

In addition to utterance length, we looked at the relative rates at which utterances resulted in system errors. We examined *rejection* (REJ) errors—instances when the utterance returned a recognition result below a default confidence threshold—*speech too early* (STE)—cases when the system detected speech before it was expected, usually because the caller was continuing to speak after a rejection had occurred—*no speech* (NSP) errors—pauses during which speech was expected, but not detected—and "null" recognition results—cases where the recognizer returned a result that did not match a valid routing destination.[3]

When looking at average errors per subject, we find that utterances in response to *Preceding* examples ($M = 1.08$) produce significantly fewer system errors than *Following* examples ($M = 2.22$), F (1,68) $= 5.16$, $p = 0.03$. There is no effect of phrasing and no significant interaction.

It is worth noting that the only significant difference in overall error rate for *Preceding* versus *Following* conditions occurs with regard to "null" recognition results, F (1,68) $= 6.05$, $p = 0.02$. Table 9 shows the average number of errors by type.

As with error rates, we looked at the average number of utterances per subject that included instances of disfluency—repetition of words or phrases, and hesitations, whether vocalized ("uh" and "um") or silent. Table 10 shows the relative rates of disfluency, as a percentage of all utterances, by prompt type. Similar to error rates, utterances in response to *Preceding* examples ($M = 0.18$) resulted in fewer disfluencies than responses to *Following* examples ($M = 0.31$), although the results did not quite reach significance: F (1,68) $= 3.85$, $p = 0.05$.

Table 9. Average number of errors, by error type per subject, for systems with examples *Preceding* and *Following* the initial query.

Average number of errors by type	Position of example	
	Following	*Preceding*
Null	1.53	0.64
REJ	0.28	0.33
STE	0.30	0.06
NSP	0.11	0.03

Table 10. Percentage of disfluent utterances in response to initial prompts with different phrasing and position of examples.

	Phrasing and position of examples			
	Keyword/ Following	*Keyword/ Preceding*	*Natural/ Following*	*Natural/ Preceding*
Percentage of disfluent utterances	30.13	22.62	30.53	15.94

Like errors, disfluency seems to be another indicator of the advantages of priming the caller with examples indicative of the *form* of request most likely to result in successful task completion.

User Experience

Beyond examining quantitative metrics around task completion, efficiency, mastery, and utterance characteristics, we were interested in determining which, if any, strategy or strategies provided for a better user experience. This we attempted to infer from subjective responses (on a seven-point Likert scale, where 1 was a low score and 7 a high score) to a series of questions pertaining to perceived recognition accuracy, ease of use, simplicity or clarity of instructions, efficiency, and helpfulness. Table 11 shows the scores in the individual question categories, along with the combined average score, for *Following* and *Preceding* systems.

Callers perceived the systems with *Preceding* examples to be more accurate than *Following* versions, as might be expected given task completion rates, F (1,68) $= 10.12$, $p = 0.00$. With all questions combined, *Preceding* examples are rated significantly better than *Following* examples, F (1,68) $= 4.16$, $p = 0.04$. Table 12 shows the breakdown of scores for *Keyword* systems versus *Natural* systems.

Table 11. Subjective usability scores (7-point scale) by category for *Following* versus *Preceding* versions of the initial prompt.

Score by category (7-point scale)	Position of example	
	Following	*Preceding*
Accuracy	4.19	5.35
Easy to use	5.47	5.92
Simple instructions	5.67	5.58
Efficient	5.49	5.78
Helpful	5.25	5.79
Combined	5.21	5.68

Table 12. Subjective usability scores (7-point scale) by category for *Keyword* versus *Natural* versions of the initial prompt.

Score by category (7-point scale)	Position of example	
	Keyword	*Natural*
Accuracy	4.72	4.82
Easy to use	5.42	5.97
Simple instructions	5.44	5.81
Efficient	5.58	5.68
Helpful	5.56	5.49
Combined	5.34	5.55

Here we see a trend that systems with *Natural* examples are rated as easier to use compared to systems with *Keyword* examples, $F\ (1,68) = 3.66$, $p = 0.06$.

Conclusions

Our findings demonstrate that correct call routing is significantly higher, and incomplete routing significantly lower, when callers are exposed to an initial prompt with examples that *precede* the initial query. We also found that, when examples are provided prior to the initial query (i.e. routinely heard), the resulting caller utterances provoke significantly fewer system errors, with a stark reduction in "null" recognition results, and instances of disfluency. *Preceding* examples also elicit significantly better subjective usability ratings. Moreover, when the examples heard prior to the initial query are phrased *naturally*, as opposed to taking the form of *keywords*, correct routing is significantly increased, by 50–90% over alternative versions of the system.

