LDC: RESEARCH

Fieldwork

Introduction

Data collection in the field continues to advance linguistic research. Researchers identify a language problem, have the inspiration to address it and then take the initiative to solicit interviews and interactions from subjects in order to accomplish their goal.

Data Collection Basics

Some data collections are conducted in locales close to the researcher, while others send researchers far afield. Though each setting has its own issues, all fieldwork requires advance planning, including the following:

- Developing a research plan
- Obtaining approvals from any bodies that regulate such research
- Securing adequate funding
- Choosing collection methods appropriate to the field conditions
- Selecting method(s) for data distribution

Some academic institutions and regulators may consider linguistic fieldwork to be human subjects research requiring certain protections. Those can include procedures for obtaining subjects' informed consent to the study, for documenting and protecting personally identifiable information about the subjects and for securing data in the field.

Many funding agencies require data to be published at the end of a collection. The planning phase is therefore not too soon for considering the ways and means of data distribution.

In the Field

Equipment. Microphones, recording devices and laptop computers are among the kinds of equipment used in fieldwork. Field geography and climate may be beyond the fieldworker's control, but equipment choices and placement can ameliorate conditions such as external noise, reflection and the distance between the subject, microphone and interviewer.

Some recent approaches in **fieldwork documenting endangered languages** incorporate simple technologies like handheld recorders and smartphones to allow large numbers of community members to capture speech for respeaking, transcription and translation. LDC supports two such studies in Papua New Guinea and Brazil funded by the National Science Foundation (BCS-0951651, IIS-0964556).

LDC also supports ethological research, including animal vocalizations, which shares methodological issues with linguistic fieldwork.

Collaborative fieldwork in Papua New Guinea
Courtesy: Steven Bird, LDC, University of Melbourne

Software. There are a variety of user-friendly, available software tools to aid fieldwork. These include:

- AGTK (Annotation Graph Toolkit) -- LDC's suite for transcription, observational coding and syntactic annotation http://agtk.sourceforge.net/
- PRAAT -- designed for phonetics research with capabilities for, e.g., speech analysis, labeling and segmentation and speech synthesis http://www.fon.hum.uva.nl/praat/
- The Linguist's Shoebox -- sorts, selects and displays text field data http://www-01.sil.org/computing/shoebox/
- The Penn Phonetics Lab Forced Aligner -- aligns transcribed speech to corresponding audio http://www.ling.upenn.edu/phonetics/p2fa/
- TranscriberAG -- transcription software http://transag.sourceforge.net/
- XTrans -- LDC's multilingual, multichannel transcription tool http://www.ldc.upenn.edu/tools/XTrans

In the Field, Continued

Data Management. Maintaining data integrity in the field requires a plan that incorporates procedures for each step of the collection, annotation and storage process. For instance, data may be collected to recorders or directly to central computers used by fieldworkers, uploaded to a database and backed up on mass storage devices. Subjects' personal identifying information may be collected in logbooks and then keyed into an encrypted spreadsheet on a laptop computer and backed up on a mass storage device. Otherwise it may be keyed in directly. Mass storage devices and logbooks are typically stored securely and under fieldworkers' control until they can be transferred to a secure network and secure storage location, respectively.

Recording Essentials at a Glance

Equipment should be adequate to support the immediate research goal but also eventual data sharing. Some things to consider when purchasing recording equipment:

- Sampling Rate - ≥16kHz is preferable
- Sample Size - ≥16 bits, if appropriate, for data source
- Compression - Lossless or none, some lossy compressions purport to be harmless but are rarely needed given storage costs
- Storage - needs are calculated as sampling rate * sample size/8 per second; 96,000 bps * 24/8 bit sample * 60 min/hr * 60 sec/min = ~1GB/hour
- Be familiar with technical software requirements for speech data

