WHAT WILL PEOPLE SAY?
SPEECH SYSTEM DESIGN AND LANGUAGE/CULTURAL DIFFERENCES

Kristin Precoda[1] and Robert J. Podesva[1,2]

[1]SRI International, Menlo Park, CA
and [2]Stanford University, Dept. of Linguistics, Stanford, CA

ABSTRACT

This paper evaluates the effectiveness of three speech system design strategies in Pashto, a little-studied language of Afghanistan and Pakistan, drawing comparisons with English where possible. The strategies discussed are using (1) prompts at the ends of questions to constrain user responses, (2) specific lexical items in system prompts to encourage user echoing, and (3) introspection as a method for building recognition grammars. It was found that Pashto speakers were strikingly less influenced by system utterances than American English speakers were, and that introspection grammars, even though constructed by a speaker with unusually broad dialect exposure, had both many too many, and a few too few, choices. We conclude that the effectiveness of these and perhaps other design strategies, many of which derive from work on English, may vary along linguistic or cultural lines, and new strategies may need to be explored for languages where these do not work well.

1. INTRODUCTION

In spoken language systems, the performance of automatic speech recognition engines rests in part on accurately predicting what people will say, or, alternately, on constraining users' responses to those the system will be able to recognize. Designers of such systems have developed a body of knowledge about procedures for building recognition grammars and designing system utterances to maximize recognition performance and acceptability to the user. Some of these procedures involve carefully wording system utterances to limit likely user responses; constructing initial recognition grammars based on reasonable guesses, or, if available, on previously observed data; and modifying the grammars during or after trial usage to better match collected data.

Most of these design strategies derive from work with a few major languages and applications of commercial interest. However, in recent years interest in other, less well-studied languages has increased, because of the need to provide humanitarian assistance to speakers of some of those languages, disseminate information to and collect information from previously unreachable populations, establish better communication in areas of military interest, and so on. In these languages, the design strategies developed for major languages may or may not be equally effective, and they may well need to be adapted to linguistic and cultural circumstances and to the availability of the needed language resources.

This paper will describe the effectiveness of several system design strategies in Pashto, a language spoken by approximately 19 million people [7] in Afghanistan and Pakistan and in émigré communities elsewhere, and where possible will compare the strategies' effectiveness in Pashto and in English. The strategies to be discussed are (1) using prompts at the ends of questions to constrain user responses, (2) using specific words in system prompts to encourage lexical echoing by the user, and (3) using introspection as a method for building a very simple recognition grammar.

2. SYSTEM GOALS, LINGUISTIC CONTEXT, AND INTERACTION WITH DESIGN

The work described in this paper was performed in the course of a U.S. government-funded project on building a limited-domain, phrase-based translation system [5] between English and Afghani Pashto. The phrases to be translated pertain to refugee registration and processing, public health, basic medical interviews, social niceties, and so on. The system is intended for use by English-speaking aid workers gathering information from and assisting Pashto speakers, primarily refugees. These circumstances influence the goals of the system. First, while the English speaker is likely to use the translation system repeatedly and become experienced in working with it, each encounter is assumed to be the Pashto speaker's first and last interaction with the system. In addition, in many refugee situations getting required information extremely rapidly is a necessity. The Pashto speaker must therefore be able to communicate successfully with no prior training with the system and with as few time-consuming errors as possible. A second consideration is that user acceptance and satisfaction, particularly on the part of the Pashto speaker, are defined very differently than is often the case in many commercial speech applications, because of the different nature and availability of alternatives to using the spoken translation system. Finally, literacy in Pashto is very low (estimated at 5% among Afghani Pashtuns [10]), providing a compelling case for spoken communication.

Like most languages of the world, Pashto has been the subject of little work of any kind and there is little supporting material such as dictionaries, linguistic descriptions, or textual or audio corpora. In addition, cultural and physical conditions have given rise to many dialects with substantial variation. This paucity of resources and complex dialect situation directly impact on system design methods in a variety of ways. The effects of wording of system utterances — in our case, prerecorded phrase translations — on user responses can initially only be assumed to be similar to effects observed in other languages, as there is no language-specific or culture-specific data. Initial recognition grammars must be based purely

on introspection, since there is no existing corpus to build from, but at the same time the dialect situation probably renders introspection by one or only a few speakers less effective than it might be in a more homogeneous or standardized language. As the language is poorly documented from any point of view, observed data is all the more important, even for answering the most basic questions of syntax or lexicon; yet the current political and economic situation of Afghanistan renders large field trials and iterative development of the system nearly impossible. There is also no one who combines language competence with familiarity with speech and language technology, and the impact of this knowledge gap is an important one.

