ELICITING NATURAL SPEECH FROM NON-NATIVE USERS: COLLECTING SPEECH DATA FOR LVCSR

Laura Mayfield Tomokiyo and Susanne Burger
Interactive Systems Laboratories
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
{laura,sburger}@cs.cmu.edu

Abstract

In this paper, we discuss the design of a database of recorded and transcribed read and spontaneous speech of semi-fluent, strongly-accented non-native speakers of English. While many speech applications work best with a recognizer that expects native-like usage, others could benefit from a speech recognition component that is forgiving of the sorts of errors that are not a barrier to communication; in order to train such a recognizer a database of non-native speech is needed. We examine how collecting data from non-native speakers must necessarily differ from collection from native speakers, and describe work we did to develop an appropriate scenario, recording setup, and optimal surroundings during recording.

1 Introduction

As part of work in improving speech recognition performance for non-native speakers, we wanted to develop a database that captures ways in which non-native language use differs from native language use in a specific domain. Features we were interested in include pronunciation, lexical choice, syntax, expressive goals, and strategies speakers use when they are unsure of the appropriate English expression. We wanted the recorded data to be appropriate for LVCSR system training, which means that the signal quality should be good and the speech should be as close as possible in terms of style and content to speech that will be used in the target application, a tourist information query system. We also wanted to elicit data which would contain examples of systematic and unsystematic variation in the speech of low- to mid-fluency non-native speakers.

One of the most interesting aspects of these experiments was the ways in which we found ourselves needing to adapt our usual data collection strategies to the needs of our speakers, whose English abilities varied from beginning to near-native. It is important to be aware of a number of assumptions that are commonly made which do not necessarily hold for non-native speakers, and which it is important to address when designing a data collection protocol.

The act of speaking is not difficult. When recording native speakers speaking spontaneously for standard LVCSR projects (that is, not projects geared towards special populations or difficult tasks), it is assumed that the the act of speaking does not in and of itself represent a major cognitive load for the speaker. This can be very untrue of non-native speakers, and we had several speakers ask to quit in the middle of the recording because they felt unable to continue. The researcher needs to make a decision about what to do in such a situation, and possibly prepare an alternate task.

There is little risk of alienating the community. Local communities of non-native speakers are not always large, and if it is close knit, word can quickly spread if the task is too hard or embarassing. Also, it is important to de-emphasize the fact that we are interested, among other things, in imperfections in the speaker's speech, or risk offending the community.

The task is not perceived as a test. Again, when speaking spontaneously, few native speakers of nonstigmatized varieties of English would feel that they are being evaluated on the correctness of their speech. Many non-native speakers will feel tested, and as this can make them nervous and affect their speech, it is important to reassure them as far as possible that they are not being tested and that the data is being anonymized.

The speaker knows what to say. Most spontaneous collection tasks are chosen because they are tasks speakers can be expected to have done before and be comfortable with. Although a non-native speaker has probably made an airplane reservation in his native language before, it is entirely possible that he has never done so in the target language, and does not have a good idea of what he should say in that situation. If he were really planning to make an airplane reservation in the target language, he would probably think about what to say in advance and might even ask someone, which he may not have a chance to do during the data collection. This undermines the representativeness of the database.

We carried out a number of exploratory experiments to try to determine the format which was the most comfortable for the speakers and which resulted in elicitation of the most natural data; two of these experiments are described in Section 3. For these experiments we worked with native speakers of Japanese. The protocol that we settled on, which we feel is very effective for non-native speakers, is described in Section 4. Although transcription and analysis of this data is at the beginning stages, we have already seen patterns that will be useful for developing acoustic and language models. Examples are shown in Section 5.

2 Related Work

Byrne et al.(Byrne and others, 1998) describe a conversational English data collection protocol with native speakers of Spanish as its targets. They identified their speakers with one of three skill levels and had them perform level-appropriate tasks designed to elicit specific grammatical structures. Participants spoke over the telephone with other non-native speakers, forcing them to communicate using speech. They found that this was an effective way to elicit spontaneous speech from non-native speakers of all fluency levels in a purely conversational domain.

