Non-Native Users in the Let's Go!! Spoken Dialogue System: Dealing with Linguistic Mismatch

Antoine Raux and **Maxine Eskenazi**
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15232, USA
{antoine+,max+}@cs.cmu.edu

Abstract

This paper describes the CMU Let's Go!! bus information system, an experimental system designed to study the use of spoken dialogue interfaces by non-native speakers. The differences in performance of the speech recognition and language understanding modules of the system when confronted with native and non-native spontaneous speech are analyzed. Focus is placed on the linguistic mismatch between the user input and the system's expectations, and on its implications in terms of language modeling and parsing performance. The effect of including non-native data when building the speech recognition and language understanding modules is discussed. In order to close the gap between non-native and native input, a method is proposed to automatically generate confirmation prompts that are both close to the user's input and covered by the system's language model and grammar, in order to help the user acquire idiomatic expressions appropriate to the task.

1 Introduction

1.1 Spoken Dialogue Systems and Non-Native Speakers

Spoken dialogue systems rely on models of human language to understand users' spoken input. Such models cover the acoustic and linguistic space of the common language used by the system and the user. In current systems, these models are learned from large corpora of recorded and transcribed conversations matching the domain of the system. In most of the cases, these corpora are gathered from native speakers of the language because they are the main target of the system and because developers and researchers are often native speakers themselves. However, when the common language is not the users' native language, their utterances might fall out of this "standard" native model, seriously degrading the recognition accuracy and overall system performance. As telephone-based information access systems become more common and available to the general public, this inability to deal with non-native speakers (or with any "non-standard" subgroup such as the elderly) is a serious limitation since, at least for some applications, (e.g. tourist information, legal/social advice) non-native speakers represent a significant portion of the everyday user population.

1.2 Previous Work on Non-Native Speech Recognition

Over the past ten years, extensive work has been done on non-native speech recognition. Early research aimed at endowing Computer Assisted Language Learning software with speech recognition capabilities (e.g. (Eskenazi and Hansma, 1998), (Witt and Young, 1997)). Usually such systems are targeted at one specific population, that is, people who share the same native language (L1). Thus, most research in non-native speech recognition uses knowledge of the L1, as well as databases of accented speech specially recorded from speakers of the target population. Ideally, by training acoustic models on target non-native speech, one would capture its specific characteristics just as training on native speech does. However collecting amounts of non-native speech that are large enough to fully train speaker-independent models is a hard and often impractical task. Therefore, researchers have resorted to using smaller amounts of non-native speech to retrain or adapt models that were originally trained on large corpora of native speech. As for native speech, such methods were mostly applied to read speech, with some success (e.g. (Mayfield Tomokiyo and Waibel, 2001)).

Unfortunately, we know from past research on native speech recognition that read speech models perform poorly on conversational speech (Furui, 2001), which is the style used when talking to spoken dialogue systems. A few studies have built and used databases of non-native conversational speech for evaluation (Byrne et al., 1998), and training (Wang and Schultz, 2003).

In all those cases, the native language of the speaker is known in advance. One exception is (Fischer et al., 2001) who apply multilingual speech recognition methods to non-native speech recognition. The authors train acoustic models on a database comprising native speech from five European languages (English, Spanish, French, German and Italian) and use them to recognize non-native English from speakers of 10 European countries. However, their task is the recognition of read digit strings, quite different from conversational speech.

Also, because of the difficulty researchers have to record large amounts of spontaneous non-native speech, no thorough study of the impact of the linguistic differences between native and non-native spontaneous speech has been conducted to our knowledge. The two spontaneous non-native speech studies cited above, report perplexity and out-of-vocabulary (OOV) word rate (for (Wang and Schultz, 2003)) but do not provide any analysis.

In this paper, while acknowledging the importance of acoustic mismatch between native models and non-native input, we focus on linguistic mismatch in the context of a task-based spoken dialogue system. This includes differences in word choices which influences the number of OOV words, and syntax which affects the performance of the speech recognizer's language model and of the natural language understanding (NLU) grammar.

