LDC Data Clinic

Christopher Cieri, Denise DiPersio, James Fiumara

NWAV 48, 2019, Eugene, OR
Goals

- Discuss data issues arising from our interaction with this community; intended for students, early careers, people newly facing corpus issues

- Our Credentials
  - LDC = first and largest (also best) data center devoted to language
  - 820 corpora in 90+ languages of which
  - > 187,000 copies distributed to more than 6400 Organizations in 100 Countries
  - Used in > 10,000 Papers
  - Supporting ~10 common task evaluations annually – a model for reuse, replicability
  - Chris
    - PhD in Lx from Penn under Labov, Liberman; LDC Exec. Dir. since 1998, Adj. Assoc, Prof. Lx, 15 additional years exp in LRs, IT
  - Denise
    - JD, 14 Years managing LDC External Relations, (IPR, IRB, distribution, curation)
  - James
    - PhD in English, MA in Lx, 1st Fellow at Penn’s Price Lab for DH Fellows, NIEUW project manager
Models of Data Intensive Research

Plan Research → Write Proposal → Collect Data → Code Data → Analyze Data → Write

Plan Research → Write Proposal → Collect Data → Code Data → Analyze Data → Share
Why Share Data? Why Use Shared Data?

- Funders require it
  - U.S. National Science Foundation, Canada Social Sciences and Humanities Research Council
  - White House Office of Science and Technology Policy (OSTP)
- Journals and Conferences beginning to encourage/require it
- Promotes Good Science
  - replication
  - comparison
  - study of change over time
- Increases Collaboration
- Permits Specialization
- Lower Barriers to Entry
- The research community desires it:
  - The previous, dominant model of considering sociolinguistic data as too valuable to ‘part with’ or to share appears to be giving way to a model where sociolinguistic data is considered to be too valuable not to share. (Kendall, Van Herk 2011: 3)
- Sharpens your own research practices; imposes discipline
- A Good Thing
5 Conditions of Shared Data

- **Discoverable**
  - user can find, evaluate relevance (absent personal relationship with owner)

- **Accessible**
  - open terms, procedures stated clearly and in advance for direct access via persistent URL

- **Interpretable**
  - independently understandable to target community, w/o special resources (e.g. consultation)

- **Portable**
  - interoperate in user’s working environment (hardware, software, community, practices) using open, transparent, widely supported formats

- **Preserved**
  - faithful copy of original (meta-)data, constantly verified, persists, protected against contingencies (e.g. reliable backup, multiple sites, migration to new media), fixes as patches
Hello World!
Does collection adequately represent research purpose?

Underlying assumptions about collection
- control over self
- control over speaker, Observer’s Paradox, over-sampling
- control over situation
  - room
  - noise

Recording Parameters
- sampling rate, sample size, compression, format

Microphone / Recorder
- usability & compatibility (especially connectors and power supply)
- operating parameters such as placement distance & direction relative to type
- make, model, pick-up pattern

Comparability
In addition to shareability, methodology impacts:
- efficiency & scale thus results
- accuracy & consistency
- balance, completeness & dynamics
- (your) reuse for another purpose & replicability

text, (e.g. transcript, generally human) required for efficient methods

Decision Points & Comparability:
- nature of dependent variable
- modeled: continuous/discrete/categorical, separate/part of general process
- independent variables, or factor groups, considered
- values (or variants or factors), assigned to each of factor group
- sampling: talkers, situations, tokens considered/excluded, features by which
- elicitation method
Coding Differences & Comparability

- **t/d following environment**

<table>
<thead>
<tr>
<th>C</th>
<th>C</th>
<th>~V</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>~C</td>
<td>V</td>
</tr>
</tbody>
</table>

- **t/d preceding environment**

<table>
<thead>
<tr>
<th>pb td kg ?</th>
<th>stop</th>
<th>stop</th>
<th>stop</th>
<th>stop</th>
<th>stop</th>
<th>stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>fv θδ h</td>
<td>fricative</td>
<td>fricative</td>
<td>fricative</td>
<td>fricative</td>
<td>stop</td>
<td>spirant</td>
</tr>
<tr>
<td>sz ñ3</td>
<td>sibilant</td>
<td>sibilant</td>
<td>sibilant</td>
<td>sibilant</td>
<td>sibilant</td>
<td>sibilant</td>
</tr>
<tr>
<td>m(n)η</td>
<td>nasal</td>
<td>nasal</td>
<td>nasal</td>
<td>nasal</td>
<td>nasal</td>
<td>nasal</td>
</tr>
<tr>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>l</td>
<td>L</td>
<td>l</td>
</tr>
</tbody>
</table>

black = not reported  same shading within column = not reported as statistically significant.
Coding Metadata