A number of studies discuss techniques for collecting spoken data from non-native speakers in the context of a language tutoring system. Most such systems ((Eskenazi, 1997; Witt and Young, 1997; Kawai and Hirose, 1997) are examples) ask users to read a prompt or narrowly constrain what the user is allowed to say. Neumeyer et al. (Neumeyer et al., 1998) describe a system that evaulates students' pronunciation in text-independent speech. They collected a database of read speech, both newspaper and conversational sentences, and imitated speech, in which students imitated the speech of native speakers; as subjects, they used American students of French.

Aist et al. (Aist and others, 1998) discuss considerations in collecting speech from children, pointing out that children may be uncooperative and easily bored, and may have difficulty reading. They describe an unsupervised data collection method in which recognized speech is compared to the transcript that the child is expected to read, and utterances in which part or all of hypothesis match the transcript are used for additional system training. This type of technique is not as effective for a system that handles completely spontaneous queries, but their observations about children's abilities (especially articulatory and reading difficulties) and reaction to formalized data collection parallel ours in our study of non-native speakers.

Outside the field of speech recognition, much research has been done into methods for eliciting natural speech. Briggs (Briggs, 1986) emphasizes the importance of understanding the meaning of the speech event for the speaker. Recording for a research project may be a familiar event for the researcher, but not for the speaker. Reading aloud is commonplace in American schools, but participants of different backgrounds may be intimidated or even offended when asked to read aloud. While native speakers of English certainly vary in their comfort reading and speaking, when the researchers are also native speakers of English, there are far fewer cultural variables that can lead to misunderstanding and compromise the integrity of the data.

In his description of the field methodology in the project on linguistic change and variation, Labov (Labov, 1984) describes a number of issues in spoken data collection, mentioning among other things the long-term relationship with the speaker pool. This is of course important for both longitudinal studies; also, when studying the speech of a restricted group, it is important that people do not come out of the data collection experience feeling that they have been objectified or misunderstood. Labov returns to this point in the context of ethical considerations in data collection.

What exactly does "natural speech" mean in the case of the non-native speaker? Wolfson (Wolfson, 1976) defines the notion of natural speech "as properly equivalent to that of appropriate speech; as not equivalent to unselfconscious speech." That is, in some situations, it is *natural* to speak carefully, and that careful speech in such contexts should not be considered unnatural. For semi-fluent non-native speakers, whether they are at a real information desk or recording a contrived scenario, their speech will most likely be planned.

3 Pilot Experiments

3.1 Recording Setup

All recordings were taken by a DAT recorder; speakers wore a Sennheiser headset. Recordings were done in a small computer lab with some incidental noise but no excessive outside noise. On some occasions there were other people in the room when the recording was being done; this will be discussed further below. In non-interactive recordings, users were seated at a table with the instruction sheets, pen or pencil, and water. Speakers were permitted to stop and restart recording at any time.

We did two pilot experiments which greatly helped us to understand the needs of our speakers and how we could make them more comfortable, in turn improving the quality of our data. For these experiments, we recorded native speakers of Japanese.

3.1.1 Pilot experiment one

In the first experiment, we drew from a human-machine collection task that we had had success with for native speakers in a similar application in another domain. Speakers were provided with prompts such as the following:

- Ask how to get to the museum
- Find out where you can exchange money
- Ask where to get a ticket for the subway

Speakers came in on two different occasions and gave us feedback after both. The first time they came in, they were given the prompts in English. As we had predicted, they were strongly influenced in their word choice by the phrasings used in the prompts. The second time they came in, they were given the prompts in their native language . They felt that this task was much harder; they perceived it as a translation task in which they were expected to give a correct answer, whereas with the English prompts they were effectively given the correct answer. Their productions, however, were more varied, different both from each other and from the original English prompt.

In addition to the prompt-based task, we had speakers read from a local travel guide, specifically about the university area so that the context would be somewhat familiar. We found that there were indeed reading errors of the type that would not occur in spontanous speech.

We observed that some speakers were stumbling over words that they obviously didn't know. We attempted to normalize for this by having them read utterances that had been previously recorded and transcribed, hoping that they would be more likely to be familiar with words that other speakers of similar fluency had used. We still found that they had some difficulty in reading. Our speakers were native speakers of Japanese, however, which has a different writing system; this would have some influence.

There was also a fair amount of stumbling over words in the prompted tasks, especially with proper nouns, and we have not yet looked at the correspondence between stumbling in read speech of familiar words and stumbling in spontaneous speech. It may be the case that they are more closely related than they are for native speakers.

