Corpus Development Activities at the Center for Spoken Language Understanding

Ron Cole, Mike Noel, Daniel C. Burnett, Mark Fanty,
Terri Lander, Beatrice Oshika, Stephen Sutton

Center for Spoken Language Understanding
Oregon Graduate Institute of Science and Technology
Portland, Oregon 97291

ABSTRACT

This paper describes eight telephone-speech corpora at various stages of development at the Center for Spoken Language Understanding. For each corpus, we describe data collection procedures, methods of soliciting callers, protocol used to collect the data, transcriptions that accompany the speech data, and the expected release date. The corpora are available at no charge to academic institutions.

1. INTRODUCTION

The Center for Spoken Language Understanding (CSLU) collects and transcribes telephone-speech data to enable research activities at CSLU and elsewhere. Corpus development activities are performed by four full-time staff, aided by graduate students and part-time employees. In 1994, we anticipate collecting and transcribing speech from 10,000 callers in twenty languages. Corpus development activities are supported by industrial memberships and research grants.

Corpus development activities at CSLU include: (a) collecting telephone speech data in different languages; (b) transcribing speech at word and phonetic levels; (c) developing and documenting transcription conventions for each level; (d) measuring the level of agreement among transcribers; (e) developing interactive speech tools for labeling; (f) distributing the speech corpora to academic institutions free of charge; and (g) placing speech tools and labeling conventions in the public domain for use by others.

In this section, we present some general information about our corpus development activities. In the following sections, we will describe individual corpora.

Data Collection. telephone-speech data are collected over analog and digital telephone lines. Prior to November, 1993, speech data were collected over analog lines using several Gradient Technology Desklabs. Since November, 1993, the majority of our data has been collected using a 24 channel T1 line connected to three LINKON FC3000 Communication Boards. We are also using an Apple GeoPort Telecom Adapter connected to a Macintosh Quadra A/V to collect analog speech data for one of the corpora to be described.

Transcription. Each call is processed by one or more listeners. Calls are verified to determine that the caller followed instructions and in some cases, transcribed at some level.

Transcription of corpora occurs at three different levels: non-time-aligned word level, time-aligned word level, and time aligned phonetic level. Non-time-aligned word level transcription involves producing an orthographic representation of the utterance, including indications of extra-speech events such as breathes or lip smacks, without time markings. Time-aligned word level transcription provides the same orthographic transcription augmented with time alignment markings. Time-aligned phonetic transcription involves aligning phonetic symbols to the acoustic signal.

A precise description of the conventions used for all levels of labeling, including a complete list of all phonetic labels for each language, is presented in the CSLU conventions document[1]

Transcription Reliability. We are conducting experiments to determine the level of agreement among labelers. In these experiments, CSLU staff and professional phoneticians are using Worldbet [5] to transcribe the same intervals of speech. Initial results for English indicate overall agreement of approximately 80% across all labels, ranging from approximately 70% for vowels to greater than 90% for stops and nasals.

Speech Tools. The OGI Speech Tools support data manipulation, analysis and display[2]. All corpus development activities are performed using these tools. They were developed at CSLU, then made portable and documented for distribution with support from NSF. The tools have been made available to the research community through anonymous ftp.

2. CORPORA

The first three corpora described in this section are considered to be complete and are now available from CSLU. They were collected over an analog telephone line using a Gradient Technology Desklab connected via the SCSI port to a workstation. The data were digitized at 8000 samples per second with a 14 bit resolution. All data are stored in the NIST wav file format, some with MIT shortpack compression. The remaining corpora are under development and estimated release dates are provided for each.

2.1. Spelled and Spoken Names Corpus

The Spelled and Spoken Names Corpus [3] contains utterances from 3667 calls. Callers were solicited through computer newsgroups and a public relations campaign initiated by OGI. The majority of callers were from the Pacific Northwest. The proportion of male to female callers is 1.15 : 1.

The goal was to collect samples of spoken English letters and spoken words to support a research project funded by U S WEST. Callers received the following prompts:

- What city are you calling from?
- What is your last name?
- Please spell your last name.
- Please spell your last name with short pauses between letters.
- Does your last name contain the letter A as in apple?
- What is your first name?
- Please spell your first name, with short pauses between letters.
- What city and state did you grow up in?
- We will now ask you to say the alphabet. We need you to pause briefly between letters, like this: A B C D E F G. You may hang up when you are finished. Please begin speaking now.
- Would you like to receive more information about the results of this project?
- If you would like more information about this project, please leave your name and address at the tone.