1.3 Non-Native Speakers as Language Learners

All the research on non-native speech recognition described in the previous section sees non-native speakers as a population whose acoustic characteristics need to be modeled specifically but in a static way, just like one would model the acoustics of male and female voices differently. A different approach to the problem is to see non-native speakers as engaged in the process of acquiring the target language's acoustic, phonetic and linguistic properties. In this paradigm, adapting dialogue systems to non-native speakers does not only mean being able to recognize and understand their speech as it is, but also to help them acquire the vocabulary, grammar, and phonetic knowledge necessary to fulfill the task the system was designed for.

This idea follows decades of language teaching research that, since the mid sixties, has emphasized the value of learning language in realistic situations, in order to perform specific tasks. Immersion is widely considered as the best way to learn to speak a language and modern approaches to foreign language teaching try to mimic its characteristics. If the student cannot be present in the country the language is spoken in, then the student should be put into a series of situations imitating the linguistic experience that he/she would have in the target country. Thus, most current language teaching methods, following the Communicative Approach (Littlewood, 1981) have focused on creating exercises where the student is forced to use language quickly in realistic situations and thus to learn from the situation itself as well as from reactions to the student's actions.

From a different viewpoint, (Bortfeld and Brennan, 1997) showed in a psycholinguistic study that non-native speakers engaged in conversation-based tasks with native speakers do not only achieve the primary goal of the task through collaborative effort but also acquire idiomatic expressions about the task from the interaction.

The research described in this paper, has the dual goal of improving the accessibility of spoken dialogue systems to non-native speakers and of studying the usability of a computer for task-based language learning that simulates immersion.

The next section gives an overview of the CMU Let's Go!! bus information system that we built and use in our experiments. Section 3 describes and analyzes the results of experiments aimed at comparing the accuracy of speech recognition and the quality of language modeling on both native and non-native data. Section 4 describes the use of automatically generated confirmation prompts to help the user speak the language expected by the system. Finally, section 5 draws conclusions and presents future directions of research.

2 Overview of the System

2.1 The CMU Let's Go!! Bus Information System

In order to study the use of spoken dialogue systems by non-native speakers in a realistic setting, we built Let's Go!!, a spoken dialogue system that provides bus schedule information for the Pittsburgh area(Raux et al., 2003). As shown in Figure 1, the system is composed of five basic modules: the speech recognizer, the parser, the dialog manager, the language generator, and the speech synthesizer. Speech recognition is performed by the Sphinx II speech recognizer (Huang et al., 1992). The Phoenix parser (Ward and Issar, 1994) is in charge of natural language understanding. The dialogue manager is based on the RavenClaw framework (Bohus and Rudnicky, 2003). Natural language generation is done by a simple template-based generation module, and speech synthesis by the Festival speech synthesis system (Black et al., 1998). The original system uses a high quality limited-domain voice recorded especially for the project but for some experiments, lower quality, more flexible voices

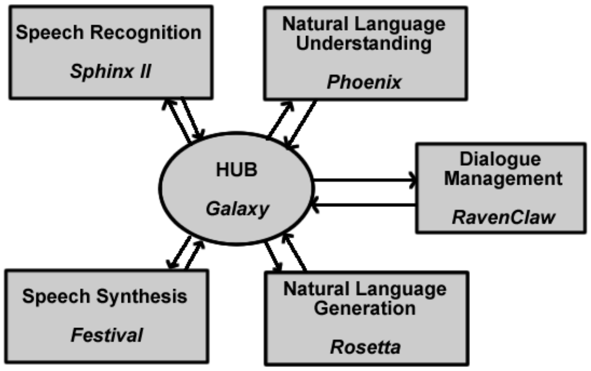

Figure 1: General architecture of the Let's Go!! bus information system.

have been used. All modules communicate through the Galaxy-II (Seneff et al., 1998) framework.