3.1.2 Pilot experiment two

In the second pilot experiment, we attempted a wizard-of-oz collection using an interactive map; the speakers could ask for locations and routes to be highlighted in the map, and there was a text screen to which the wizard could send messages to answer speaker queries. Instead of a list of prompts, the speakers were given a sheet of paper listing some points of interest in the city, hotel names, some features that they could ask about (business hours, location, etc.) and the dates that they would be in the city. Their task was to plan a weekend, finding hotels, restaurants, and things to do. Our thought was that perhaps speakers would speak more naturally in an information-gathering task, where they are actually trying to communicate instead of simply producing sentences.

Our general impression was that although the visual highlighting of the locations was a feature that the users enjoyed, and which helped them to become involved in the task, the utterances could not be characterized as more natural than those given in the prompted task. It was also our feeling that speakers were less sure of what to do in a less structured task; both lack of confidence in speaking and unfamiliarity with a "just say whatever comes to mind" approach contributed to their general discomfort. It took time to read and understand the responses from the wizard; also, speakers were aware that someone (the wizard) was listening in. Both of these factors were additional sources of self-consciousness. Although we thought that the repair dialogues that came about when the wizard misunderstood the speaker were valuable data, and that someone trained to provide responses geared toward the fluency level of the speaker would have more success as a wizard, it was our opinion that given the range of fluency levels we were targeting, wizard-of-oz collection would not be ideal for the following two reasons:

- communication and speaker confidence break down when the speaker is really having trouble expressing himself and the wizard cannot understand
- simulating a real-life experience, such as making a hotel reservation, without the real goal of wanting to stay in a hotel and background knowledge about the trip, can be very difficult depending on language ability and cultural background

4 Final Protocol

The final data collection protocol that we settled on has three parts. The first is a series of scenarios, in each of which a situation is described in the speaker's native language (L1) and a list is given in bullet form of things relevant to the situation that the speaker is to ask about. For instance, if the situation is a Pittsburgh Steelers game, the speakers would see the bullets

- arena location
- ticket price
- seat availability
- transportation
- game time

The bullets are made as short as possible so that the speakers absorb them in a glance and can concentrate on formulating an original question instead of on translating a specific phrase or sentence.

The second part is a read task. There was no doubt left after the pilot experiments that the amount of patience speakers had with the prompted task was limited; after the novelty wore off speakers tired quickly. Although spontaneous data would be better than read data, read data would be better than no data, and speakers seemed willing to continue at least as long again reading as they had with the prompted task. We considered two types of material for the reading. Some sort of phonetically balanced text is often used for data collection, so that the system is trained with a wide variety of phonetic contexts. Given that our speakers are even more restricted in their phrasings than native speakers are in conversational speech, it is likely that some phonetic contexts are extremely sparsely represented in our data. However, it may be the case that semi-fluent speakers avoid some constructions precisely because they are difficult to pronounce, and a sparsity in the training data probably is a good predictor of a sparsity in unseen data; even with new words, which may have as-yet-unseen phonetic contexts, non-native speakers may not pronounce them at all in the way that the designer of the phonetically balanced text had anticipated. We chose a 1000-word version of the fairy tale Snow White for our read texts; it had the highest syllable growth rate of any of the fairy tales we looked at and we augmented the syllable inventory by replacing some words with others, trying to ensure at the same time that all of the words were ones our speakers were likely to have encountered before.

Finally, we ask speakers to read a selection of previously recorded and transcribed utterances from the prompted task, both by native speakers and non-native speakers, randomly selected and with small modifications made to preserve anonymity. Our objective here was threefold: to quantify the difference between read dialogues and spontaneous dialogues; to quantify the difference between read dialogues and read prose; and to compare the performance of the end recognizer on native grammar with non-native pronunciation with performance on non-native grammar with non-native pronunciation.

We have recorded 23 speakers so far in the post-pilot phase of data collection, and all have expressed satisfaction with the protocol.

5 Analysis and Examples

Although transcription and analysis of the data we have collected so far is in the beginning stages, we have observed patterns that lead us to believe that our protocol is meeting our goals of eliciting speech from non-native speakers that is representative of what they would use in a real system and that begins to uncover patterns that are different from those native speakers use and will be useful in acoustic and language modeling.