Documentation of the Spelled and Spoken Name Corpus includes a speaker-by-speaker log file containing orthographic transcriptions of each utterance. Each utterance was transcribed by two separate listeners. The log also contains the global judgments of gender, age, connection quality, accent, and intelligibility. In addition, occurrences of extraneous speech, environmental noise, excessive breath, or line noise are indicated in the log file for each utterance.

A subset of the data was transcribed at the time-aligned phonetic level. The utterances were labeled by hand then labels and time-alignments with the speech spectrogram were verified by an expert spectrogram reader. The subsets of phonetically labeled utterances available to date are as follows:

Type	Number
alphabet	100
hometown	1359
callfrom	693
say first name	100
say last name	101
spell last name with pause	300

2.2. Enhanced OGI Multi-Language Corpus

The OGI Multi-Language Telephone-Speech Corpus [4] consists of telephone-speech from 10 languages: English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The initial corpus included 900 calls—90 calls for each language.

Callers were solicited through computer newsgroups. Each caller was asked to respond to the following prompts:

- What is your native language?
- What language do you speak most of the time?
- Please recite the seven days of the week.
- Please say the numbers zero through ten.
- Tell us something that you like about your hometown.
- Tell us about the climate in your hometown.
- Describe the room that you are calling from.
- Describe your most recent meal.

In addition, unconstrained speech was obtained by asking callers to speak for 1 minute on any topic of their choice.

Each utterance was listened to by a native speaker of the language to verify that the caller responded appropriately. The native speaker also made judgments concerning the caller's gender, the caller's age, and the line quality.

The enhanced corpus is augmented with: (a) 200 Hindi calls; (b) speech files that were collected during the original collection but were not included in the original distribution; and (c) time-aligned phonetic transcriptions of over five hours of speech (up to 50 sec per call) in six languages—English, Japanese, German, Spanish, Hindi, and Mandarin. For the broad phonetic transcription, we have adopted the Worldbet labeling scheme, a set of orthographic symbols for multi-language transcription that correspond to IPA symbols [5]. The rationale for using Worldbet and the inventory of symbols for each language is provided in [1].

2.3. Stories Corpus

Collection for the OGI Multi-Language Corpus produced additional calls from English speakers not included in the Multi-Language Corpus. The Stories Corpus consists of up to 50 sec of spontaneous speech (hereafter "stories") from 692 English calls. All 692 calls have been transcribed at the non-time-aligned word level, 300 at the time-aligned word level, and 200 at the time-aligned phonetic level.

2.4. Twenty-one Language Corpus

CSLU plans to collect and verify calls from at least 200 fluent native speakers in 21 languages—Eastern Arabic, Cantonese, Czech, Farsi, French, German, Hindi, Hungarian, Japanese, Korean, Malay, Mandarin, Italian, Polish, Portuguese, Russian, Spanish, Swedish, Swahili, Tamil, and Vietnamese. Verification and global judgments will be performed by native speakers.

The following is the English version of the protocol for the twenty-one language corpus. The protocol will be presented to the caller in their language.

- **Thank you for calling the Oregon Graduate Institute language database. We are currently recording speech in ⟨language⟩. We are studying the different languages of the world. To do this, we need to record samples of speech from fluent speakers of ⟨language⟩. Please respond to the following questions and instructions in ⟨language⟩**

only. This will take about 7 minutes. Please wait for the beep before speaking.

- What is your native language?
- What language do you speak most of the time?
- What language do you speak at home?
- What other languages do you speak and understand?
- How old are you?
- What is your date of birth?
- Are you male or female?
- How long have you been in the United States?
- What city and state did you spend most of your childhood?
- What is your zipcode?
- What area code are you calling from?
- What day is today?
- What time is it?
- Say a familiar telephone number.
- How would you ask someone if they speak ⟨language⟩?
- Give us the greeting you usually use when answering the phone.
- **For each of the following descriptions, we will record the first ten seconds of your answer. Begin speaking at the beep. A second beep will indicate when we have finished recording your answer to each question.**
- Describe the route you take to work or to the store.
- Tell us something that you like about your hometown.
- Tell us about the climate in your hometown.
- Describe the room you are calling from.
- Describe your most recent meal.
- **We now want you to talk for a longer period of time. We do not care what you say as long as you keep talking. You can tell us anything about yourself, your hobbies and interests, the city that you live in, and the sports that you like. Or you can make up a story, tell a fairy-tale or recite a poem. You will have 1 minute to speak. We will now give you 10 seconds to think about what to say. Please do not read anything, we would prefer you make something up.**
- Please begin talking at the beep. You will hear a second beep when you have 10 seconds left.
- For the last question, we would like you to tell us something about yourself in English. If you do not speak English, you may push any button on your phone, or simply wait for 20 seconds. At the beep, please tell us something about yourself in English.
- If you are calling from a touch tone phone, please push the number 2 button.
- Would you like to receive a gift certificate for McDonalds or for TCBY frozen yogurt?
- Thank you for your participation. If you would like a gift certificate please leave your name, address, and gift certificate selection. Your name and address will be kept confidential.