2.2 Definition of the Domain

The Port Authority of Allegheny County, which manages the buses in Pittsburgh provided the full database of bus routes and schedules. Overall, this database contains more than 10,000 bus stops but we restricted our system to 5 routes and 559 bus stops in areas where international students are likely to travel since they are our main target population at present.

In order to improve speech recognition accuracy, we concatenated the words in the name of each bus stop (e.g. "Fifth_and_Grant") and made them into a single entry in the recognizer's lexicon. Because there are usually several variant names for each bus stop and since we included other places such as landmarks and neighborhoods, the total size of the lexicon is 9914 words.

2.3 Data Collection Experiments

To gather enough data to train and test acoustic and language models, we had the system running, advertising it to international students at our university, as well as conducting several studies. In those studies, we gave scenarios to the participants in the form of a web page with maps indicating the places of departure and destination, as well as additional time and/or route preferences. There was as little written English as possible in the description of the scenarios to prevent influencing the language habits of the participants. Participants then called the system over the phone to get the required information. One experiment conducted in June 2003 netted 119 calls from 11 different non-native speakers (5 of them were from India and 6 from Japan), as well as 25 calls from 4 native speakers of American English. Another experiment in August 2003 allowed the collection of 47 calls from 6 non-native speakers of various linguistic backgrounds. The rest of the non-native data comes from unsollicited individual callers labelled as non-native by a human annotator who transcribed their speech. The total size of the spontaneous non-native corpus is 1757 utterances.

	Native	Non-Native
Word Error Rate	20.4 %	52.0 %

Table 1: Word Error Rate of the speech recognizer with a native language model on native and non-native data.

3 Recognition and Understanding of Non-Native Speech

3.1 Recognition Accuracy

We used acoustic models trained on data consisting of phone calls to the CMU Communicator system(Rudnicky et al., 2000). The data was split into gender specific sets and corresponding models were built. At recognition time, the system runs the two sets of models in parallel and for each utterance selects the result that has the highest recognition score, as computed by Sphinx. The language model is a class-based trigram model built on 3074 utterances from past calls to the Let's Go!! system, in which place names, time expressions and bus route names are each replaced by a generic class name to compensate for the lack of training data.

In order to evaluate the performance of these models on native and non-native speakers, we used 449 utterances from non-native users (from the August experiment and the unsollicited calls) and 452 from native users of the system. The results of recognition on the two data sets are given in Table 1. Even for native speakers, performance was not very high with a word error rate of 20.4%. Yet, this is acceptable given the small amount of training data for the language model and the conversational nature of the speech. However, performance degrades significantly for non-native speakers, with a word error rate of 52.0%. The two main potential reasons for this loss are acoustic mismatch and linguistic mismatch. Acoustic mismatch arises from the variations between the native speech on which the acoustic models were trained and non-native speech, which often include different accents and pronunciations. On the other hand, linguistic mismatch stems from variations or errors in syntax and word choice, between the native corpus on which the language model was trained and non-native speech.

3.2 Impact of Linguistic Mismatch on the Performance of the Language Model

To analyze the effect of linguistic mismatch, we compared the number of out-of-vocabulary words (OOV) and the perplexity of the model on the transcription of the test utterances. Table 2 shows the results. The percentage of

	Native	Non-Native	Difference	Significance
% OOV words	1.2 %	3.09 %	157.5 %	$p < 10^{-4}$
% utt. w/ OOV words	5.9 %	14.0 %	174.5 %	$p < 10^{-5}$
Perplexity	22.89	36.55	59.7 %	–
% words parsed	63.3 %	56.0 %	56.0 %	$p < 10^{-9}$
% utt. fully parsed	56.4 %	49.7 %	49.7 %	$p < 0.05$

Table 2: The native language model and parsing grammar applied to native and non-native speech transcriptions. The statistical significance of the difference between the native and non-native sets is computed using the chi-square test for equality of distributions.

