Welcome

1st Workshop on Data Citation and Attribution in Linguistics

Introductions & Acknowledgements

Andrea Berez-Kroeker, University of Hawai‘i at Mānoa,
Kaipuleohone Language Archive

Gary Holton, Alaska Native Language Archive & University of Hawai‘i at Mānoa

Peter Pulsifer, National Snow & Ice Data Center

Susan Smythe Kung, Archive of the Indigenous Languages of Latin American,
University of Texas at Austin

Special thanks to NSF for supporting this project, SMA-1447886.

Part of a larger cross-cutting NSF initiative: *Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution*

More Introductions

Our graduate student assistants:

Meagan Dailey (U Hawai‘i)

Ryan Henke (U Hawai‘i)

Jaime Perez Gonzalez (U Texas Austin)

Nick Williams (U Colorado Boulder)

Goals of the project

Develop and promote standards for data citation & attribution for linguistic data sets.
Shift the field toward a more scientific, data-driven model which results in reproducible research.
Hold three workshops and one panel presentation:

- Workshop 1: Boulder, Colorado, 18-20 September 2015
- Workshop 2: Austin, Texas, 8-10 April 2016
- Panel Presentation & Workshop 3 (1-day): LSA Annual Meeting, Austin, TX, January 2017

Output at the end of the project:

- Submit a proposal for a Resolution on citation and attribution to the LSA.
- Write a position paper on standards for citation and attribution in linguistics

On reproducibility in science and linguistics

- In science, claims must be falsifiable, verifiable, and *reproducible*.
- Reproducibility is similar to replicability
  - Replicability: recreation of controlled study > new data > confirm (or disconfirm) conclusions
    - Belly button microbe study (Hulcr et al. 2012)
  - Reproducibility: recreation of study not controllable > reuse of another's data > confirm (or disconfirm) conclusions
    - Chimpanzee tool-use study (Tomasello & Call 2011)
    - Linguistics, e.g., Choice of inflected form in spontaneous conversation

On reproducibility in science and linguistics

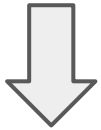
- Valued by the *Reproducible Research* movement:
  - “The product of academic research is the paper *and the full data* so that claims can be reproduced.” http://biostatistics.oxfordjournals.org/content/10/3/405.full
- Grew out of computer science
  - provide software *and* the underlying code
- Linguistic science values reproducibility too.
- Open Science Project:

On reproducibility in science and linguistics

"If a scientist makes a claim that a skeptic can only reproduce by spending three decades writing and debugging a complex computer program that exactly replicates the workings of a commercial code, the original claim is really only reproducible in principle."

Dan Gezelter. 2009. http://www.openscience.org/blog/?p=312.

On reproducibility in science and linguistics



"If a linguist makes a claim that a skeptic can only reproduce by spending three decades working in the same language community in the same sociolinguistic and fieldwork conditions, the original claim is really only reproducible in principle."

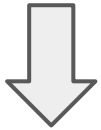
Dan Gezelter. 2009. http://www.openscience.org/blog/?p=312.

On reproducibility in science and linguistics

"Our view is that it is not healthy for scientific papers to be supported by computations that cannot be reproduced except by a few employees at a commercial software developer. [...] It may be research, and it may be important, but unless enough details of the experimental methodology are made available so that it can be subjected to true reproducibility by skeptics, it isn't Science."

Dan Gezelter. 2009. http://www.openscience.org/blog/?p=312.

On reproducibility in science and linguistics



"Our view is that it is not healthy for linguistics papers to be supported by examples that cannot be reproduced except by doing one's own fieldwork. [...] It may be research, and it may be important, but unless enough details of the utterances in context are made available so that it can be subjected to true reproducibility by skeptics, it isn't Science."

Dan Gezelter. 2009. http://www.openscience.org/blog/?p=312.

On reproducibility in science and linguistics

- The debate over reproducibility has been prominent in Language Documentation:
  - “[Language] documentation will ensure that the collection and presentation of primary data receive the theoretical and practical attention they deserve” (Himmelmann 1998:164)
  - “[...] it is our professional responsibility to provide the data on which our claims are based. [...] It enhances the scientific basis of the linguists’ work.” (Thieberger 2009:365-6)
  - “Establishing open archives for primary data is in the interest of making analyses accountable.” (Himmelmann 2006:6)

On reproducibility in science and linguistics

Potentially relevant to all of linguistics:

Data can be made available,

And data preparers can receive attribution for it,

And data can be cited,

And skeptics can confirm claims,

And linguistics becomes a more data-driven science.

But what do we mean by *citation*, *attribution* and *data*?

Citation and Attribution

Citation refers to the practice of identifying the source of linguistic data

Can be more or less granular but crucially needs to identify data within a larger context.

Attribution refers to the practice of giving people credit for collecting (and providing access to?) data

requires developing protocols for assessing and evaluating research outputs what should people get credit for?

Types of data

- Raw data
- Primary data
- Secondary data
- Derivative products
- Database -- means a lot of different things (but it is NOT a corpus)
- Corpora
- Experimental data (reaction time, eye-tracking, etc.)
- Phonetic data
- etc.

Overview of Workshop 1

Format: Combination of presentations and working group discussion & presentations.

Friday: Presentations on current state of the art in data citation and attribution and citations practices in linguistic journals and subfields.

Saturday: Mini-presentations on evaluation, data packaging, archives, the digital humanities, and communication, plus working group presentations (morning and afternoon).

Sunday: Recap, make work plans for Workshop 2

A few logistics

Restrooms are directly below us (and a small one is down the hall).

Bagged lunches will be brought in here today and tomorrow; feel free to take them outside.

Dinner tonight is at Aji Restaurant--it's about a 1.2 mile walk from here.

Saturday lunch there will be a meeting for DELAMAN members.

Saturday dinner is on your own (see list in your folder).

Introductions