- **Demographic**
  - association with multiple groups
  - immigrant communities
  - disenfranchised groups

- **Situation**

- **Attitude**

- **Nature of elicitation matters**
  - mined from discussion: in what context
  - elicited by written form or by live questions
    - form of the question, number of categories and which
    - who asked
  - any merging of categories (e.g. due to imbalance in sample)
  - change over time

- **When is a simplifying assumption a distorting practice?**
Creation of LRs limited as it employs finite resources (e.g., grant funds) for nearly infinite problem

Proposal: use novel incentives to harness the renewable resources of the human drives to learn, compete and make meaningful contributions.

Novel Incentives: information, entertainment, self-expression, developing skills, socializing, competition, status, recognition, contributing to greater cause or good

Successful examples:
- LibriVox: public domain audiobooks, >9,400 volunteers, >13,000 books, 39 languages
- Zooniverse: Citizen Science portal founded 2009, over 200 total projects (105 active), >1,900,000 contributors
- Game with a Purpose (GWAP): ESP Game, Google Image Labeler, Great Language Game
NSF CISE Research Infrastructure (CRI) grants
- Planning Grant: 2016 - 2017
- NEW Grant: July 2017 – June 2020

Develop and build infrastructure:
- Software toolkit to build Ix activities and projects
- Web portals for hosting projects, crowdsourcing, community building

Build upon LDC’s WebAnn (web annotation) tools
- Simplify tools to allow researchers to build Ix annotation activities online
- Simplify annotation task design & interface for non-experts
- Create multiple portals geared towards different incentives / workforces
  - gamers, citizen scientists, language teachers/students
Multiple web portals

Cognitive Health

Games

Citizen Linguists

Language Pros / Students

Transcribing Labov Archive

Web portal for Students, teachers
lingboingo.org
Italian: Dialects, Regional and Standard
Classifying Italian Speech

Listen to the audio clip and classify the speech as Standard or Regional Italian or Dialect using the buttons below.
LingoBoingo (games portal)
- deployed
- link farm to partners’ games
- Name That Language

LanguageARC (citizen science portal)
- Deployed soft release
- Official release within ~1 month

LanguagePro (teachers, students portal)
- development in progress
- transcription tool prototyped
- otherwise used as staging ground for task types
- CogHealth, developed with non-NSF funds
Navigation Tips

- Human Subjects Collection
- Privacy/Ethics
- Intellectual property
- Finding an archive
- Choosing a license
- Distribution plan
- The Data Management Plan
Human Subjects Collection 1/2

- Triggered by obtaining data from humans by intervention, interaction
  - Requires a protocol approved by an Institutional Review Board
  - Procedures vary by university – check department, research resources

- Based on the Common Rule
  - Respect for persons, beneficence, justice
  - Focus on vulnerable populations, informed consent, confidentiality

- The Linguist Fieldwork Problem: IRBs as research obstacles
  - Survey suggested otherwise (Bowern 2010)
  - Clinical orientation of some IRBs; aversions to data sharing

- Recent changes to the Common Rule
  - Some studies are exempt from human subjects research requirements, although consent should be obtained to confirm data use & sharing
    - “Benign behavioral interventions”
    - Oral histories exempt, but not ethnographic studies
    - Tribal rights – additional review, approval under tribal laws
Most language-related studies are considered minimal risk and subject to expedited IRB review

IRB submissions should cover:
- Study description: collection method (field, crowd, campus)
- How subjects’ informed consent will be obtained
  - Disclosure about use in research and how data will be shared
- How privacy will be protected (e.g., anonymization) if this is a goal
- How data will be stored, secured, maintained and shared

Consent methods
- Written consent, recorded verbal consent, click-through consent on web-based registration system, consent through action
- Parent/guardian consent, minor assent

US open data initiatives promote data sharing: impetus for Common Rule changes
Privacy/Ethics

- No single US data protection/privacy law (as opposed to GDPR)

- Sector-specific solutions: credit reporting, health care (HIPAA)