OOVs is 3.09% for non-native speakers, more than 2.5 times higher than it is for native speakers, which shows the difference in word choices made by each population. Such differences include words that are correctly used but are not frequent in native speech. For example, when referring to bus stops by street intersections, all native speakers in our training set simply used “A and B”, hence the word “intersection” was not in the language model. On the other hand, many non-native speakers used the full expression “the intersection of A and B”. Note that the differences *inside* the place name itself (e.g. “A and B” vs “A at B”) are abstracted away by the class-based model, since all variants are replaced by the same class name (words like “intersection” and “corner” were kept out of the class to reduce the number of elements in the “place” class). In other cases non-native speakers used inappropriate words, such as “bus timing” for “bus schedule”, which were not in the language model. Ultimately, OOVs affect 14.0% of the utterances as opposed to 5.9% for native utterances, which is significant, since an utterance containing an OOV is more likely to contain recognition errors even on its in-vocabulary words, since the OOV prevents the language model from accurately matching the utterance. Differences between the native and non-native set in both OOV rate and the ratio of utterances containing OOVs were statistically significant.

We computed the perplexity of the model on the utterances that did not contain any OOV. The perplexity of the model on this subset of the non-native test set is 36.55, 59.7% higher than that on the native set. This reflects differences in syntax and selected constructions. For example, although native speakers almost always used the same expression to request a bus departure time (“When does the bus leave ...?”), non-natives used a wider variety of sentences (e.g. “Which time I have to leave?”, “What the next bus I have to take?”). Both the difference between native and non-native and the larger variability of non-native language account for the larger perplexity of the model over the non-native set. This results seems to disagree with what (Wang and Schultz, 2003) found in their study, where the perplexity was larger on the native set. Unfortunately, they do not describe the data used to train the language model so it is hard to draw any conclusions. But one main difference is that their experiment focused only on German speakers of English, whereas we collected data from a much more diverse population.

3.3 Impact of the Linguistic Mismatch on Language Understanding

The Phoenix parser used in the natural language understanding module of the system is a robust, context-free grammar-based parser. Grammar rules, including optional words, are compiled into a grammar network that is used to parse user input. When no complete parse is found, which is often the case with spoken language, Phoenix looks for partial parses and returns the parse forest that it is most confident in. Confidence is based on internal measures such as the number of words covered by the parses and the number of parse trees in the parse forest (for an equal number of covered words, a smaller number of parse trees is preferred).

The grammar rules were hand written by the developers of the system. Initially, since no data was available, choices were made based on their intuition and on a small scale Wizard-of-Oz experiment. Then, after the first version of the system was made available, the grammar was extended according to actual calls to the system. The grammar has thus undergone continuous change, as is often the case in spoken dialogue systems.

The grammar used in this experiment (the “native” grammar) was designed based for native speech without adaptation to non-native data. It provides full parses of sentences like “When is the next bus going to the airport?”, but also, due to the robustness of the parser, partial parses to ungrammatical sentences like “What time bus leave airport?”. Once compiled, the grammar network consisted of 1537 states and 3076 arcs. The two bottom rows of Table 2 show the performance of the parser on human-transcribed native and non-native utterances. Both the number of words that could be parsed and the number of sentences for which a full parse was obtained are larger for native speakers (resp. 63.3% and 56.4%) than non-native (56% and 49.7%), although the relative differences are not as large as those observed for the lan-

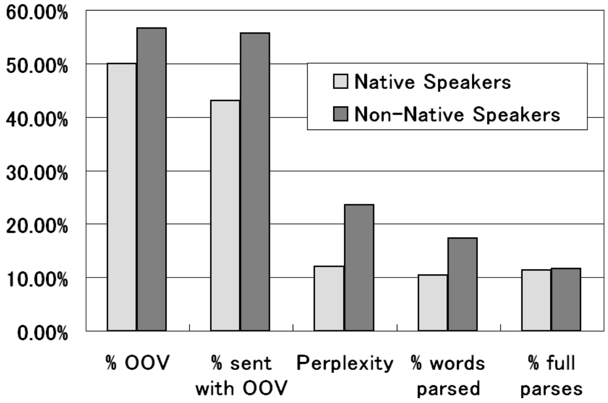

Figure 2: Comparison of the relative gain obtained by using a language model and grammar that includes some non-native data over the original purely native model, on transcribed native and non-native speech.

