THE OGI 22 LANGUAGE TELEPHONE SPEECH CORPUS

T. Lander *R. A. Cole* *B.T.Oshika* *M. Noel*

email: tlander@cse.ogi.edu http://www.cse.ogi.edu/CSLU/
Center for Spoken Language Understanding, Oregon Graduate Institute
P. O. Box 91000, Portland, Oregon 97291-1000 USA

ABSTRACT

The 22 Language Telephone Speech Corpus is the newest effort in CSLU multi-language corpus development. Since November of 1994 we have been collecting calls from speakers of 22 languages. The completed corpus will contain at least 200 speakers per language. All calls are verified by native speakers to insure that callers followed instructions. In addition a subset of the calls have been transcribed by native speakers of the different languages. Conventions are being developed for phonetic transcription of a portion of the calls in each language. This corpus is useful for language identification research, as well as research into spoken language systems.

1. DATA ACQUISITION

Calls were collected using a 24 channel T1 line connected to three LINKON FC3000 Communication boards. Speakers are solicited through advertisements posted on various net sites and in foreign language newspapers. All callers are offered an award for participating, either a coupon redeemable at a national food chain or entry into a drawing. Each language is associated with a toll free phone number accessible from anywhere within the United States.

2. CHOICE OF LANGUAGES

This corpus includes the following languages: Arabic, Cantonese, Czech, English, Farsi, French, German, Hindi, Hungarian, Italian, Japanese, Korean, Malay, Mandarin, Polish, Portuguese, Russian, Spanish, Swahili, Swedish, Vietnamese, and Tamil. These languages were selected to represent large speaker populations and also reflect diversity with respect to historical relationships and phonological structure.

3. DATA COLLECTION PROTOCOL

A protocol was developed in English and then translated and recorded by native speakers for the other 21 languages. The protocol includes prompts which elicit a total of two to three minutes of speech from each caller. The responses fall into three categories: (a) Requests for specific information; (b) Continuous speech on selected topics; (c) Extemporaneous speech.

The prompts used to elicit responses in each category are as follows:

- **Requests for specific information**

  The following prompts were designed to elicit sufficient speech in a structured way to determine the dialect of the speaker and idiosyncratic speech characteristics the caller might manifest. They also elicited forms of numbers because of the usefulness of number recognition in many applications.

  1. What is your native language?
  2. What language do you speak most of the time?
  3. What language do you speak at home?
  4. What other languages do you speak and understand?
  5. How old are you?
  6. What is your date of birth?
  7. Are you male or female?
  8. How long have you been in the United States?
  9. What city and state did you spend most of your childhood?
  10. What is your zip code?

11. What area code are you calling from?
12. What day of the week is today?
13. What time is it?
14. Please say a familiar telephone number.
15. How would you ask someone if they speak English?
16. Please say the greeting you usually use when answering the phone in English.

- **Continuous speech on selected topics**

  Each of the following questions elicited narratives of up to ten seconds of continuous speech on specified topics.

  1. Please describe the route you take to work or to the store.
  2. Tell us something that you like about your home town.
  3. Tell us about the climate in your home town.
  4. Describe the room you are calling from.
  5. Describe your most recent meal.

- **Extemporaneous speech**

  The final two questions were designed to elicit longer segments (up to 20 or 60 seconds) of extemporaneous speech. The first of these asked the caller to speak in English for twenty seconds. Callers had the option to press a button to continue if they did not feel comfortable speaking in English. Responsiveness to this question varied, but a significant portion of the callers did speak for the entire time. The purpose of this question was to provide data for research into phonological characteristics of non native speakers of English. This question was not a part of the English Protocol.

  The second question asked the caller to speak for one minute on any topic. Suggestions for possible topics were given, but callers were not restricted to those topics. They were given ten seconds to plan what they were going to say. Most callers did not experience difficulty in speaking for the entire time.

4. CURRENT STATUS

Table 1: *Number of complete calls per language, Number of calls verified, Number of stories transcribed at the word level*

language	# Calls	# Ver	# Trans
Arabic	27		
Cantonese	50		
Czech	38	38	38
English	526	21	
Farsi	25		
French	101	101	52
German	72	50	28
Hindi	173	25	
Hungarian	91		
Italian	44		
Japanese	98	74	52
Korean	30		
Malay	20		
Mandarin	100	34	33
Polish	30		
Portuguese	497	84	
Russian	76		
Spanish	329	300	51
Swahili	33		
Swedish	40		
Vietnamese	68	64	50
Tamil	231	220	11
Total	2,699	1,011	315

To date, we have received an estimated 2,699 complete calls. Calls are considered complete if there is sufficient response to one of the last two questions. A person who is not necessarily a speaker of the language listened to these responses to verify completeness. See Table 1 for the total number of complete calls received and the number of calls or files that have been verified or transcribed in each language to date.

5. VERIFICATION

Two native listeners in each language verify each call. They resolve differences in order to ensure reliability of the verification procedure. The listeners are hired part-time and are trained in the verification process.

Verification involves making judgments as to whether each utterance is a complete answer to the question and whether the caller has followed instructions. In addition, judgments are made for each utterance regarding the presence of background noise, breath noise, cut off speech, or speech that is in a language other than that which the caller was asked to speak. The verifiers can hear the files as many times as necessary.

In addition to the comments on the appropriateness of the responses, other characteristics are noted: (i) is speaker native (ii) describe dialect if not standard (iii) describe any non-native accent (iv) describe any background speech (v) describe any environmental noise (vi) gender (male, female, unknown) (vii) age (child, adult) (viii) connection quality (poor, average, good); and (ix) speaker intelligibility (poor, good).

