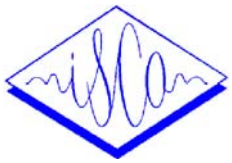

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

# ISSUES IN DATABASE CREATION: RECORDING NEW POPULATIONS, FASTER AND BETTER LABELLING

*M. Eskenazi, C. Hogan, J. Allen, R Frederking*
Language Technologies Institute
Cyert Hall
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, Pa. 15213 USA.
Tel. +1 412 268 3858, FAX: +1 412 268 6298, E-mail: max@cs.cmu.edu



## ABSTRACT

As speech recognition systems become more accurate, they are used for more diverse applications. These applications often involve populations who never used a recogniser before and for whom the standard data for adult male, adult female, or mixed adult speech is not very representative. This paper will deal with issues concerning the collection and processing of data from those new speaker populations and from speakers of different languages. It deals with data collected for various projects, such as the KIDS database [1] and the Diplomat project [2]. It specifically discusses issues related to obtaining quantitatively and qualitatively sufficient amounts of speech from diverse speaker populations. Since the speech of these individuals is very different from the speech collected in the past, we assume that some hand labelling may be necessary and therefore also address the issue of ameliorating the labelling process.

## 1. ADAPTATON TO NEW APPLICATIONS

As speech recognition systems become more accurate, they are ported to more diverse applications. Changing domains involves changes in many levels of processing. Data obtained in the past has varied from large populations of speakers carefully reading relatively small amounts of text (TIMIT), smaller populations reading larger amounts of text in a defined application domain (DARPA RM), heavily constrained, but not read, speech from a relatively small population (ATIS) to more spontaneous speech in a less restrained domain from a fairly small number of speakers (Broadcast News). When a new application is defined, large amounts of speech data typical of that type of variability are collected for training. The speakers have generally been adult natives.

As the data for automatic speech recognizers (ASRs) has changed, each newly-defined hurdle has revealed new data-gathering issues. Some of the issues in Broadcast News concerned obtaining the broadcast signal and choosing a subset of all that is broadcast. Once the signal was recorded, other issues surfaced, such as segmenting the signal into usable chunks.

With new populations of users, such as children, other issues have come up. The information drawn from our new populations will hopefully aid the reader in preparing to deal with yet other populations in the future, and in anticipating issues that have not yet been encountered.

The increase in the amounts of data needed for training requires better processing methodologies. To address part of this issue, we will also discuss a new approach to data labelling.

### 1.1. Description of the projects and their data

The few applications of ASRs that presently have children for users have little or no children's speech data at their disposal. Instead, like Project LISTEN at Carnegie Mellon University [3], they have had to use adult female speech models. In order to furnish more appropriate data, the KIDS database recorded 76 children. Since Project LISTEN aims at helping children learn to read, the data consists of text read aloud. There were 2 populations of speakers. First, a population of good readers (SUM95) was recorded in order to obtain as much speech data as possible. Then, children from a school where reading scores are especially low were recorded (FP) in order to get data representative of local dialect and reading hesitations.

The DIPLOMAT project [4] is designed to test the feasibility of rapid-deployment, wearable speech translation systems. This means developing a machine translation system that performs initial translations at a useful level of quality between a new language and English within a matter of days or weeks, with continual, graceful improvement to a good level of quality over a period of months. A potential use for DIPLOMAT is to allow English-speaking soldiers on peace-keeping missions to interview local residents. So far, Diplomat has worked with Serbo-Croatian, Creole, and Korean.

Since rapid deployment is central to the project, read speech is used. It is faster and less labor-intensive to develop than spontaneous speech. At present, there are 13 speakers for Haitian Creole (hereafter, Creole) (10m, 3f) with 99 to 231 sentences each. For Korean there are 8 speakers (5m., 3f) with 118 to 180 sentences each. Recordings are still underway in both languages.

## 2. NEW SPEAKER POPULATIONS

We group our observations of new populations according assumptions researchers made in the past. We examine how they are no longer valid, and note how we dealt with them.

They are grouped in: speaker competence, quality control of the utterances, language consistency and text processing.

### 2.1. Speaker competence assumptions

We often assume universal competence in the linguistic skills learned in grade school such as reading or punctuating. Yet all students do not do equally well at this, and these skills are not taught everywhere. Examples below concern speech production, but this holds for perception as well (e.g. summarising a story that was just read to you).