guage model. This can be attributed to the original difficulty of the task since even native speech contains a lot of disfluencies that make it difficult to parse. As a consequence, robust parsers such as Phoenix, which are designed to be flexible enough to handle native disfluencies, can deal with some of the specificities of non-native speech. Yet, the chi-square test shows that the difference between the native and non-native set is very significant for the ratio of words parsed and mildly so for the ratio of fully parsed sentences. The weak significance of the latter can be partly explained by the small number of utterances in the corpora.

3.4 Effect of Additional Non-Native Data on Language Modeling and Parsing

In order to study the improvement of performance provided by mixing native and non-native data in the language model, we built a second language model (the "mixed" model), using the 3074 sentences of the native model to which were added 1308 sentences collected from non-native calls to the system not included in the test set. Using this model, we were able to reduce the OOV rate by 56.6% and perplexity by 23.6% for our non-native test set. While the additional data also improved the performance of the model on native utterances, the improvement was relatively smaller than for non-native speakers (12.1%). As can be seen by comparing Tables 2 and 3, this observation is also true of OOV rate (56.6% improvement for non-native vs 50.0% for native) and the proportion of sentences with OOVs (43.1% vs 55.7%). Figure 2 shows the relative improvement due to the mixed LM over the native LM on the native and non-native set.

We also evaluated the impact of additional non-native data on natural language understanding. In this case, since we wrote the grammar manually and incrementally over time, it is not possible to directly "add the non-native data" to the grammar. Instead, we compared the June 2003 version of the grammar, which is mostly based on native speech, to its September 2003 version, which contains modifications based on the non-native data collected during the summer. This part is therefore an evaluation of the impact of the human grammar design done by the authors based on additional non-native data. At that point, the compiled grammar had grown to contain 1719 states and 3424 arcs which represents an increase of respectively 11.8% and 11.3% over the "native" grammar. Modifications include the addition of new words (e.g. "reach" as a synonym of "arrive"), new constructs (e.g. "What is the next bus?") and the relaxation of some syntactic constraints to accept ungrammatical sentences (e.g. "I want to arrive the airport at five" instead of "I want to arrive at the airport at five"). Using this new grammar, the proportion of words parsed and sentences fully parsed improved by respectively 10.4% and 11.3% for the native set and by 17.3% and 11.7% for the non-native set. We believe that, as for the language model, the reduction in the number of OOVs is the main explanation behind the better improvement in word coverage observed for the non-native set compared to the native set. The reduction of the difference between the native and non-native sets is also reflected in the weaker significance levels for all ratios except that of fully parsed utterances, in 3, larger p-values meaning that there is a larger probability that the differences between the ratios were due to spurious differences between the corpora rather than to their (non-)nativeness.

This confirms that even for populations with a wide variety of linguistic backgrounds, adding non-native data does reduce the linguistic mismatch between the model and new, unseen, non-native speech. Another explanation is that, on a narrow domain such as bus schedule information, the linguistic variance of non-native speech is much larger than that of native speech. Therefore, less data is required to accurately model native speech than non-native speech. It also appears from these results that, in the context of task-based spoken dialogue systems, higher-level modules, such as the natural language understanding module, are less sensitive to explicit modeling of non-nativeness. This can be explained by the fact that such modules were designed to be flexible in order to compensate for speech recognition errors. This flexibility benefits non-native speakers as well, regardless of additional recognition errors.

3.5 Effect of Additional Non-Native Data on Speech Recognition

Unfortunately, the reduction of linguistic mismatch was not observed on recognition results. While using the new language model improved word error rate on both native

	Native	Non-Native	Difference	Significance
% OOV words	0.6 %	1.34 %	123.3 %	$p < 0.05$
% utt. w/ OOV words	2.9 %	6.2 %	113.8 %	$p < 0.01$
Perplexity	20.12	27.92	38.8 %	–
% words parsed	69.9 %	65.7 %	65.7 %	$p < 10^{-3}$
% utt. fully parsed	62.8 %	55.5 %	55.5 %	$p < 0.05$

Table 3: The mixed language model and parsing grammar applied to native and non-native speech transcriptions. Significance is computed using the chi-square test, except for perplexity where the relative difference is reported.