Also of note is the discovery that utterance duration within the ranges produced by our tests, whether measured in seconds or in number of words, has no significant relationship to successful task completion. Perhaps even more striking is that we found no connection between the length, in words, of the examples provided and the utterances made in response. So, while *Keyword* examples produced shorter utterances, on average, than *Natural* examples (7.74 versus 9.57 words in the *Preceding* systems, where examples were heard in 100% of task attempts), these shorter utterances were still markedly longer than would be expected if callers had internalized a *Keyword* command style of interaction. In fact, this discrepancy between utterance length and example length suggests that callers have a more or less innate tendency to couch their goals in fully structured requests.

With regard to both task completion and the characteristics or quality of caller utterances, our findings indicate the advantage of linguistically "priming" callers by presenting examples up front instead of offering them only when callers are struggling. Given that preceding *Natural* examples resulted in better task completion, we conclude that the form that these examples take is also important. It should be recalled that in no instance were the tasks assigned to the test subjects mentioned in the examples provided. Therefore, absent a direct reference to a caller's goal, the most advantage is gained by indicating to the caller *how*, rather than *what*, to speak. From these results we conclude that call-routing applications generally will benefit from rethinking the currently employed strategies. Namely, that providing callers with examples that they are likely to hear (i.e., *prior* to the initial solicitation for a request) and focusing on imparting a sense of the expected *form* of the request, as opposed to an arbitrary semantic category label, should be the normative prompting strategy.

Future Work

Since no qualitative difference in callers' responses to *Natural* examples, as compared with *Keywords,* can be inferred from utterance length, in further research we plan to analyze the callers' utterances with regard to linguistic structure to understand qualitatively why responses to *Natural* prompts preceding the initial query resulted in better task completion.

Also, while our findings offer compelling evidence in support of offering examples that promote an understanding of the utterance structure expected by the system, there remain serious questions requiring further study. It is unclear, for example, how promotion of a natural utterance structure could affect task mastery for frequent callers to a system. It is conceivable that while correct completion rates for naïve callers is improved by this strategy (effectively the result we found), overall efficiency gains in successful task completion could be hindered because callers never learn to employ highly accurate verbal shortcuts. Further experiments may also be warranted to determine if learning to use an application on the basis of a 'natural' interaction style might hinder callers' abilities to conform to more highly constrained recognition grammars of the type that could be used in sub-applications of a larger call-router (e.g., 'yes/no' confirmations, sub-menus, digit or character string capture, etc.).

Notes

1. 'Clarion Wireless' is the name of a fictional wireless telephony carrier. Test subjects were made aware that this was a test application (as opposed to a deployed system). The branding message was included to provide a more realistic context for the subjects' interactions.
2. The effects of speaker state (tone of voice) and speaker style (fast track) in dialogue prompts. Dialogues Spotlight, University of Edinburgh. http://spotlight.ccir.ed.ac.uk
3. Because the test applications used a statistically modeled SayAnything™ recognition grammar, the recognizer was capable of returning a recognition result that did not match a valid routing destination.

References

Boyce, S.J. (1999). Spoken natural language dialogue systems: User interface issues for the future. In D. Gardner-Bonneau (Ed.), *Human Factors and Voice Interactive Systems*, Norwell, MA: Kluwer Academic Publishers, pp. 37–61.

Boyce, S.J. and Gorin, A.L. (1996). User interface issues for natural spoken dialog systems. *Proceedings of ISSD 96*.

Brems, D.J., Rabin, M.D., and Waggett, J.L. (1995). Using natural language conventions in the user interface design of automatic speech recognition systems. *Human Factors*, *37*(2):265–282.

Carpenter, B. and Chu-Carroll, J. (1998). Natural language call routing: A robust self-organizing approach. *ICSLP 98*, Sydney.

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

Gorin, A.L., Riccardi, G., and Wright, J.H. (1997). How may I help you? [natural spoken dialog system for automated services]. *Speech Communication*, *23*(1/2):113–127.

Lee, C.H., Carpenter, B., Chou, W., Chu-Carroll, J., Reichl, W., Saad, A., and Zhou, Q. On natural language call routing. *Speech Communications*, *31*:309–320.

Suhm, B., Bers, J., McCarthy, D., Freeman, B., Getty, D., Godfrey, K., and Peterson, P. (2002). A comparative study of speech in the call center: Natural language call routing vs. touch-tone menus. *Proceedings of the SIGCHI conference on human factors in computing systems: Changing our world, changing ourselves*. Minneapolis, Minnesota, USA.