3.   SOME PRIOR RESULTS

Empirical results have been reported in several studies concerning the effect of explicit instructions in a system utterance, and the percent of user responses which contain a key word to be recognized, either alone or with other words. The results cited here all derive from English-language experiments.

In a study on automatic telecommunications services, Brems et al. [3] examined the elicitation of two kinds of desired information, one yes/no responses and the other a choice from a short menu of options. Forty-eight users used a test speech recognition system to make or receive nine calls each. For yes/no questions, presenting only the question with no prompting as to expected answers obtained responses containing a key word and no more than two extra words 97.2% of the time; other responses either did not contain a key word or included more than a sentence of additional speech. For menu questions, presenting a similar open-ended question without prompts resulted in 61.3% of responses containing a key word and no more than two extra words, or 98.7% containing a key word and no more than one sentence. When a yes/no question included a prompt such as "Please say yes, no, or operator, now", 97.3% of the responses were a key word with no more than two extra words, and 100% contained a key word and no more than one sentence. When a menu question included a similar prompt (e.g. "Please say collect, calling card, third number, person to person, or operator, now"), 98.7% and 100% of the responses contained a key word and no more than two words or one full sentence, respectively.

Basson et al. [1] experimented with several systems including an automated customer service center application presenting queries such as "Please select one of the following: sales, service, or billing information". In 9000 calls, "approximately 84% of the compliant key word responses were spoken in isolation", though they do not note how many responses did not contain a key word. In 174,000 calls to a directory assistance application, 14% of users gave a complete listing including name and address in response to "What city, please? <beep>", while none gave a full listing to "After the beep, please say just the name of the city".

Marcus et al. [9] found that "Are you returning to the same location?" in a rental car reservation system elicited "almost 100% simple affirmative/negative responses".

Turning to another phenomenon of system utterance wording, it has long been noted that the choice of lexical item in an utterance may influence the respondent to repeat (a form of) the same word. For example, using data collected in the DARPA Communicator project evaluation in 2000 [11], we looked at several words and counted the number of times they appeared in a user utterance immediately following a system utterance using the same or a semantically similar word [2]. The results, given in Table 1, show that words heard in system utterances strongly influenced users. For example, users were unlikely to say "depart" after hearing "leave" in a system utterance (9 times out of 100), and much more likely to say "depart" when it appeared in the system utterance as well (49/110). Similarly, users responded to system utterances containing the word "right" by echoing "right" half the time, even though users never said "right" when the system utterance contained "correct". Finally, speakers were most influenced by system utterances with regard to "fly" and "go", for which speakers repeated the word heard in the system utterance 92% and 87.5% of the time, respectively.

Table 1. Use of the same or a different word in system utterances and user responses. Shading indicates cells where the question and response share a word.

	System	
User	"leave"	"depart"
"leave"	91	61
"depart"	9	49
	"correct"	"right"
"correct"	20	8
"right"	0	8
	"fly"	"go"
"fly"	92	5
"go"	8	35

4.   DATA COLLECTION AND METHODS

To look at the effect of the wording of system utterances on user responses, we recorded spoken responses to prerecorded utterances in both Pashto and English. The same set of phrases and prompts were presented in both languages; speakers heard the phrases in their primary language. Speakers were told that they would hear a series of phrases or questions pertaining to refugee or medical situations, and that they should answer each while trying to imagine appropriate scenarios. It was emphasized that whether or not answers were truthful was irrelevant. Pashto speakers also responded to a number of utterances to which only an acknowledgment was expected, for example, "I'm a member of the refugee assistance team". Responses to these utterances are not studied here.

Speakers sat at a computer and heard the utterances played to them. After each utterance, the speaker pressed a button to record her or his response using a head-mounted microphone. Speakers were free to replay the utterance if desired and to take breaks as needed. The set of phrases was presented starting from various different points for different speakers. However, each English speaker heard the phrases in exactly the same order as the Pashto speaker to whom s/he was matched.

