Sharable Resources for Sociolinguistic Research

Christopher Cieri
Stephanie Strassell
William Labov
{ccieri@ldc.upenn.edu, strassel@ldc.upenn.edu, labov@central.cis.upenn.edu}
DASL: An investigation of best practices in the use of digital data & tools to support empirical linguistic inquiry and documentation. Funded by the National Science Foundation (BCS-998009, KDI, SBE) under the Talkbank project (www.talkbank.org) and by the Linguistic Data Consortium (www.ldc.upenn.edu).

1. Vision for empirical, quantitative research that is robust, collaborative, and accountable.

2. Review of previous DASL (www.ldc.upenn.edu/Projects/DASL) analyses using corpora; discussion of resources and tools developed.

3. **SLX Corpus of Classic Sociolinguistic Interviews** conducted by William Labov and his students available for distribution 1/2003.
Methodologies

1970

Interviews are recorded but not always transcribed; even then transcripts are often only partial.

Analytical tools are not integrated.

The presentation is an independent artifact.

After 30 years of technological advance, our use of data is largely unchanged; only the components differ.

2000
Audio File

Transcript

start stop uttID utterance speaker

Speaker Table

speaker sex age occupation education region

Search Terms
Exclusion List

Tokens are parts of an interview that match search criteria.

Token Table

uttID tokenId token Dep. Variable Ind. Variables...

Lexicon

lexical item pronunciation word class

Electronic lexicons simplify many kinds of coding tasks.

Papers published on CD/Internet refer back to coded tokens.
Corpora

- Originally created for linguistic technology development
- Selected for range of styles, availability of time-aligned transcripts
- Contain basic speaker demographics

<table>
<thead>
<tr>
<th>Corpus</th>
<th>ISBN</th>
<th>Minutes</th>
<th>Type of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>1-58563-019-5</td>
<td>6300</td>
<td>Phonetically Rich Sentences</td>
</tr>
<tr>
<td>Switchboard-1</td>
<td>1-58563-121-3</td>
<td>12000</td>
<td>Short Conversations with Constrained Topics among Strangers</td>
</tr>
</tbody>
</table>

- t/d deletion case study
  - Well-documented and well understood, stable indicator
  - Are corpus data results comparable to traditional studies?
  - Linguistic and social factors
    - morphological, preceding & following phonological environments, stress, cluster complexity
    - age, gender, education, region, race
  - Results are substantially similar to previous t/d studies
    - See Strassel - NWAV2001
**Tools**

- **Concordance** identifies tokens of interest through regular expression query
- **Filters** remove additional non-tokens
- **Tag set** specifies factors to code
- **Web browser** displays annotation file
  - Listen to audio
  - Code tokens quickly
  - View demographic information
- **Save results and output to text file for further analysis**

**TIMIT Corpus**

- 55,000 words concordance 3154 words filters 2059 words annotate 1578 t/d tokens

**Switchboard Corpus**

- 3,217,800 words concordance 100,048 words filters 45,164 words annotate 26,733 t/d tokens
**DASLTrans** - transcription and coding tool

- **Audio**: handles arbitrary length audio files
- **Text**: AG compliant XML
- **Tag set**: user defined
- **Integration**
  - Listen to audio
  - Segment easily
  - Transcribe directly maintaining time-alignment
  - Code using same tool
  - Output results in table format for further analysis
- **Free**
- **Extensible** via distributed source code
SLX Corpus

- **SELECTION:** Sessions selected by William Labov where observation effects are minimized, style more closely approximates vernacular and sound quality is high. Interviews conducted between 1963 and 1973 primarily by Labov.