As a part of this process, native talkers compare the speech of each caller to a named standard dialect. The standard dialect is one that was chosen based mainly on number of speakers or association with a major metropolitan area. Verifiers describe dialects differing from the standard by citing regional or salient, speech characteristics.

For each call a log file is created containing the above judgments.

6. ORTHOGRAPHIC TRANSCRIPTIONS

Native speakers have produced some non-time aligned word transcriptions of caller responses using standard (romanized) orthographic conventions. Orthographic transcriptions allow easy access to the data at a lexical level, and include notations for non-speech information such as pauses and background noise.

There are eight languages for which we are developing romanizations: Arabic, Cantonese, Farsi, Hindi, Japanese, Korean, Mandarin, and Tamil. Where possible we use standard romanizations, such as Pin Yin for Mandarin Chinese.

For languages whose orthographies may contain diacritics not found on a standard English language keyboard, there is a mapping between the diacritic and an ASCII symbol such that the value of the original diacritic can be unambiguously recovered. For example, the cedilla in French is represented by a comma, as in as in *franc,ais*, "French." Orthographic conventions appear in an updated version of "The CSLU Labeling Guide." [2]

The goal is to have two labelers transcribe a subset of the data in common in order to ensure reliability. The duplicate transcriptions are useful for determining labeler reliability. Informal spot checking of orthographic transcriptions during the transcriber training period indicates a reasonably high level of consistency.

7. PHONETIC TRANSCRIPTIONS

7.1. Transcription Conventions

Phoneticians are developing broad phonetic transcription conventions for each language in the corpus. A portion of the story responses are hand labeled at the broad phonetic level in order to develop these conventions. All phonetic transcriptions will be done in Worldbet, a phonetic alphabet developed by James Hieronymus to capture the sounds in all of the world's languages using ASCII symbols [3]. Segmentation principles and label usage have been described in [2].

7.2. Reliability

As a part of the process of convention development, a subset of the newly collected data has been transcribed at the phonetic level by trained phoneticians. Approximately three minutes in each of six languages have been transcribed: English, German, French, Vietnamese, Japanese and Spanish. We have plans to label similar amounts of speech in each of the 22 languages.

Results of a formal study on inter-labeler reliability on these data were presented in [4] and summarized in Table 2 In this preliminary study, labeler agreement across languages ranged from 41 percent with detailed label to label comparisons to 91 percent when less fine comparisons were made.

Inter-transcriber agreement was measured in terms of the number of substitutions, deletions and insertions required to map one transcription to another. The "reference" transcription was chosen arbitrarily.

The average accuracy for the set of files in each language was computed using the average number of reference segments, substitutions, insertions and deletions over both of the files.

Column one of Table 2 displays agreement in each language when the original label files were compared. Column two displays agreement when diacritics were stripped from the phonetic labels. Column three displays agreement when all phonetic labels were mapped to broad categories. The number in parenthesis in each box refers to the number of labels in the reference string. As expected, agreement improves as labels become more general.

Table 2: *Percent Average transcriber agreement at three levels of precision: 1) original labels 2) diacritic stripped 3) broad category (number of reference labels appears in parenthesis)*

	1	**2**	**3**
Eng	55 (143)	67 (124)	83 (143)
Fre	59(119)	60 (97)	83 (119)
Ger	41 (122)	52 (105)	74 (122)
Jap	72 (143)	77 (118)	91 (143)
Span	71 (107)	78 (94)	86 (107)
Viet	60 (104)	68 (84)	84 (104)
ave	59	67	84

8. USAGE

the OGI 22 Language Telephone Speech Corpus is an extension of the OGI Multi-language Telephone Speech Database, which is the NIST standard for evaluation in Automatic Language Identification. The additional 12 languages in the corpus provide a richer phonological variety for language identification research, and the increased number of calls per language will yield more data for training and testing.

9. AVAILABILITY

When complete, the corpus will be placed in the public domain. We expect to complete data collection by April 1996. The corpus will be donated to the National Institute for Standards and Technology in the USA and to the Linguistic Data Consortium. It will also be distributed free of charge throughout the CSLU. It is CSLU's desire to provide this corpus to the research community as a resource for the study of different languages and the development and evaluation of speech recognition and language identification algorithms.

To obtain this corpus, please contact Mike Noel at noel@cse.ogi.edu.

10. ACKNOWLEDGMENTS

We are indebted to the Linguistic Data Consortium for providing support for this project. We would also like to acknowledge Dr. James Hieronymus for his contribution of Worldbet.[3]

REFERENCES

[1] B. Comrie. *The World's Major Languages.* Oxford, Oxford University Press, 1990.
[2] T. Lander, S.T. Metzler. *The CSLU Labeling Guide.* CSLU, Oregon, February, 1994.
[3] J.L. Hieronymus. *ASCII phonetic symbols for the world's languages: Worldbet.* AT&T Bell Laboratories Technical Memo, 1994.
[4] T. Lander, B.T. Oshika, R.A. Cole, and M. Fanty. Multi-language Speech Database: Creation and Phonetic Labeling Agreement. *Proceedings of the International Congress of Phonetic Science.* Stockholm Sweden, August 1995.
[5] Cole, R. A., M. Fanty, M. Noel and T. Lander, "Telephone Speech Corpus Development at CSLU," Proceedings of the 1994 International Conference on Spoken Language Processing, Yokohama, Japan, Sept. 18-22, 1994. pp 2131-2134.
[6] Y.K. Muthusamy R.A. Cole, and B.T. Oshika, "The OGI Multi-language Telephone Speech Corpus, *The International Conference on Spoken Language Proceedings,* Banff, Canada, Oct 1992. pp 895-898.