Data types across linguistic subfields

Kavon Hooshiar
Andrea L. Berez-Kroeker
University of Hawai‘i at Manoa
Our homework assignment

To find out what counts as “data”

In different linguistic subfields

For the purpose of making linguistic research reproducible.

In other words

What is the range of materials for which we need to standardize
   how to store,
   how to share,
   how to cite
in a reproducible linguistics.
Our starting point
Himmelmann (2012)

Examines **data levels** in language documentation

Using an analogy to philology. Raw > Primary > Secondary

<table>
<thead>
<tr>
<th>Data level</th>
<th>Examples from Philology</th>
<th>Example from LangDoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data <strong>Particular, unique</strong></td>
<td>Inscription, original manuscript</td>
<td>Audio/video recordings, (Non-recorded) observations</td>
</tr>
<tr>
<td>Primary data <strong>Particular, non-unique</strong></td>
<td>Critical edition</td>
<td>Transcript with translations, Field notes</td>
</tr>
<tr>
<td>Secondary/Structural data <strong>General, replicable</strong></td>
<td>Statements about language history (e.g. Old High German ø &gt; Middle High German ö)</td>
<td>Glossing, descriptive statement, dictionary entry, entry in a typological DB, implicational universal...</td>
</tr>
</tbody>
</table>
Our starting point

Hypothesis based on Himmelmann

Are data levels the key to cross-subfield reproducibility?

Can we identify data levels for all subfields?
Can we then postulate:
“Across the board, linguists should archive their raw/primary/secondary data.”

(No, it’s not that simple)
What we did

We kept our research close to home: Interviewed UHM linguistics faculty:

Victoria Anderson - Laboratory Phonetics
Robert Blust - Historical Linguistics
Kamil Deen - Language Acquisition
Katie Drager - Sociolinguistics/Sociophonetics
Patricia Donegan - Phonology
William O’Grady - Syntax & Acquisition
Yuko Otsuka - Formal Syntax
Amy Schafer - Psycholinguistics
What we did

We asked them

- What is your subfield’s prototypical research? (Questions, Methods, Data)
- What is raw data to you? What is primary data to you?
- What level of detail of data do you wish was recoverable?
- What are practices for data sharing/reproducibility in your subfield?

Freeform interviews lasting 30-60 minutes
Summary findings

Most subfields would need a mixture of data levels and types
  Eg., natural language data + experimental measurements + scripts
Need to work out hierarchical relationships between data types

Most subfields don’t have a system in place
  but agree that having one would be beneficial to all
  (less clear with sociolinguistics and syntax)

Our solution should be robust and broadly inclusive
  Having a flexible process in place will make ideology easier
Psycholinguistics: Eye tracking

Position of eyes (x- and y- axes) over time

Converted to attending to regions of interest
  Non-trivial calculation
  Script-based calculations

Analyzed with statistics
  Also script-based

Data not published, researcher obligated to maintain it
Could be asked to share with other researches
Would be useful to check statistical methods

Necessary data include: linguistic stimuli, visual stimuli
Sociolinguistics: Quantitative variation

Typical research question: How is the realization of a specific linguistic phenomenon affected by other linguistic and social variables?

Necessary data include:
Natural language data + transcripts + speaker meta data (c.f. documentation)
   Plus coding for individual tokens

Processing of raw data (natural language) can include:
   Coding (non-trivial and hard to automate)
   Acoustic analysis (both scripting and manual measurements)
   Auditory judgements
Sociolinguistics: Quantitative Variation

Data shared “in house”

Scripts might be viewed by some as a type of value added intellectual property

Example of online database: SOLIS

Anonymity is especially relevant for sociolinguists
  Larger amounts of publicly available data make it easier to ID individuals
Acquisition: Picture tasks

Often combines natural language and experimental data
Production vs perception tasks

Common example: Picture tasks
Production: describe this scene / describe the specific relationship between these objects
Perception: which picture in this group corresponds to the linguistic stimulus?

Necessary data for perception example: visual stimuli, linguistic stimuli, responses

Consultants by definition include children

Resources exist for natural language data (CHILDES), but not experimental
Acquisition: Picture tasks

Often combines natural language and experimental data
Production vs perception tasks

Common example: Picture tasks
Production: describe this scene /
    describe the specific relationship between these objects
Perception: which picture in this group corresponds to the linguistic stimulus?

Necessary data for perception example: visual stimuli, linguistic stimuli, responses

Consultants by definition include children

Resources exist for natural language data (CHILDES), but not experimental
Acquisition: Picture tasks

Often combines natural language and experimental data
Production vs perception tasks

Common example: Picture tasks
Production: describe this scene / describe the specific relationship between these objects
Perception: which picture in this group corresponds to the linguistic stimulus?

Necessary data for perception example: visual stimuli, linguistic stimuli, responses

Consultants by definition include children

Resources exist for natural language data (CHILDES), but not experimental
Conclusion: Data bundles?

Data levels is too simplistic a concept to really achieve reproducibility.
   Data are multiple and varied.
   A data bundle seems to make sense.

All the necessary raw and/or primary data
   plus everything used in the derivation.

The bundle would (could?) be
   A digital object with a DOI
   Containing recordings, images, scripts, measurements, descriptions, etc.
   With a usage guide
   And standardized metadata
   ...lodged in a digital repository.
Conclusion: Data bundles?

Our task here could be
   To provide guidelines/instructions for creating bundles
      Flexible for all subfields
   To provide advice for finding a repository
   To develop standardized metadata for linguistics research bundles
   To develop citation formats for the entire bundle
      And for subparts of the bundle
         And for subparts of subparts of the bundle…?