The Pashto speakers were recorded first, and then English speakers matched in gender and approximate age were recruited and recorded. The Pashtuns' educational levels and exposure to technology are probably quite unrepresentative of Pashtuns in Afghanistan, but we expect that their linguistic behavior should be at least as similar to Americans' as that of Pashtuns in

Table 2. Speaker information

Pashto speakers			Number of matched questions answered	Matched American English speakers	
Gender	Approximate age	Years in U.S.		Gender	Approximate age
m	55-60	21	120	m	50s
m	65	17	661	m	66
m	24	3.5	383	m	25
m	55-60	5	295	m	60s
f	26	7	338	f	22
m	45	10	350	m	40
f	35	17	67	f	30
m	25	2	324	m	25

Afghanistan would be. No speaker, Pashtun or American, had any familiarity with speech and language technology other than what they might have encountered during daily life. Information on the speakers is given in Table 2.

Many of the prerecorded utterances ended with a prompt, explicitly indicating the form of response desired. Several prompts corresponding to different kinds of expected answer were used. The English versions of the prompts whose responses will be analyzed were the following:
- Say just yes or no
- Say just yes, no, or a little
- Say just yes, no, or I'm not sure (literal translation of Pashto analog: "I'm not informed")
- Say just the number
- Say just the number of (noun)

The phenomenon wherein a speaker repeats specific lexical items heard in a system utterance, which we term "echoing", was examined through words for two concepts, "person/people" and "times" (e.g. times per day). There are several Pashto words for each of these concepts, in some cases differing by dialect or shades of meaning. Several alternative words were included in system utterances; each such utterance contained either one or two of the alternatives.

Initial recognition grammars were built by introspection for possible responses to several question types; grammars for "yes" and "no" equivalents will be discussed here. The person asked to introspect was a native speaker of Pashto who had lived in Afghanistan and Pakistan until early adulthood, and spent the last few years in the U.S., speaking more Pashto than English. This speaker had had quite considerable and unusual exposure to Pashtuns from many areas of Afghanistan and Pakistan, through several years' affiliation with a medical clinic in a refugee camp. The speaker was given some sample yes/no questions, and asked to make a list of ways that people might answer those questions with the meaning of "yes" or "no". The sample questions included questions of fact ("Do you have children?") and questions asking for agreement or permission ("Is it okay if I examine you?").

Data for comparison to the introspection grammars was collected from the recordings described above and from a corpus of about 7000 additional recordings. This additional corpus was much less structured and included responses to questions presented in English or Pashto, as well as translations and unscripted, minimally prompted speech. To evaluate the match of the recorded data with the introspection grammars, only the key words in a recorded response were considered. To give an English example, if the response to "Are you hungry?" were "Yes, I'm hungry", only the "yes" would be counted. If the response were "I'm very hungry", this would be considered to not include a key word.

5. RESULTS

5.1. Effect of prompts following questions

Figure 1 summarizes types of responses to questions requesting a number, in Pashto and English. Questions included either no prompt, or a prompt of the form "Say just the number" or "Say just the number of (noun)". As there was little difference in responses to the two prompt types, their results have been pooled. Responses which included a number were categorized into four types: number alone; number plus a noun; number in an adpositional (circumpositional, prepositional, or postpositional) phrase in Pashto or prepositional phrase in English; or number in another type of phrase.

Figure 1. Counts for different types of responses to questions requesting a number, for Pashto (top) and English (bottom), with a prompt (left) or without (right).
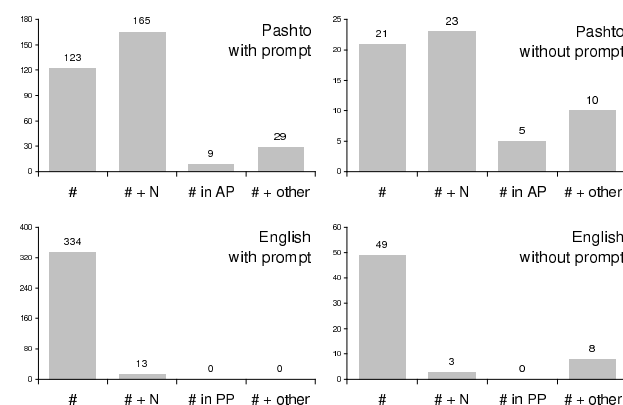

Figure 2 shows how responses differed according to whether or not a yes/no question contained a prompt. Questions included either no prompt or one of the following: "Say just yes or no", "Say just yes, no, or a little", or "Say just yes, no, or I'm not sure". As with number questions, responses to the three different prompts behaved quite similarly. Three kinds of responses are tabulated: isolated key word or phrase; "yes" or "no" (i.e. the specific words used in the prompt) with other words; or phrases without a key word or phrase. For "yes" and

"no", only the forms used in the prompt are counted as key words. For "I'm not sure" in Pashto, different word orders and pro-drop forms are counted as key phrases.