To date, the prompts for several of the languages have been recorded by native speakers. We expect to begin collection for five languages in March 1994 and then will add five more languages every two weeks until the collection is finished. The expected completion data is yet to be determined.

2.5. English Census Corpus

In conjunction with the U.S. Bureau of the Census, CSLU is collecting data to develop a prototype automated census system. Callers were solicited by the Census Bureau; a memorandum was sent to regional offices asking Census Bureau employees, their family members and family friends to call an 800 number on a voluntary basis to provide speech data for the study. A different 800 number was provided for each city. The cities are Dallas, Chicago, Boston, Charlotte, Atlanta, Philadelphia, Denver, Kansas City, Detroit, and Seattle.

Two protocols were used that differed in the wording of some of the prompts. Each protocol was recorded by both male and female speakers. In addition, male and female synthesized voices were used. Incoming calls were assigned to the eight conditions (prompt X gender X source) in rotation.

An interesting feature of the data collection was the use of automatic recognition to control the protocol. Recognition of "yes," "no," "other," and "American Indian" was performed at certain decision points to determine subsequent prompts. This is illustrated in the following protocol:

- **Thank you for calling the OGI census project. We appreciate your help. The goal of this study is to determine the feasibility of using a computerized questionnaire for the Year 2000 Census. This research is sponsored by the United States Census Bureau. The answers you give to the following questions will be kept confidential. Afterwards we will ask you some questions to help us evaluate this questionnaire. It will take approximately four minutes to complete. Please wait for the tone before answering each question.**
- Please say your first name.
- Please spell your first name.
- Please say your last name.
- Please spell your last name.
- Please say your middle initial. If you have no middle initial, say "none".
- What is your sex, female or male?
- We will now ask about your marital status. Have you ever been married? Please say yes or no.
- *(if yes, then)* Which one of the following options best describes your current marital status: now married, widowed, divorced, or separated?

- We will now ask about your date of birth. What month were you born?
- What day of the month?
- What year?
- We will now ask about your origin. Are you of Spanish or Hispanic origin? Please say yes or no.
- *(if yes then)* Are you of Mexican, Mexican-American or Chicano origin? Please say yes or no.
- *(if no then)* Are you of Puerto Rican origin?
- *(if no then)* Are you of Cuban origin?
- *(if no then)* Please say what other Spanish or Hispanic group is your origin.
- Please spell that.
- We will now ask about your race. Are you: White, Black or Negro, American Indian, Eskimo, Aleut, or other?
- *(if American Indian, then)* What is the name of your tribe?
- Please spell that.
- *(if other, then)* Okay. Are you: Chinese, Japanese, Asian Indian, Korean, Vietnamese, or other?
- *(if other, then)* Okay. Are you: Filipino, Hawaiian, Samoan, Guamanian, or other?
- *(if other, then)* Please say the name of your race.
- Please spell that.
- Is that the name of an Asian or Pacific Islander race?
- Do you have a telephone at home? Please say yes or no.
- *(if yes, then)* Please say your home telephone number, area code first.
- Finally, we'd like some additional information to help us with our study. What is your native language?
- In what city and state did you spend most of your childhood?
- Are you a Census Bureau employee?
- **This concludes the questionnaire portion. We will now ask you some questions to help us evaluate this questionnaire.**
- Would you be willing to provide census information using a questionnaire of this type over the telephone?
- In this questionnaire, we asked about your name, sex, marital status, date of birth, origin, race and telephone number. Please tell us about any questions you found unclear or poorly worded.
- What, if anything, did you like about this questionnaire?
- What, if anything, do you suggest we do to improve this questionnaire?
- We would like to hear any further comments you may have. You may begin speaking at the tone. When you're through, if you would like a gift certificate to either Baskin Robbins, TCBY Yogurt, B. Dalton Books, McDonald's, or Blockbuster Video, please say which one and leave your mailing address. Thank you for your help.