#### *2.1.1. People can read aloud*

The BREF database [5] from LIMSI consists of sentences read aloud from the Le Monde newspaper. A pretest was administered to the speakers: ten sentences were sent out by mail and the potential subject was asked to practice reading them aloud. At pretest time, only 35% of the subjects were able to read all ten sentences with less than 3 errors and only 50% could read the sentences with between 4 and 10 errors.

For DIPLOMAT, 55% of the Haitian population is illiterate [7]. And those who can read typically learn to read French, not Creole. To record 23% of the speakers, a Creole-literate Haitian sat next to the speaker and *read* him the sentence, simply having the latter repeat and then record it.

For the KIDS database, we expected some of the children to not be fluent readers, and wanted to record examples of their disfluencies for later modelling. But 25% of the FP children provided us with only a dozen painfully read sentences each.

We learned to question the capability of any population to read aloud. Older subjects, like younger ones, new language populations and any "average" reader should be pretested. Other solutions range from using spontaneous speech instead of read speech, to having the speaker prompted, or using elicitation techniques.

### 2.2. Quality control of the utterances assumption

#### *2.2.1. The spoken form obtained will be consistent*

We often presume that the variants included in the lexicon of the ASR will represent the main variations of the language. In Creole, the goal was to represent Creole speech, yet French is taught in the schools and many speakers will change registers toward French in a formal situation. Consider the example of "Clinton'', borrowed from English. To a speaker of Creole, "Clinton'' will be pronounced either as in English: [klInt@n] or as approximated by the Creole phonology: [klinton], or as in French: [klE~to~] or [klinto~] with nasalization. If the speaker uses the more formal French register for the recording (or changes to it in the middle of the recording!), only the third and fourth pronunciations would be possible.

There is also the minor problem of always finding someone who is fluent in the target language to do the recording, in order to catch any reading flaws.

### 2.3. Language consistency assumptions

#### *2.3.1. The written form will be legible for all the speakers of the language*

Several stages were required to form an acceptable text corpus, legible for all Creole speakers. We first took texts translated into Creole by three native speakers, but there was lack of consensus by all three. Second we chose the translated data provided by the most reliable speaker but editing of this material by other Haitians and a Creole-proficient linguist indicated that the corpus was also unacceptable. Careful revision of existing data proved to be too slow a process (25 sentences/day) to correct all the data quickly. Finally, we asked five native speakers about the quality of the Creole section of the "Haiti Progres" newspaper. They unanimously agreed on the high quality of the sentences here and demonstrated the ability to read sentences aloud without difficulty. This became the base text Compared with the sentences used in the first two attempts (when speakers stopped recording frequently to comment on the text), the final corpus provoked very few interruptions. Acquiring data from reputable outside resources, rather than depending on internally-translated materials for data, provided text that was legible for all.

### 2.4. Text processing assumptions

#### 2.4.1. *The written form obtained will be consistent*

Of the eleven orthographic systems proposed for Creole through 1980, three have been the most used in educational and literacy programs. The divergent orthographic systems are the cause of inconsistencies in the text when it is revised by several native speakers. Examples of the variants are given in the table below.

| IPA | Institut Pedagogique National | Faublas-Pressoir | McConnel-Laubach |
|---|---|---|---|
| [a~] | an | an | a^ |
| [an] | an | a-n | an |
| [a~n] | ann | ann | a^n |
| [õ] | on | on | o^ |
| [on] | o`n | onn | o^n |
| [õn] | onn | o`n | o`n |

**Table 1. Examples of variations of Creole orthography in three of the systems used.**

The data in Figure 1 was collected by grouping words into classes according to the number of spelling variations each word exhibited. For Creole this was done automatically, by grouping words according to common variations in spelling, such as "r" for " w" or "e" for "e`". For English, this was done by identifying misspellings (as noted by the UNIX spell program) with the correct spellings, and matching variants of proper names. Data were collected from a bilingual corpus built from English newswire stories and their Creole translations. The Creole side of the corpus had 186644 instances (tokens) of 11599 distinct words (types). The

English side had 185133 instances representing 12871 distinct words.