Figure 2. Counts for different types of responses to yes/no questions, for Pashto (top) and English (bottom), with a prompt (left) or without (right).
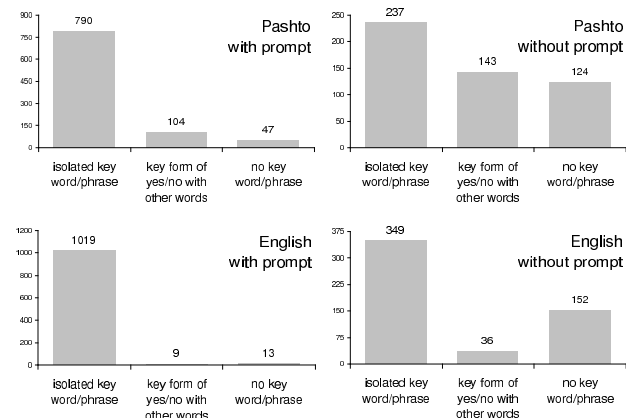
5.2. Echoing of lexical items in questions or prompts

Table 3 shows results for user responses when the previous system utterance included one or two words for "person" or "people"; cells where the question and response shared a word are shaded. There are some meaning and usage differences between these words, as follows. *nafar* means "person" or "people" if plural. *khalk* is a mass noun meaning "people". *kas* means "person/people" and may be used more or less according to dialect. *ghattAn* means "adults", is somewhat awkward, and was used to translate "adults" in the original English sentences only because no better translation could be found. *dAna*, originally from Farsi, means "unit" or "piece". Some speakers can use it to refer to humans while others use it only for inanimates; this usage appears to be associated with dialect.

Table 4 shows the analogous results for "times". One speaker we consulted indicated that *dzal*, *daf'a*, and *wAr* (in one of its senses) are entirely synonymous.

5.3. Introspection grammars and observed data

The Pashto grammars for affirmative and negative responses built from introspection contained, respectively, 95 and 41 different possible phrases or words. Table 5 shows keywords in responses by Pashto speakers to yes/no questions, with the number of times each was used. This data included 664 affirmative responses, expressed in ten different words or phrases. Of these ten, only seven were found in the introspection grammar. The same corpus included 633 negative responses, expressed in six different ways. The introspection grammar contained only one of these six, although the second most frequent one was probably omitted only by accident.

6. DISCUSSION

6.1. Effect of prompts following questions

The results reported in §5.1 indicate that, overall, including prompts in system utterances does shorten responses. For questions requesting a number response, after a prompt speakers of both languages were less likely to embed numbers in longer phrases. For yes/no questions as well, the addition of a prompt resulted in more isolated key word/phrase responses.

The degree to which user responses were constrained to the form requested in the prompt, however, was language-specific. For number responses, English speakers showed an overwhelming preference to respond with a number alone, even without prompting, and a prompt further increased the likelihood that they would say a number in isolation. Though Pashto speakers often responded with a number alone, with or without prompting, they more commonly responded to number questions with a number plus a noun, even when the system told them to say just a number. The yes/no responses pattern similarly. Although English speakers frequently failed to say just "yes" or "no" when unprompted (in contrast to the results reported in [9]), they failed to comply only 2% of the time when a prompt was included, similar to figures reported in [3]. Pashto speakers, like English speakers, frequently responded to yes/no questions without using key words or phrases, though to a greater extent. Although the addition of prompts led to a dramatic increase in responses of isolated key words or phrases (from 47% to 84%), a substantial number of yes/no responses still differed from what the prompt requested.

A post-hoc comparison of the proportions of isolated key word/phrase responses in Pashto and English can be performed, though statistical conclusions should be drawn only with care. In the case of responses to questions which included prompts requesting a number, a 95% confidence interval for the proportion of isolated number responses out of all responses including a number is (.326, .431) for Pashto, as compared with (.937, .978) for English. For responses to yes/no questions which included prompts, 95% confidence intervals for the

Table 3. Counts for word used for "person/people" in system utterance versus speaker response. Grammatical variants by case or number are grouped into the citation form. Shading indicates cells where the question and response share a word.