Each call will be transcribed at the time-aligned word level, including indications of filled pauses and other non-speech events. Each utterance will also be assigned a behavior code which characterizes the usability of the response. The behavior codes are described in the following table.

Code	Full Name	Meaning
AA1	Adequate Answer 1: concise	Answer is concise and responsive.
AA2	Adequate Answer 2: usable	Answer is usable but not concise
AA3	Adequate Answer 3: responsive	Answer is responsive but not usable
QA	Qualified Answer	An adequate answer in which respondent expresses uncertainty.
IA1	Inadequate Answer 1: unresponsive	Answer does not seem to be responsive
IA2	Inadequate Answer 2: says nothing	Respondent says nothing at all (may have hung up, or may be lurking).
RC	Request for Clarification	A request for clarification as to the meaning of a concept of survey question. Not used for respondent asking for a repeat due to background noise, etc.
IN	Interruption	Respondent interrupts the speaking of the question. This code implies a second code to account for the content of the interruption.
DK	Don't Know	"I don't know" or any other equivalent formulation.
RF	Refusal	Respondent refuses to answer.
O	Other respondent behavior	Respondent behavior not captured in codes listed above. Also include request for repetition based on not hearing the question.

We are in the process of transcribing the calls that have been collected. We expect that the transcriptions will be completed and the corpus ready for distribution by September 1st, 1994.

2.6. Cellular Words, Numbers and Alphabet Corpus

This corpus will consist of up to 600 calls made from cellular phones. Each caller answers nine questions, says words that might be used in voice messaging applications, says a familiar

phone number, and recites the letters of the English alphabet. Callers are being provided by a private company who helped fund the data collection.

The corpus is being collected using the Gradient Technology Desklab over an analog line. Non-time-aligned word level transcriptions are being produced.

The protocol for the corpus is:

- Are you calling from a cellular phone?
- If you happen to know if you are calling from an analog or digital phone, please say which one.
- Are you using a speaker phone?
- What is your native language?
- Where were you born?
- Where did you spend your childhood?
- What is the month day and year of your birth?
- Please say your name.
- Please say the name of the company or organization you are with.
- **We will now say a set of words, and would like you to repeat each word after you hear it. The words that you speak are intended to be commands to a voice processing system. When you say each command, try to imagine that you are telling the system what to do.**
- *The caller was prompted for the following words one at a time. Each word was presented in the carrier phrase "Say ____ now".*

  Cancel, Change Greeting, Continue, Copy, Erase, Help, Listen, No, Operator, Pause, Replay, Rerecord, Reply, Resume, Review, Save, Send copy, Yes, Add, Dial , Call, Edit, Callback, Change, Delete, Phonebook, Beginning, Choices, End, Directory assistance, Customer support, Next, Repeat, Replay message, Return call, Skip, Tutorial, Customer care, Verify, Scan, Messages, Message, List, Rewind, Fax, Voice, Print.
- Please say a familiar phone number, one digit at a time.
- We would now like you to recite the English alphabet with a brief pause between letters, like this: A B C D E. Please hang up when you are finished. Thanks again.

Currently, approximately 300 calls have been collected and transcribed. We estimate that the corpus will be ready for distribution May 1994.

2.7. Words, Numbers and Phrases Corpus

With support from Apple Computer, CSLU is collecting both analog and digital speech data for utterances related to voice messaging and voice control of computer applications. Callers are being provided both by Apple Computer and by CSLU through newspaper advertisements.

The protocol consists of two questions to help determine the caller's language background, followed by instructions to repeat 35 words or phrases given in the prompt. To increase the usefulness of the corpus, several sub-vocabularies, including first names, last names, digits, numbers and days of the week were inserted into the prompts. For example, the phrase "phone ⟨first name⟩" is expanded to 50 different phrases using 50 common first names.

There are about 350 different phrases that will be recorded from different speakers.

The goal is to collect 1000 speakers using an Apple Macintosh Quadra A/V and 2000 speakers on the digital T1 system using the LINKON setup.