The analysis in this section is based on transcribed data from 12 speakers. For comparison, we recorded three native speakers doing the same task the non-native speakers did (with English prompts). This is not a large sample, but gives us some evidence to support our intuitions about what native speakers would be likely to say.

5.1 Qualitative Analysis

Examples 1-3 show some sample utterances produced by the non-native speakers. In each example, the first sentence represents the prompt that would have been used for elicitation (speakers were actually given short bullets). Example 1 was selected to exemplify how speakers were influenced in their use of phrasal and colloquial verbs when given an English prompt. We observed that when prompted to ask for directions or travel time, native speakers almost always used the expression "get to." Non-native speakers often used this form when given an English prompt containing it, but almost never when given an L1 prompt.

1. **Ask how to get to the aquarium.**
   How do I get the aquarium?
   Please let me know how do you go the aquarium?
   I'd like to go to Aquarium.
   I want to go to the aquarium so please let me know how to go to there

In the data we have transcribed so far, 25 of 55 uses of *get to* were by non-native speakers, while 45 of 56 uses of *go to* were by non-native speakers.

Example 2 illustrates how number agreement can be influenced by the English prompt. Although native speakers often misspeak and disobey agreement rules in conversational speech, there are situations in which we observed that they are consistently careful, and the pattern any + N_{pl}, when appropriate, was one. The non-native speakers, on the other hand, consistently produced any + N_{sing} when not primed by an English prompt. "Any" was also often used where a native speaker would use "a."

2. **Ask if there are any [restaurants nearby / tickets available...].**
   Is there any restaurant around here?
   is there any good place to visit
   is there any available ticket
   do you have any special exhibition now
   is there any subway around

Of the 105 instances of use of the word "any," 52 were followed inappropriately by a singular noun. When the pattern "any place" is removed from the list, 52 out of 81 instances were grammatically incorrect in this way. To compare, 1 of 21 instances in the native sample were grammatically incorrect. Prescriptively incorrect grammar is expected in spontaneous speech even by native speakers. However, when non-native speech consistently strays from patterns observed in native speech, the bigram and trigram contexts used to model language at the sentence level can no longer be relied upon.

Of course, by using an L1 prompt we are influencing the speakers in the opposite direction, priming them to produce a translation of an L1 word and form an awkward English sentence around it when they might not do so in spontaneous system use. It is difficult to know whether this is the case with example 3. On the one hand, the speaker is clearly translating the Japanese term *nyuujouryou* (entrance fee). On the other hand, speakers consistently built a sentence around the word "fee" where a native speaker would use the pattern "how much does X cost" regardless of what Japanese term was used.

3. **Ask how much admission costs**
   How much is the fee for entrance?
   How much is fee for entering?
   How much is the fee for admission?

Although it was the element of the task that the speakers liked the least, the handling of unfamiliar expressions showed us how important it was to prompt users with specific queries that they might not know how to express. In real-world use, an application would have to handle such utterances, but in a more free-form data collection scenario speakers might avoid asking such questions altogether. We included among the Japanese prompts expressions which have no obvious English equivalent in order to observe how speakers expressed themselves when they did not know what the right English expression would be. Speakers were very inventive and almost always came up with an understandable English utterance, as shown in Figure 1 (displayed on the following page).

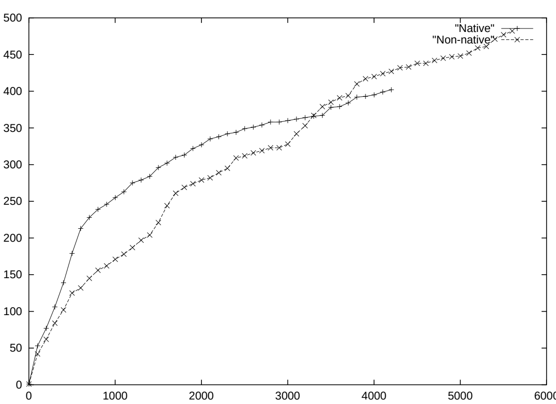

Figure 2: Vocabulary growth for native and non-native speakers in the tourist information task. Corpus size is displayed on the x axis and vocabulary size is displayed on the y axis.