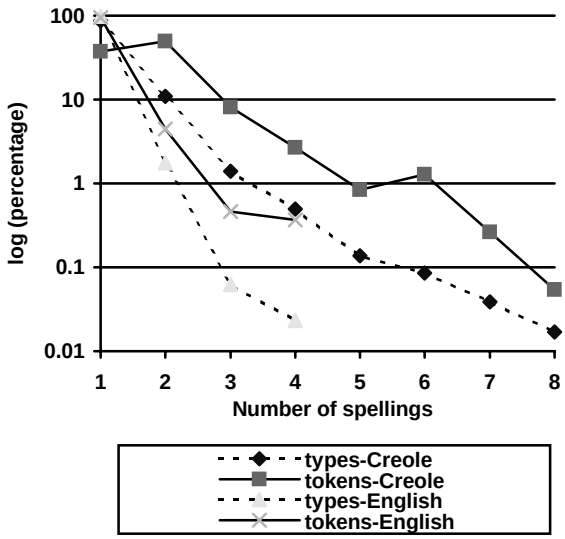


**Figure 1. Creole vs. English: relative sizes of classes of spelling variants. Percentage of words with given number of variants - log scale.**

Figure 1 is a histogram, where the x-coordinate identifies words with a given number of spelling variants. For Creole, no word had more than 8 variants, while for English, no word had more than 4. Y-coordinates are given as a percentage, plotted on a log scale. For example, if we compare type counts, English words with only one spelling variant accounted for approximately 98% of all English words, whereas in Creole, this class of words accounts for only 86% of all Creole words.

The data confirm the hypothesis that a standardised language like English not only has fewer spelling variants, but also that these variants are less frequent. Thus, English word classes with more than one variant are always a much smaller part of the data than in Creole. Of particular interest are the graphs of word instances (tokens): in English, the majority of word occurrences (94%) are from a class that does not vary (1-variant). In Creole, on the other hand, not only do the invariant words comprise a smaller percentage of the whole (37%) but words with two variants actually occur more frequently (50%) than any other class. Thus, more than half the words in Creole encountered in our corpus systematically vary in spelling.

### 2.5. Comments

We learned that we cannot assume everyone can achieve a cognitive task with the same level of success, and that all languages are not equally consistent in their written and spoken forms as English is. We have also learned that the apparent proximity of two languages and cultures (Creole seemed closer to American English than Korean) is not a valid measure of the ease that we will have in acquiring and processing data from a new language. Rather, whether the language has a standardised spelling, and whether the target language comes from an industrialised society or an emergent one, seem to be the predominant factors in processing new data.

## 3. LABELLING

When processing data from speakers similar to those already represented in an ASR, it is often sufficient to have a word-level transcription and then feed the transcription and the signal to the recognizer in forced alignment mode. If read speech is being collected, the text on the screen is often sufficient - no manual transcription is needed. However, when the population differs in the characteristics of its speech from what we already have, for example for foreign speakers and children, then some form of phone labelling is usually necessary. Labelling is often a one-pass operation. This is partially due to the fact that several passes involving more man-hours would be more costly. One exception was the work on the TIMIT corpus [6] where several labellers transcribed the speech and the final transcription represented agreement amongst the transcribers.

Due to the increase in the amounts of data that we need to process, problems arise concerning the quality of the labels obtained. We have broken the labelling process down into a series of separate cognitive tasks to address this.

In the KIDS database project, the children's speech could theoretically have been labelled using forced alignment of the speech signal against the text shown on the screen. However, the young readers had many errors and false starts. After processing we found that only 1067 out of 5414 utterances exactly followed the text and could be forced aligned. For all of the other sentences, we decided to phonemically label only the part of the speech signal that did not correspond to the text.

The task was divided in two parts:

1) listen to each sentence and decide if it follows the text exactly (can be forced aligned)
    a) YES -> automatic labelling (bin1)
    b) NO -> make text file pointing to where the differences are, send to hand labellers (bin2)

2) take the bin2 sentences and phonemically label them only at the points indicated in the text file.

We then assessed the quality of the labels and the consistency of the labellers. Although we have no precise numbers to show, we noted that it took the labellers about 50 % less than twice the time we had estimated only one pass would take. This process could will not gain speed in labelling, but rather, insure better quality.

### 3.1. Assessing the labels

After phonemic labelling was complete, we performed forced alignment on the bin2 utterances. Of the 4347 bin2 utterances, only 80 failed to be aligned. Of these, closer examinations showed that noise over speech was the probable cause of failure for 60 utterances.

We conclude that the phonemic labelling was of sufficient quality to enable automatic labelling.

### 3.2. Assessing the labellers

We cannot compare the individual labellers' actual performances to what they would have done if they only used one pass. But we can assess across the five labellers.