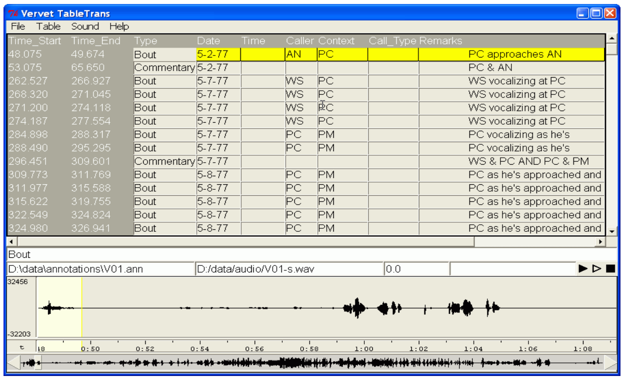
LDC Resources Based on Field Data

LDC's catalog contains several resources based on field collections that are useful for language research as well as human language technology applications.

USC-SFI MALACH Interviews and Transcripts English and Czech *LDC2012S05*, *LDC2014S04*: interviews and transcripts from Holocaust survivors and witnesses collected by the University of Southern California's Shoah Foundation Institute annotated by an international research team for speech recognition experiments.

Malto Speech and Transcripts *LDC2012S04*: Eight hours of speech and transcripts collected in India from native speakers of this minority language by Mosato Kobyashi (University of Tokyo) and Bablu Tirkey (Ranchi University).

Digital Archive of Southern Speech *LDC2012S03*: 370 hours of speech from the US Gulf states with associated metadata collected as part of the Linguistic Atlas Project (University of Georgia).



Time_Start	Time_End	Type	Date	Time	Caller	Context	Call_Type	Remarks
48.075	49.674	Bout	5-2-77		AN	PC		PC approaches AN
53.075	65.650	Commentary	5-2-77					PC & AN
262.527	266.927	Bout	5-7-77		WS	PC		WS vocalizing at PC
268.320	271.045	Bout	5-7-77		WS	PC		WS vocalizing at PC
271.200	274.118	Bout	5-7-77		WS	PC		WS vocalizing at PC
274.187	277.554	Bout	5-7-77		WS	PC		WS vocalizing at PC
284.898	288.317	Bout	5-7-77		PC	PM		PC vocalizing as he's
288.490	295.295	Bout	5-7-77		PC	PM		PC vocalizing as he's
296.451	309.601	Commentary	5-7-77					WS & PC AND PC & PM
309.773	311.769	Bout	5-8-77		PC	PM		PC as he's approached and
311.977	315.588	Bout	5-8-77		PC	PM		PC as he's approached and
315.622	319.755	Bout	5-8-77		PC	PM		PC as he's approached and
322.549	324.824	Bout	5-8-77		PC	PM		PC as he's approached and
324.980	326.941	Bout	5-8-77		PC	PM		PC as he's approached and

Bout

D:\data\annotations\V01.ann D:/data/audio/V01-s.wav 0.0

Focus on Ethology: Field Recordings of Vervet Monkey Calls LDC2004S12: 30 hours of digitized vervet monkey recordings annotated with AGTK TableTrans to show metadata about the situation surrounding a particular recording

Asian Elephant Vocalizations *LDC2010S05*: 57 hours of elephant recordings made by researcher Shermin de Silva (University of Pennsylvania) in Sri Lanka annotated with PRAAT to support research in acoustic features and repertoire diversity.

Grassfields Bantu Fieldwork: Dschang Tone Paradigms *LDC2003S02*, Dschang Lexicon *LDC2003L0*, Ngomba Tone Paradigms *LDC2001S16*: Tonological and phonetic transcription for portions of tone paradigms recorded by Steven Bird (LDC, University of Melbourne) in Cameroon in Bantu languages Dschang (Yémba) and Ngomba and a dictionary of sound files and related material collected from Dschang (Yémba) speakers.

Publishing Field Data

Making fieldwork data collections broadly available advances science by allowing for replication and benchmarking, comparison of results across studies and, over time, re-annotation and reuse for new purposes, among other things.

Data sets that are documented in standard formats and whose contents conform to any regulatory or proprietary constraints are candidates for distribution. LDC provides guidelines for corpus preparation and welcomes submissions of field data. For more information visit **https://www.ldc.upenn.edu/data-management/providing** or contact **ldc@ldc.upenn.edu**.