Word in speaker response	Word(s) in system utterance					
	nafar	nafar + khalk	khalk	kas + khalk	ghattAn	ghattAn + dAna
نفر nafar	12	2	2	2	7	3
خلك khalk			2			
کس kas	14		2	2	8	5
غټان ghattAn						
دانه dAna					1	2

Table 4. Counts for word used for "time/times" in system utterance versus speaker response. Shading indicates cells where the question and response share a word.

Speaker		System	
		dzal	dzal + daf'a
څَل	dzal	12	10
دفعه	daf'a	2	2
وار	wAr	7	3

Table 5. Usage of affirmative and negative forms in observed data. Responses also found in the introspection grammars are indicated with shaded cells.

Observed affirmative responses		Times used
هو	ho	591
بلي هو	bale ho	46
آو	Aw	13
بلي	bale	6
نه	xa	2
نه دی	x@ day	2
هوکي	hoke	2
ډير نه	dder x@	1
هو ولي نه	ho wale na	1
ولي نه	wale na	1
Observed negative responses		**Times used**
نه	na	578
یا	ya'a	51
نه خیر	na khayr	1
نه ډه سمه	n@ da sama	1
نه بیخي نه	na bekhi na	1
بیخي نه	bekhi na	1

proportion of times the response was an isolated key word/phrase, out of all responses described in Figure 2, are (.815, .862) for Pashto and (.968, .986) for English. The large regions of nonoverlap between the Pashto and English confidence intervals indicate there is little likelihood that the true proportions of isolated responses after prompts are the same for Pashto and English, and thus defy a reasonable a priori expectation.

These patterns suggest that Pashto speakers were much less influenced by prompts than English speakers. There are many explanations for why this might be the case, and we can offer only a little speculation here.

It is possible that Pashto and English speakers differ in the relative importance they grant to clarity on one hand and conciseness on the other. The results suggest that among the

English speakers participating in this study, conciseness is the more important factor, since even when not prompted to say just a number, English speakers most often responded with a number in isolation. Pashto speakers might have used longer responses to ensure that their utterances were clearly understood, indicating a preference for clarity. Stated in terms of the maxims of Grice's Cooperative Principle [6], the Pashto speakers may have relaxed the requirements of Quantity (say neither less nor more than is necessary) in favor of those of Manner (be clear), whereas the English speakers may have been influenced more heavily by Quantity.

Politeness considerations could also be playing a role, as terse responses, though efficient, can also be interpreted as rude in some circumstances. Though the Pashtuns and Americans were matched as closely as possible, and no one was particularly familiar with speech applications, the English speakers probably had greater familiarity with technology, and perhaps along with it held the belief that when interacting with computers efficiency can outweigh politeness.

6.2. Echoing of lexical items in questions or prompts

The results in Table 3 and Table 4 provide no evidence for a tendency by Pashto speakers to repeat words heard in system utterances, for these semantic items. No matter what word or words for "people" were heard in the system utterances, speakers responded mostly with forms of *nafar* or *kas*, in roughly similar proportions for all system word choices. The Pashto speakers also appeared not to have been influenced by the word they heard for "times"; their word choice is fairly similar independent of what word the system used. The number of responses tabulated is too small to draw firm conclusions, but they fail to show support for the expected pattern of echoing. These results contrast sharply with those reviewed in §3 for English, in which speakers were influenced by the words used in system utterances for all three word pairs under analysis. A possible interpretation of that data is that English speakers have an a priori preference for "correct" and "leave" in this task, but can be swayed by the system. No clear preference between "fly" and "go" is shown in this data, only a moderate willingness to reflect what the system said.

The Pashto results are surprising to us. But again, the results only fail to show echoing for these sets of words. It is possible that echoing occurs for other words, or that the same social goals are served in some other way, or that our expectations were simply inappropriate.

It is also true that failure to show echoing is consistent with the results on the effect of prompting: overall, Pashto speakers appeared to be strikingly less influenced by what the system said than American English speakers are. We can only speculate as to why.

6.3. Introspection grammars and observed data

The responses to yes/no questions display a poor match between predicted possible and actual responses. The grammars built from introspection were both too inclusive (with 128 of 136 predicted responses unattested in the collected data) and not inclusive enough (as several of the small number of observed responses did not appear in the introspection grammars). We do

not believe these discrepancies between the predicted and observed responses show any inadequacy on the part of the Pashto language consultant; we feel confident of that speaker's extensive exposure to other regional forms, heightened attention to language through working with us, and diligence. Rather, we suspect that other speakers would not produce much better results. Even with the best of experience and intentions, the introspection grammars have both many too many, and a few too few, choices, and in testing recognition performance, these grammars produced clearly suboptimal results and many confusions.