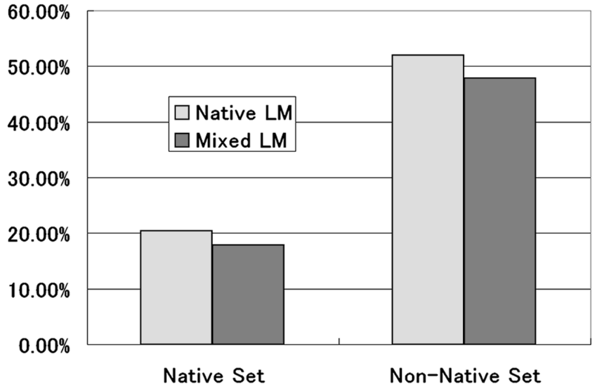

Figure 3: Word Error Rate on Native and Non-Native Data using a Native and a Mixed Language Model

and non-native utterances (resp. to 17.8% and 47.8%, see Figure 3), the impact was relatively larger for native speech. This is an indication that acoustics play a prominent role in the loss of accuracy of speech recognition on non-native speech. Acoustic differences between native and non-native speakers are likely to be larger than the linguistic ones, since, particularly on such a limited and common domain, it is easier for non-native speakers to master syntax and word choice than to improve their accent and pronunciation habits. Differences among non-native speakers of different origins are also very large in the acoustic domain, making it hard to create a single acoustic model matching all non-native speakers. Finally, the fact that additional non-native data improves performance on native speech is a sign that, generally speaking, the lack of training data for the language model is a limiting factor for recognition accuracy. Indeed, if there was enough data to model native speech, additional non-native data should increase the variance and therefore the perplexity on native speech.

4 Adaptive Lexical Entrainment as a Solution to Linguistic Mismatch

4.1 Gearing the User To the System's Language

The previous section described the issue of recognizing and understanding non-native speech and solutions to adapt traditional systems to non-native speakers. Another approach is to help non-native users adapt to the system by learning appropriate words and expressions. Lexical entrainment is the phenomenon by which, in a conversation, speakers negotiate a common ground of expressions to refer to objects or topics. Developers of spoken dialogue systems frequently take advantage of lexical entrainment to help users speak utterances that are within the language model of the system. This is done by carefully designing the system prompts to contain only words that are recognized by the recognition and understanding modules (Gustafson et al., 1997). However, in the case of non-native users, there is no guarantee that users actually know the words the system wants them to use. Also, even if they do, some non-native speakers might prefer to use other words, which they pronounce better or that they better know how to use. For those reasons, we believe that to be optimal, the system must try to match the user's choice of words in its own prompts. This idea is motivated by the observations of (Bortfeld and Brennan, 1997), who showed that this type of adaptation occurs in human-human conversations between native and non-native speakers.

The role of the system's "native" prompts is to take the users through the shortest path from their current linguistic state to the system's expectations. In fact, this is not only true for non-native speakers and lexical entrainment is often described as a negotiation process between the speakers (Clark and Wilkes-Gibbs, 1986). However, while it is possible for limited-domain system designers to establish a set of words and constructions that are widely used among native speakers, the variable nature of the expressions mastered by non-native speakers make adaptation a desirable feature of the system.

