Digitizing the Audio Archive of Linguistic Fieldwork at the Berkeley Language Center

Mark Kaiser, Associate Director
with the help of
Marianne Garner, Librarian
Susan Stone, Programmer
Historical Background

- Fieldwork began in the 1950’s
- Collection at the UCB Language Lab dates from the 1960’s
- 1979-1980 UCB Language Lab received NEH funding to organize the archive
  - Archive mastered to reel-to-reel tape
  - Paper catalog created
Examples of the various media on which original materials are deposited (various reel-to-reel formats, magnetic wire, phonographs)
Storage of the original materials (left) and the mastered reel-to-reel tapes (right)
State of the Collection Late 1990’s

• Factors influencing decision to digitize
  – Exercising the masters
  – Tape and tape machine manufacturers were leaving the business
  – Limited access to information about the archive
  – Distribution was limited
  – Patrons’ needs
Technical Issues

• File format
  – Archive: 24 bit / 96 kHz, uncompressed .wav
  – Production: 16 bit / 44 kHz, uncmp. .wav
  – Web: Windows Media Audio (.wma)

• “Master” and backup
  – Hard drive array
  – Tape and DVD, 2 copies each
Legacy Permission Forms

- Collectors sign off on extent of access, contractual arrangement
- Three interest groups with the potential for disparate opinions
- Permission forms did not anticipate Internet
- Permission forms did not anticipate repackaging media
- In many cases original parties now deceased
Resegmentation

- Segmentation of the analog recordings was based on time constraints of media, done by audio engineers.
- Segmentation of the digital recordings is based on the content of the media, done by graduate students in linguistics.
Item-level Cataloging

• Creation of segment titles
  – Used collector’s data when available
  – Graduate students apply a limited uniform set of descriptors
  – Searchable in the online catalog
• Modification of content type terms
  – Controlled vocabulary, mappable to OLAC
• Lessons learned
Conforming to OLAC Standards

• OLAC as a moving target
• Methodology to conform
  – Database mapping tables where our vocabulary does not conform to OLAC’s
    • Personal roles, content type terms, language names
  – Java program generates an OLAC static repository, using SQL database tables from OLAC and a Sybase stored procedure in our database
Current Status

• Online catalog is up and running
  – Currently contains a mix of batch-generated data from the old catalog and newly digitized items together with the linked audio files
  – We are 6 months into the grant and on schedule to complete digitization and cataloging on time
  – We will register our collection with OLAC shortly
http://blc.berkeley.edu
http://blc.berkeley.edu
**Contents for Maidu (LA 30)**

<table>
<thead>
<tr>
<th>Segment</th>
<th>Title</th>
<th>Language/Comment/Note</th>
<th>Personal Names</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Miscellaneous words especially animals, numbers, and body parts includes simple phrases, e.g., &quot;I drink water.&quot;</td>
<td>Maidu, Northeast / linguistic data / Paynes Creek, CA. Formerly numbered 1.1 English glosses.</td>
<td>Bernier, Lissa T (linguistic consultant); Shapley, William L. (collector).</td>
</tr>
<tr>
<td></td>
<td>Words and phrases especially kinship terms.</td>
<td>Maidu, Northeast / linguistic data / Paynes Creek, CA. Formerly numbered 1.2 English glosses.</td>
<td>Bernier, Lissa T (linguistic consultant); Shapley, William L. (collector).</td>
</tr>
<tr>
<td></td>
<td>General talk includes some words and phrases pertinent to discussion.</td>
<td>Maidu, Northeast / Paynes Creek, CA. Formerly numbered 1.3. English glosses for some.</td>
<td>Bernier, Lissa T (linguistic consultant); Shapley, William L. (collector).</td>
</tr>
</tbody>
</table>

[http://blc.berkeley.edu](http://blc.berkeley.edu)
Contact Information

• Mark Kaiser, Associate Director, BLC
  – mkaiser@socrates.berkeley.edu

• Marianne Garner, Librarian
  – ll-lib@socrates.berkeley.edu

• Susan Stone, Programmer
  – sstone@socrates.berkeley.edu