PROJECTS

LVDID: Language Variety and Dialect Identification

LVDID Project and Goals

The Language Variety and Dialect Identification (LVDID) project supports language recognition and speaker recognition technologies by creating and sharing language resources, including data and specifications.

Within this project, LDC designed and executed several studies that created speech corpora for multiple languages and speech acts using an increasingly diverse arsenal of microphones and recording technologies.

Each LVDID study addressed a different set of language technology needs: multiple speakers, sessions, languages, situations and recording devices over time and across dialect regions.

Mixer 3

As systems improve in accuracy, LDC correspondingly increases the degree of difficulty associated with processing these data. Collecting various dialects of a given language as well as mutually unintelligible languages continues to provide challenging test conditions.

Mixer 3 built a corpus that not only addressed the traditional concerns of speaker, topic and handset variation, but also examined bilingualism in multiple languages paired with English, cross-channel speaker recognition and the effect of extended data on system performance.

In 2004, LDC collected speech from more than 3900 speakers in nineteen different languages: Bengali, four Chinese varieties, three English dialects, Farsi, Hindi, Italian, Japanese, Korean, Russian, Spanish, Tagalog, Thai, Urdu and Vietnamese. Participants completed 19,951 calls that were then used for speaker and language identification. Of this group, 1867 subjects (47.7%) completed fifteen additional phone calls.

Mixer 4

Mixer 4 collected multiple telephone calls, a subset of which were simultaneously recorded using a cross-channel system with eight microphones. The study focused on native speakers of American English and had a goal of 400 participants.

Each participant made ten ten-minute phone calls. Approximately half of the subjects also visited one of the collection sites, LDC or ICSI, where they placed two of the required ten calls that were recorded on the cross-channel platform. As a result, 25% of the participants were from the Philadelphia area, 25% were from near Berkeley and 50% were recruited from the remainder of the United States.

Mixer 5

In a departure from previous studies that focused on telephone conversations, Mixer 5 also collected cross-channel recordings of face-to-face interviews. Mixer 5 elicited speech within a variety of situations; 300 subjects each completed ten telephone calls and six personal interviews. The interviewer engaged subjects in conversation and guided them through a series of speech elicitation exercises including high and low vocal effort telephone calls, therefore providing a variety of speech styles from each participant.

The chart below indicates the minutes allocated to each type of situation per session.

Situation/Session number	1	2	3	4	5	6	Total
Repeating Questions	1	1	1	1	1	1	6
Warm-up	4						4
Family, Personal	5						5
Informal Conversation	20	9	14	9	9	10	71
Transcript Reading		20	15	10	15	10	70
Story Reading				5			5
Sentence Reading					5		5
Phrase/Word List Reading						5	5
Low Vocal Effort				5			5
High Vocal Effort						4	4
Total/Session	30	30	30	30	30	30	180

Linguistic Data Consortium 3600 Market St., Suite 810, University of Pennsylvania, Philadelphia, PA 19104-2653 USA
Telephone: +1.215.898.0464 • Fax: +1.215.573.2175 • ldc@ldc.upenn.edu • www.ldc.upenn.edu

Greybeard

LDC's first longitudinal speech study leveraged a rich network of past participants to create a unique corpus. Greybeard included speech samples taken over a seventeen-year span, from 1991-2008. Subjects of previous telephone collections, from Switchboard 1 through Mixer 3, were asked to complete twelve new ten-minute phone calls.

The following table summarizes the total number of phone calls made by Greybeard participants: those collected in Greybeard as well as in previous studies.

Project	Number of Calls	Year of Calls
Greybeard	1009	2008
Mixer 3	979	2004
Mixer 1 & 2	2358	2001-2003
Switchboard 2	362	1997
Switchboard 1	36	1990-1991

BNBS

The Broadcast Narrow Band Speech (BNBS) study was based on a resource not previously utilized in large-scale language recognition research and technology development. Audio clips were created by identifying narrow band speech in three different epochs of Voice of America (VOA) transmissions.

LDC identified, segmented and audited approximately 400 to 600 thirty-second segments for NIST's 2009 Language Recognition Evaluation in each of the following linguistic varieties: Amharic, Azerbaijani, Belorussian, Bosnian, Bulgarian, Cantonese, Creole (Haitian), Croatian, Dari, English (American and Indian), Farsi, French, Georgian, Hausa, Hindi, Japanese, Korean, Mandarin, Pashto, Portuguese, Romanian, Russian, Spanish, Swahili, Tibetan, Turkish, Ukrainian, Urdu, Uzbek and Vietnamese.

The addition of BNBS data has proven a cost-effective way to significantly expand the quantity of usable data available to language recognition system developers. Due to the success of this study, LDC is planning to expand its BNBS collection efforts.

MIXER 6

Similar in design to Mixer 5, Mixer 6 elicits speech in a variety of speaking styles, including interviews and telephone conversations. This study is currently underway and LDC's goal is to engage 600 participants in three onsite interviews and three to four onsite telephone calls that are simultaneously recorded over twelve to sixteen microphones.

Participants are also required to place calls to LDC's robot operator that pairs them and records their conversations. Mixer 6 focuses on native speakers of American English local to the Philadelphia area.

Mixer 6 is expected to conclude on January 1, 2010.

LVDID Collection History

The following table compares the features of LDC hosted telephone collection projects from Mixer 1 to Mixer 5 as well as Switchboard 1.

Features/Projects	SB1	M1	M2	M3	M4	M5
Core Calls (8+)	•	•		•	•	•
Variable Environments	•					
Unique Handsets (4+)	•	•	•	•	•	•
Extended Data (20+)		•	•	•	•	
Multilingual (4+)		•		•	•	
Cross Channel (4+)		•	•		•	
Transcript Reading (2+)		•				•
Interviews (6)						•

All LVDID data will be released to the research community after its use in various NIST technology evaluations. Learn more about NIST and its language evaluations at www.itl.nist.gov/iad/mig/

Please contact LDC to learn more about its collection capabilities.