It is possible that speakers could guess likelihoods of various responses, however crudely, and these estimates could be used to limit introspection grammar size. Whether the estimates yielded by such a task would be meaningful is a question.

Another strategy for obtaining grammars in the absence of sufficient collected data would be to ask additional speakers to complete the same introspection task and examine the similarities and differences between speakers. Determining how to interpret such similarities and differences would be an interesting avenue for future research, as the best tradeoff between inclusiveness and parsimony is not obvious.

7.   CONCLUSIONS

For this application to be as successful as possible, we may need to find alternate or additional strategies to constrain Pashtuns' responses enough to maximize recognition performance, especially as the system must work with little or no training of the Pashto speaker. In general, speech systems in new languages may require exploration to find effective strategies for each language. Some possible design strategies affect the system as a whole and may call for additional kinds of system information to be tracked. Examples of system-level strategies are to explicitly instruct a speaker to give short answers if an unexpectedly long response is detected, to start an interaction with an explanation of system limitations or with instructions on how to respond, and to provide clear prompts after low-confidence recognition results. Other strategies affect the design of the system utterances themselves. One example is to collate and expand upon utterance designs described in the literature, pilot test them, and then choose the most effective and construct recognition grammars accordingly. It has also been suggested that longer system utterances may elicit longer responses [4],[8], so short or urgent-sounding utterances could be tested.

Our results show that the effectiveness of the design strategies examined here, and perhaps of others as well, appears to vary for different languages or cultures. New strategies may therefore need to be devised for new languages and cultures.

8.   ACKNOWLEDGMENTS

9.   REFERENCES

[1] Basson, S., S. Springer, C. Fong, H. Leung, E. Man, M. Olson, J. Pitrelli, R. Singh, and S. Wong, "User Participation and Compliance in Speech Automated Telecommunications Applications," *Proceedings of ICSLP '96*, Philadelphia, PA, pp. 1680-1683, Oct. 1996.

[2] Bratt, E.O., C. Culy, H. Bratt, K. Precoda, D. Brown, and C. Richey, "SRI Communicator: Toward More Natural Spoken Output," presented at the DARPA Communicator PI meeting, Philadelphia, PA, Sept. 2000.

[3] Brems, D.J., M.D. Rabin, and J.L. Waggett, "Using Natural Language Conventions in the User Interface Design of Automatic Speech Recognition Systems," *Human Factors* 37(2), pp. 265-282, 1995.

[4] Dobroth, K., "Beyond Natural: Adding Appeal to Speech Recognition Conversations?", *Speech Technology Magazine* 5(2),   www.speechtechmag.com/issues/5_2/cover/191-1.html, 2000.

[5] Franco, H., J. Zheng, K. Precoda, F. Cesari, V. Abrash, D. Vergyri, A. Venkataraman, H. Bratt, C. Richey, and A. Sarich, "Development of Phrase Translation Systems for Handheld Computers: From Concept to Field," , *Proceedings of EuroSpeech '03*, Geneva, Switzerland, pp. 373-376, Sept. 2003.

[6] Grice, H.P., "Logic and Conversation," in P. Cole and J. Morgan, editors, *Speech Acts*, Academic Press, New York, pp. 41-58, 1975.

[7] Grimes, B.F., editor, *Ethnologue: Languages of the World*, 14th edition, SIL International, Dallas, TX, 2000.

[8] Hansen, B., D.G. Novick, and S. Sutton, "Systematic Design of Spoken Prompts," *CHI 96 Proceedings*, Vancouver, Canada, Apr. 1996.

[9] Marcus, S.M., D.W. Brown, R.G. Goldberg, M.S. Schoeffler, W.R. Wetzel, and R.R. Rosinski, "Prompt Constrained Natural Language — Evolving the Next Generation of Telephony Services," *Proceedings of ICSLP '96*, Philadelphia, PA, pp. 857-860, Oct. 1996.

[10] Tegey, H., and B. Robson, *A Reference Grammar of Pashto*, Center for Applied Linguistics, Washington, D.C., 1996.

[11]Walker, M., R. Passonneau, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker, "Cross-Site Evaluation in DARPA Communicator: The June 2000 Data Collection," submitted to *Computer Speech and Language*, 2002.