4.2 Automatic Generation of Corrective Prompts

In this study, not all prompts were modified to match the user's choice of words. Instead, the focus was placed on confirmation prompts that both ensure proper understanding between the user and the system and lexically entrain the user towards the system's expected input. Two questions arise: how to generate the prompts and when to trigger them. Our approach has been to design a list of target prompts that fit the system's language model and grammar and find the closest target prompt to each user input. The distance between a user utterance as recognized by Sphinx and each of the target utterances is computed by the same dynamic programming algorithm that is traditionally used to compute word error rate in speech recognition evaluation. It determines the number of word insertions, deletions and substitutions that lead from the target prompt to the user's utterance. The target prompt that is closest, i.e. that requires the fewest operations to match the input, is selected. In addition, words that represent important concepts such as places, times or bus route numbers, are given additional weight. This follows the assumption that a target sentence is not appropriate if it has a missing or an extra concept compared to the utterance. We also used this heuristic to answer the second question: when to trigger the confirmation prompts. The system asks for a confirmation whenever a target sentence is found that contains the same concepts as the user input and differs from it by at least one word. In this case a prompt like "Did you mean ..." followed by the target sentence is generated. Finally, the dynamic programming algorithm used to align the utterances also locates the words that actually differ between the input and the target. This information is sent to the speech synthesizer, which puts particular emphasis on the words that differ. To provide natural emphasis, the intonation of all sentences is generated by the method described in (Raux and Black, 2003) that concatenates portions of natural intonational contours from recorded utterances into a contour appropriate for each prompt. Since the domain-limited voice recorded for the project does not allow us to either generate non-recorded prompts or to modify the contour of the utterances, we used a different, generic voice for this version of the system.

4.3 Application and Example

The method described in the previous paragraph was implemented in the system and tested in a small pilot study. We manually wrote 35 different target prompts describing departure and destination places, times and route numbers, based on our knowledge of the system's language model and grammar. An example of a confirmation dialogue obtained from one of these prompts is given in Figure 4. In the first user utterance, the preposition "to" is missing, either because it was not pronounced by the user or because it was not recognized by the speech recognition module. As a consequence, the utterance cannot be fully parsed by the language understanding module. In parallel, the confirmation module computes the distance between the user's input and each of the 35 target prompts, and identifies the closest one as "I want to go to the airport". At the same time it finds that the user's utterance is obtained from the target by deleting the word "to" and therefore stresses it in the confirmation prompt. Once the user answers "yes" to the confirmation prompt, the target prompt is sent to the parser as if it had been uttered by the user and the state of the dialogue is updated accordingly. If the user answers "no", the prompt is simply discarded. We found that this method works well when speech recognition is only slightly degraded and/or when the recognition errors mostly concern grammar and function words. In such cases, this approach is often able to repair utterances that would not be parsed correctly otherwise. However, when too many recognition errors occur, or when they affect the values of the concepts (i.e. the system recognizes one place name instead of another), the users receive too many confirmation prompts to which they must respond negatively. Combined with the difficulty that non-native speakers have in understanding unexpected synthesized utterances, this results in cognitive overload on the user. Yet, this method provides an easy way (since the designer only has to provide the list of target prompts) to generate adaptive confirmation prompts that are likely to help lexical entrainment.

```
S: What can I do for you?
U: I want to go the airport.
S: Sorry, I didn't get that.
   Did you mean:
   I want to go TO the airport?
U: Yes
S: To the airport.
   Where are you leaving from?
U: ...
```

Figure 4: Example of an adaptive confirmation dialogue. The capital "TO" indicate that the word was emphasized by the system.

5 Conclusion and Future Directions

In this paper, we described the Let's Go!! bus information system, a dialogue system targetted at non-native speakers of English. In order to investigate ways to improve the communication between non-native users and the system, we recorded calls from both native and non-native speakers and analyzed their linguistic properties. We found that besides the problem of acoustic mismatch that results from the differences in accent and pronunciation habits, linguistic mismatch is also significant and degrades the performance of the language model and the natural language understanding module. We are exploring two solutions to reduce the linguistic gap between native and non-native users. First we studied the impact of taking into account non-native data to model the user's language and second we designed a mechanism to generate confirmation prompts that both match the user's input and a set of predefined target utterances, so as to help the user acquire idiomatic expressions related to the task.