- Personal identifying information

- Anonymization is a good solution for language-related studies
  - Separate personal identifying information from research data; retain for contact, compensation
  - But threat of re-identification within a data set should be assessed
    - A few data points are enough to re-identify (the dark side of algorithms)
    - Potential retribution in minority language communities

- Some communities/participants want credit for their participation
Who owns a language?
- Language communities may claim ownership rights
- May include non-language community data

Copyright rights in found data collections: web-based text, speech, social media
- Favorable license (creative commons, etc.)
- Explicit permission from data provider
- Fair use (apply with caution!)
- Website terms of use can derail a collection
  - Browser-wrap terms of use
  - Regulate how material can be used, shared, modified; third party data problem

Seek guidance from university resources
Finding an Archive

- NSF programs may not recommend a repository
  - DEL program requires DSA archive
- A good choice: archives that follow accepted standards, best practices for digital resources (LDC example)
  - Data quality: independent quality checks for deposited data sets
  - Stability: evidence that the repository is established and likely to remain in existence
  - Discoverability: the ease of the finding the resource
  - Standards for metadata, formats, tools, documentation: enhancing usability
- Costs
  - Funding model based on actual costs
  - Flexibility: sponsor-funded, incremental cost sharing
- FAIR: findable, accessible, interoperable, reproducible
- LDC is a CoreTrustSeal repository – WDS/DSA
Choosing a License

◆ Confusing license nomenclature
  ● “Freely available”
  ● Open source – may not mean cost-free
    ■ Restrictions on derivatives, sharing
    ■ Viral terms
  ● “Research only”
    ■ Consider how data is used to train language technology systems

◆ LDC license model
  ● Language-related research, education and technology development
    ■ Includes commercial development by for-profit Consortium members
  ● Licensees cannot redistribute resources outside their organization
  ● Options: research-only, referrals to data owners for commercial licensing
  ● Special licenses to support shared tasks
LDC Distribution Plan Options

- Archive as host
  - Temporary v. permanent deposit
  - Arrangements regarding curation services: storage, back-up, security
    - Associated costs
  - Hosting for distribution: no distribution services; provided to users “as is”

- Archive as host and distributor
  - Moratorium on distribution pending permanent deposit
  - Distribution-related services: quality control, licensing, customer care
  - Conditions for access: license, fee, other (e.g., organization type)
  - Distribution method
    - Web download, media
    - Archive servers, third party hosted service (e.g., cloud)
    - Associated costs
The Data Management Plan

- The foregoing flows into the DMP
- NSF DMP guidelines for SBE proposals
  - Describe the data expected to be generated
  - Identify a hosting archive
  - Explain data access and sharing
  - Identify other information maintained and shared, such as metadata
  - Pertinent intellectual property rights
  - Ethical and privacy issues
  - Data, tools and documentation formats
  - Plans for archiving and preservation
- Check institutional DMP resources
  - Often maintained by the library
  - Decision trees, tools
- LDC DMP services: DMP submissions form; data sheet
LDC DMP Data Form

Please fill out the following form as completely as possible. If more than one choice applies, list other choices in the "Other/Note" box. Translational texts are marked with an asterisk (*).

**Contact Details**

Name:

Email:

**Corpus Details**

Title of corpus (proposed):

Authors (comma separated):

Languages (comma separated):

Data Type (select all that apply):
- Text
- Audio
- Video
- Lesson
- Basel

Objective(s), please specify:

Estimated Data Size (uncompressed):

**Data Sources**

Describe the nature of the source data and the methods of collection, e.g., broadcast, conversation, news, documentary, pre-recorded, spontaneous, speech, telephone, telephone, ear, field/audio recordings, interviews, e.g., telephone, etc.

Description:

Provide a short (1-2 paragraphs) of the corpus. Included here — use list the "data sources" text — should be descriptions of any presentation or usage to which the data, pertinent applications for the corpus and the intended audience. There may be some overlap between this text and the preceding text. You may use these fields in your title section in order to provide the broadest possible description of the corpus contents.

Ownership:

Do you own all of the data in this corpus?

- Yes
- No

If "No," or "Not sure," please further explain your answer:

**Distribution**

Do you wish LDC to distribute the data in addition to archiving?

- Yes
- No
- Not sure

If "Yes" or "Not sure," what, if any, external constraints on the distribution of this corpus exist? (e.g., release date, price, licensing):
Open Clinic Session

- Share your plans, questions, experiences!
- Additional drop-in session today, 2:00-4:00 pm, Oak Room

- Thank you!