- Sessions digitized from open reel tapes onto DAT/disk at 16bit, 44KHz sampling. Monaural signal passed through 2 channels at levels differing by 20% to capture best digital copy in single pass. Technician monitored recording, adjusted for sustained changes in speech levels. Digital files show no significant clipping in the digital domain.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Age</th>
<th>Speech Community</th>
<th>Occupation</th>
<th>Tapes</th>
<th>Others</th>
<th>Minutes</th>
<th>Words</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adolphus H.</td>
<td>82</td>
<td>Hillsboro, NC</td>
<td>Farmer</td>
<td>2</td>
<td>3</td>
<td>85</td>
<td>9660</td>
<td>1494</td>
</tr>
<tr>
<td>Bobbie A.</td>
<td>32</td>
<td>Ayr, Scotland</td>
<td>Saw Doctor</td>
<td>1</td>
<td>1</td>
<td>44</td>
<td>8990</td>
<td>1769</td>
</tr>
<tr>
<td>Henry G.</td>
<td>61</td>
<td>E.Atlanta, GA</td>
<td>Railroad Mechanic</td>
<td>3</td>
<td>5</td>
<td>112</td>
<td>20012</td>
<td>2372</td>
</tr>
<tr>
<td>Jerry T.</td>
<td>20</td>
<td>Leakey, TX</td>
<td>Gas Attendent</td>
<td>2</td>
<td>1</td>
<td>66</td>
<td>11264</td>
<td>1700</td>
</tr>
<tr>
<td>Joe D.</td>
<td>21</td>
<td>Liverpool, ENG</td>
<td>Docker</td>
<td>2</td>
<td>0</td>
<td>100</td>
<td>19798</td>
<td>2515</td>
</tr>
<tr>
<td>Eddie M.</td>
<td>20</td>
<td>Liverpool, ENG</td>
<td>Docker</td>
<td>2</td>
<td>0</td>
<td>100</td>
<td>19798</td>
<td>2515</td>
</tr>
<tr>
<td>Kathy D.</td>
<td>15</td>
<td>Rochester, NY</td>
<td>Student</td>
<td>2</td>
<td>2</td>
<td>64</td>
<td>29001</td>
<td>1938</td>
</tr>
<tr>
<td>Louise A.</td>
<td>53</td>
<td>Knoxville, TN</td>
<td>Mother/Domestic</td>
<td>3</td>
<td>0</td>
<td>76</td>
<td>11348</td>
<td>1521</td>
</tr>
<tr>
<td>Rose B.</td>
<td>36</td>
<td>New York, NY (LES)</td>
<td>Seamstress</td>
<td>3</td>
<td>3</td>
<td>60</td>
<td>12184</td>
<td>1938</td>
</tr>
</tbody>
</table>
**Segmentation**

- One audio file per speaker
- Distinguish target speaker from other speakers, silence, non-speaker noise

<table>
<thead>
<tr>
<th>First pass segmentation</th>
<th>Second pass segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>◆ ID basic utterance boundaries</td>
<td>◆ Finer-grained boundaries</td>
</tr>
<tr>
<td>◆ Play audio</td>
<td>◆ Additional breakpoints at</td>
</tr>
<tr>
<td>◆ Hit &lt;enter&gt; to create boundary</td>
<td>- Sentence/phrase boundaries</td>
</tr>
<tr>
<td>◆ Close to 1X real time</td>
<td>- Noticeable pauses (&gt;500ms)</td>
</tr>
<tr>
<td></td>
<td>- Breath groups</td>
</tr>
</tbody>
</table>
Transcription

First pass segmentation
- Verbatim transcript
- No “correction” of speakers' grammar, pronunciation
- Standard orthography, punctuation
- Special conventions for
  - unintelligible speech
  - non-standard lexical items
  - speaker restarts

Second pass segmentation
- Verify existing transcript
- Revisit ((unintelligible)) sections

Third pass segmentation - specialized
- Dialect-specific review

Is that ((Hugh Potty))? → Is that how you put it?
She done her lovely! → She done a wobbler!
Bloody (( )) uh. → Bloody nutters, youse are.
All ((amber)) heads. → All them birds.
What Variables Appear in the Corpus?