The protocol is as follows:

- **Thank you for calling the Center for Spoken Language Understanding speech data base. We appreciate your willingness to participate in our study. This research is directly related to developing better human computer interaction through the use of voice control. During this call we will be asking you to answer questions and repeat phrases. After each prompt please wait for the beep before responding.**
- First we would like to ask a couple of questions to help us characterize your speaking patterns. What is your native language?
- In what city or state did you spend most of your childhood?
- **For the rest of this call we will say a phrase and ask you to repeat it. For example, we would say "read this text" and you would respond by saying "read this text". Please say the phrase as if you were giving a command to a computer.**
- play previous message again
- cancel my ten AM appointment
- make a meeting for today
- what is my street address
- quit
- forward this message to my wife
- set-up a call with ⟨firstname⟩ and ⟨firstname⟩
- conference call ⟨lastname⟩ and ⟨lastname⟩
- who is at work
- stop
- what is the area code for this state
- add my son to the phone book
- remove number ⟨digit⟩ from the directory
- hello, what are my messages
- skip the next name
- help
- good-bye
- please send a car from the city
- dial ⟨number⟩
- delete my email tomorrow

- cancel
- read this text
- correct my balance
- call my daughter at eleven pm on ⟨day⟩
- erase all information
- no
- record extended phonebook
- get my office
- transfer all calls to home at twelve oclock
- use voice
- record urgent message
- yes
- find the operator
- call ⟨firstname⟩
- dial ⟨lastname⟩
- phone ⟨firstname⟩
- call ⟨number⟩
- phone ⟨number⟩
- Thank you for your participation. If you would like to receive a gift certificate for either McDonalds, TCBY yogurt, B Dalton Books, Blockbuster, or Baskin Robbins please leave your name, address, and selection. You may hang up when you are done. Thank you.

The data collection is just beginning. We expect this corpus will be available September 1994.

2.8. OPERA Corpus

CSLU is collaborating with the International Computer Science Institute (ICSI) at Berkeley to develop speech corpora for Open Performance Evaluation of Recognition Algorithms (OPERA). These corpora will be distributed with designated training and test sets to all researchers who wish to compare recognition performance on a common task. Performance evaluation and summary of results will also be provided.

The first OPERA corpus, now under development, consists of numbers taken from three of the corpora described earlier: the Spelled and Spoken Words Corpus, the Cellular Words, Numbers and Alphabet Corpus, and the English Census Corpus. We estimate the final corpus will consist of about 10,000 different numbers.

Thus far, we have created numbers files from utterances in the Spelled and Spoken Names Corpus in which the caller provided their street address and zipcode. Speech intervals containing numbers found in street addresses, street names (e.g., "fifth") and zip codes were located manually, and new files were created containing just the numbers. From approximately 1300 different speakers, 2167 files have been created. Each file has been transcribed at the non-time-aligned word level and at the time-aligned phonetic level.

3. AVAILABILITY

CSLU is dedicated to promoting progress in the field of computer speech recognition. To this end, corpora are made available at no charge to academic institutions. These data are available once they are completed. Portions of the Enhanced Multi-Language Corpus have been placed in the public domain.

For information on obtaining any of these corpora, the conventions document, or the speech tools, contact Mike Noel at noel@cse.ogi.edu.

4. ACKNOWLEDGMENTS

We are indebted to the organizations that helped fund the projects: U.S. Bureau of the Census, ONR, NSF, Linguistic Data Consortium, U S West, Digital Equipment Corporation, LINKON Corporation, and Apple Computer

Much of the corpus development would have been impossible without the dedicated efforts of the labeling and transcribing staff. Many thanks are due to Terri Durham, Vince Weatherhill, Amie Wilson, Victoria Noel, Alexandra Guerra, Troy Bailey, Johan Schalkwyk, and many others.

References

1. Terri Lander, S. T. Metzler, *The CSLU Labeling Guide*, CSLU, Oregon, February, 1994.
2. CSLU. *OGI speech tools user's manual*, Technical report, Center for Spoken Language Understanding, Oregon Graduate Institute, 1993.
3. R. A. Cole, K. Roginski, and M. Fanty, *A Telephone Speech Database of Spelled and Spoken Names*, Proceedings of the International Conference on Spoken Language Processing, Banff, Alberta, Canada, October 1992, pp 891–893.
4. Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, *The OGI multi-language telephone speech corpus*, Proceedings of the International Conference on Spoken Language Proceedings, Banff, Alberta, Canada, October, 1992, pp 895–898.
5. James L. Hieronymus, *Ascii phonetic symbols for the world's languages: Worldbet*, Journal of the International Phonetic Association, 1993.