Real-world systems like Let's Go!! are in constant evolution because the data that is collected from users calling the system is used to refine the acoustic and linguistic models of the system. In the near future, our priority is to collect more data to improve the acoustic models of the system and address the specific issues related to a general non-native population, which does not share a common native language. We will also work on integrating the confirmation prompt generation method proposed in this work with state-of-the-art confidence annotation methods.

6 Acknowledgments

The authors would like to thank Alan W Black, Dan Bohus and Brian Langner for their help with this research.

This material is based upon work supported by the U.S. National Science Foundation under Grant No. 0208835, "LET'S GO: improved speech interfaces for the general public". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

A. Black, P. Taylor, and R. Caley. 1998. The Festival speech synthesis system. http://festvox.org/festival.

D. Bohus and A. Rudnicky. 2003. Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda. In *Proc. Eurospeech 2003*, pages 597–600, Geneva, Switzerland.

H. Bortfeld and S. Brennan. 1997. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23:119–147.

W. Byrne, E. Knodt, S. Khudanpur, and J. Bernstein. 1998. Is automatic speech recognition ready for non-native speech? A data collection effort and initial experiments in modeling conversational hispanic english. In *Proc. ESCA Workshop on Speech Technology in Language Learning*, pages 37–40, Marholmen, Sweden.

H. Clark and D. Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

M. Eskenazi and S. Hansma. 1998. The Fluency pronunciation trainer. In *Proc. ESCA Workshop on Speech Technology in Language Learning*, pages 77–80.

V. Fischer, E. Janke, S. Kunzmann, and T. Ross. 2001. Multilingual acoustic models for the recognition of non-native speech. In *Proc. ASRU '01*, Madonna di Campiglio, Italy.

S. Furui. 2001. From read speech recognition to spontaneous speech understanding. In *Proc. 6th Natural Language Processing Pacific Rim Symposium*, pages 19–25, Tokyo, Japan.

J Gustafson, A. Larsson, R. Carlson, and K. Hellman. 1997. How do system questions influence lexical choices in user answers? In *Proc. Eurospeech '97*, pages 2275–2278, Rhodes, Greece.

X. Huang, F. Alleva, H.-W. Hon, K.-F. Hwang, M.-Y. Lee, and R. Rosenfeld. 1992. The SPHINX-II speech recognition system: an overview. *Computer Speech and Language*, 7(2):137–148.

W. Littlewood. 1981. *Communicative Language Teaching*. Cambridge University Press.

L. Mayfield Tomokiyo and A. Waibel. 2001. Adaptation methods for non-native speech. In *Proc. Multilinguality in Spoken Language Processing*, Aalborg, Denmark.

A. Raux and A. Black. 2003. A unit selection approach to f0 modeling and its application to emphasis. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop 2003*, pages 700–705, Saint Thomas, US Virgin Islands.

A. Raux, B. Langner, A. Black, and M. Eskenazi. 2003. Lets go: Improving spoken dialog systems for the elderly and non-natives. In *Proc. Eurospeech 2003*, pages 753–756, Geneva, Switzerland.

A. Rudnicky, C. Bennett, A. Black, A. Chotimongkol, K. Lenzo, A. Oh, and R. Singh. 2000. Task and domain specific modelling in the carnegie mellon communicator system. In *Proc. ICSLP 2000*, Beijing, China.

S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue. 1998. Galaxy-II: A reference architecture for conversational system development. In *Proc. ICSLP '98*, Sydney, Australia.

Z. Wang and T. Schultz. 2003. Non-native spontaneous speech recognition through polyphone decision tree specialization. In *Proc. Eurospeech '03*, pages 1449–1452, Geneva, Switzerland.

W. Ward and S. Issar. 1994. Recent improvements in the CMU spoken language understanding system. In *Proc. ARPA Human Language Technology Workshop*, pages 213–216, Plainsboro, NJ.

S. Witt and S. Young. 1997. Language learning based on non-native speech recognition. In *Proc. Eurospeech '97*, pages 633–636, Rhodes, Greece.