5.2 Quantitative Analysis

Figure 2 shows the vocabulary growth rate for native and non-native speakers in the tourist information task that was our domain for these experiments. Interestingly, the vocabulary growth seems to be faster for non-native speakers than for native speakers. The curve for native speakers in another similar domain (travel arrangement) for which we have much more data was similar to the curve for native speakers shown in Fig. 2; in fact, the vocabulary size for this bigger corpus did not reach the size of the non-native corpus at 5600 words until 10,000 word tokens had been seen.

We also looked at trigram perplexity of the data collected in the different pilot experiments measured with respect to a model built on the large travel arrangement data set. Although the test corpora were very small, we found that the corpus collected from non-native speakers using English prompts was very similar in terms of perplexity to the corpus collected from native speakers in the tourist information task. Conversely, the corpus collected from non-native speakers using Japanese prompts showed over 1.5 times the perplexity of the native corpus. This indicates that the character of the two non-native corpora are quite different, and that incorporating the L1-prompted data in training a statistical language model will increase the predictive power of the model with respect to non-native speakers.

6 Discussion

A final question is how many of our observations are L1-dependent. It is true that Japanese speakers show some common patterns in their speech and tend to be very self-conscious about speaking. Japanese is written with a non-roman script and this probably influences both comprehension in the spontaneous tasks and reading accuracy in the read tasks. Japanese is very different from English grammatically, pragmatically, and phonotactically. Many of our observations may not be consistent with observations in collection with native speakers of German, for example. In this respect, though, it is really an ideal case study for the purposes of uncovering all the stumbling blocks we may encounter when designing data collection for non-native speakers. We found that speakers' reading ability was generally much higher than their conversational ability; Byrne's study (1998) found that their lowest skill level speakers had some conversational ability but no reading ability. The important thing to recognize is that the reading level - speaking level correspondence is among the variables that should be evaluated in order

どんな	かっこう	で	行く	べき	か
what_sort_of	appearance	with	go	should	QUES

What should I wear?
Do we need to wear the formal dress or we can wear the casual one?
What kind of clothes do I have to wear for there?
In what kind of dresses should I go there?
Should I oh should I go formal with formal style?
What should I wear to go there?

バス・船・列車	など	の	最終 便	の	時間
bus/boat/train	etc.	GEN	last_trip	GEN	time

What time is the last return train/bus/ferry?
What time is the last train to go back to my house?
What time is the last transportation from there?
Do you know what time is the last bus ships or trains to return?
When does the final bus or ship or train?
What time is the final bus?

子供	割引
child	discount

Is there a children's discount?
Is there any discount for the for child
Do they have a discount for children
When I buy the ticket for children are there any discount
Is there special children cost
How much is the fee for children

Figure 1: Inventive expressions. The Japanese prompt and an English gloss are shown with a sample English response at the top of each series.

to design an effective data collection protocol.

References

Greg Aist et al. 1998. How Effective is Unsupervised Data Collection for Children's Speech Recognition? In *Proceedings of ICSLP*.

Charles Briggs. 1986. *Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research*. Cambridge University Press, Cambridge.

William Byrne et al. 1998. Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English. In *Proceedings of Speech Technology in Language Learning (STiLL)*.

Maxine Eskenazi. 1997. Detection of Foreign Speakers' Pronunciation Errors for Second Language Training - Preliminary Results. In *Proceedings of Eurospeech*.

Goh Kawai and Keikichi Hirose. 1997. A CALL System Using Speech Recognition to Train the Pronunciation of Japanese Long Vowels, the mora nasal and mora obstruents. In *Proceedings of Eurospeech*, Rhodes.

William Labov. 1984. Field methods of the project on linguistic change and variation. In *Language in Use: Readings in Sociolinguistics*, pages 28 – 66. Prentice-Hall.

Leonardo Neumeyer, Horacio Franco, Mitchel Weintraug, and Patti Price. 1998. Automatic Text-independent Pronunciation Scoring of Foreign Language Student Speech. In *Proceedings of ICSLP*.

Silke Witt and Steve Young. 1997. Language Learning Based on Non-Native Speech Recognition. In *Proceedings of Eurospeech*, Rhodes.

Nessa Wolfson. 1976. Speech Events and Natural Speech: Some Implications for Sociolinguistic Methodology. *Language in Society*, 5:188 – 209.