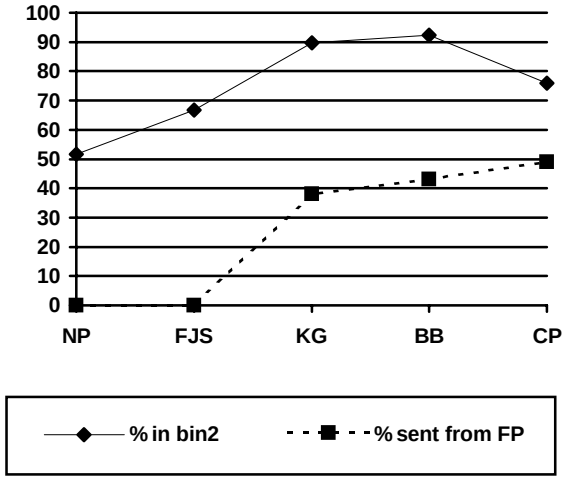


**Figure 2. Percent sentences in bin2 compared to percent sentences from FP each labeller saw.**

First, we tried to see if all the labellers put the same amounts of sentences into bin2. We reasoned that the two separate populations (SUM95 and FP) should have different amounts of bin2 data due to different reading levels (more for FP). In Figure 2, the labellers are plotted on the x axis, the percentage of sentences they saw that were put into bin2 on the y axis. The solid line represents all the sentences put in bin2 and the dotted line, the FP sentences alone. There was not clear relation between the two lines.

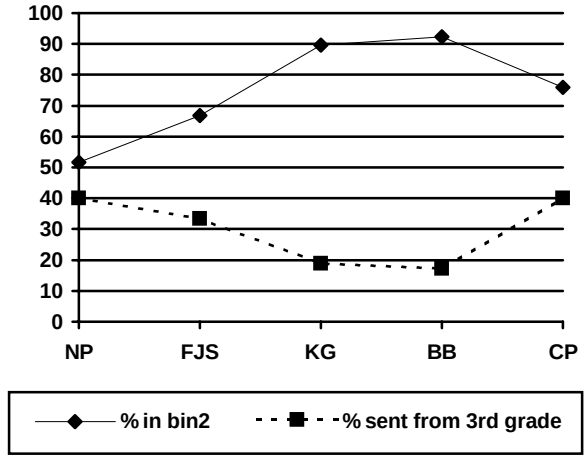


**Figure 3. Comparison of percentage of data put in bin2 to the percentage of data each labeller had from grade 3 students.**

Then, remembering that the older speakers (3rd grade) had less data in bin2 per speaker in general, we decided to compare labellers according to the amount of 3rd grader data that each had seen (Figure 3). The two sets of lines are symmetrical, and showing that the labellers are consistent *and* that grade, rather than school background makes the difference in reading well here.

## 4. CONCLUSIONS

We have presented examples of unforeseen differences encountered when dealing with new speaker populations. Our conclusions encourage the database designer to question and to test all linguistic capacities the speaker will need to call on during recording.

We also show evidence that hand labelling of large corpora should be divided into several passes, each corresponding to a different cognitive task in order to obtain consistent labelling. It is now up to the database designer to decide if the probable increase in quality is worth the slight increase in processing time.

*This research is sponsored in part by the National Science Foundation grant no. IRI-9528984 and Defense Advanced Research Projects Agency grant no. F33615-93-1330.*
*Views and conclusions in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Government.*

## 5. REFERENCES

[1] M. Eskenazi, "KIDS: A Database of Childen's Speech" , in Proc 3rd joint Meeting: Acoustical Societies of America and Japan, Honolulu, 1996.

[2] R. Frederking and R, Brown, "*The Pangloss-Lite Machine Translation System"*, in Proc. Conference of the Association for Machine Translation in the Americas (AMTA), Montreal, 1996.

[3] J. Mostow, S. Roth, A. Hauptmann, M. Kane, " A Prototype Reading Coach that Listens", Proc. 12th National Conference on Artificial Intelligence, Seattle, 1994.

[4] R. Frederking, A. Rudnicky and C. Hogan, "*Interactive Speech Translation in the DIPLOMAT Project*", Working notes of the Spoken Language Translation workshop at ACL 97, Madric, 1997. (To appear)

[5] L. Lamel, JL. Gauvain, M. Eskenazi, "BREF, a Large Vocabulary Spoken Corpus for French, Proc. Eurospeech93, Genoa, 1991, p. 505-508.

[6] L. Lamel, R. Kassel, S. Seneff, "Speech Database Development: Design and analysis of the acoustic-phonetic corpus", Proc. DARPA Speech Recognition Workshop, 1986.

[7] Central Intelligence Agency, "The World Factbook", Central Intelligence Agency, Washington, D.C., 1995.

.