First pass – variable survey to encourage future research
- Examine general and dialect-specific variables
- Determine presence/absence of each variable for all speakers

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Categories</th>
<th>Subcategory Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonological, Phonetic, Prosodical: 90 variables</td>
<td>Consonants</td>
<td>(DH) - voiced interdental fricative</td>
</tr>
<tr>
<td></td>
<td>Front Vowels</td>
<td>(ae-NAS) - tensing of short-a before nasals</td>
</tr>
<tr>
<td></td>
<td>Back Vowels</td>
<td>(ahr) - realization of /ahr/ sequence</td>
</tr>
<tr>
<td></td>
<td>General Vowels</td>
<td>(SCHWA) - realization of schwa</td>
</tr>
<tr>
<td></td>
<td>Diphthongs</td>
<td>(aw) - realization of /aw/</td>
</tr>
<tr>
<td></td>
<td>Prosody</td>
<td>(RISE) - rising final intonation</td>
</tr>
<tr>
<td>Grammatical: 60 variables</td>
<td>Prepositions</td>
<td>(PREP-DEL) - preposition deletion</td>
</tr>
<tr>
<td></td>
<td>Adjectives</td>
<td>(ADJ-WO) - non-standard ADJ word order</td>
</tr>
<tr>
<td></td>
<td>Determiners</td>
<td>(DET-DEL) - determiner deletion</td>
</tr>
<tr>
<td></td>
<td>Negation</td>
<td>(NEG-AINT) - use of ain't in neg. constructions</td>
</tr>
<tr>
<td></td>
<td>Word Order</td>
<td>(WO-LEFTDIS) - left dislocation of initial NP</td>
</tr>
<tr>
<td></td>
<td>Pronouns</td>
<td>(POS-LEV) - leveling of possessives to mine paradigm</td>
</tr>
<tr>
<td></td>
<td>Verbs</td>
<td>(COP-DEL) - copula deletion</td>
</tr>
<tr>
<td></td>
<td>Quantifiers</td>
<td>(Q-BUT) - but as quantifier</td>
</tr>
<tr>
<td></td>
<td>Agreement</td>
<td>(PLURAL) - singular ending on plural noun</td>
</tr>
</tbody>
</table>

Variable profile, examples of each variable for each speaker
Legal, Ethical Issues

- Collection now requires informed consent of subjects.
- Shared data must protect subjects’ anonymity.
- Distribution requires permission from copyright holder.
- SLX Corpus of Classic Sociolinguistic Interviews may be used only for linguistic education, research and technology development.
- Researcher is responsible to make best effort to ensure that uses of shared data respect the dignity of subjects.
- Need community specific code of ethics for shared data.
Publication Standards

- Development, productions methods fully documented
- Complete audio available in standard format (AIFF, RIFF, SPH) uncompressed or with lossless compression.
- Transcripts in XML or other standard, non-proprietary platform-independent and application-independent format.
- Consistent naming conventions for audio, transcriptions and any annotations.
- All data formats specified and confirmed.
- Inter-annotator agreement measured and published.
- Coding practice fully documented; results shared.
◆ Shared data resources and tools
  ◆ enables comparison of results across studies and over time
  ◆ serves as stable benchmark for competing theories
  ◆ allows re-annotation and reuse of existing data
  ◆ supports measurement of inter-annotator consistency
  ◆ reduces impediments facing new researchers
  ◆ allows established scholars to tackle broader issues

◆ SLX Corpus of Classic Sociolinguistic Interviews
  ◆ classic interviews, cited in literature
  ◆ demonstrate best practice in conducting sociolinguistic interviews
  ◆ represent variety of speech communities
  ◆ effect of observation minimized
  ◆ sound quality high
  ◆ demonstrate best practice in digitization, segmentation, transcription
  ◆ teaching tool for sociolinguists
  ◆ stable benchmark for training/comparing transcription and coding
  ◆ an example of a multi-community sociolinguistic corpus