Writing data citation guidelines for linguistics: Lessons learned

Helene N. Andreassen, UiT The Arctic University of Norway
Andrea Berez-Kroeker, University of Hawai'i at Manoa
Lauren B. Collister, University of Pittsburgh
Philipp Conzett, UiT The Arctic University of Norway
Christopher Cox, Carleton University
Koenraad De Smedt, University of Bergen & CLARIN
Lauren Gawne, La Trobe University
Bradley McDonnell, University of Hawai'i at Manoa

INTRO

In spite of the potential generated by recent shifts in the field toward the use of open data formats and tools, a great deal of linguistic research created today is not reproducible, either in principle or in practice. **Much linguistic data is shared without citation, and data sets are not available or accessible even when citations are given.** This project aims to develop data citation and attribution standards for the field of linguistics.

GOALS

- Create guidelines for citation of linguistic data that are applicable across all sub-disciplines.
- Provide a template for data citation in linguistics with relevant examples.
- Integrate the guidelines and template into publication style sheets.
- Recommend practices for scholars and repositories to make data more easily citable and accessible.

CURRENT STATUS

- Completed a series of meetings to identify barriers and issues. Created a Data Management Workshop to educate on data practices like citation and identify pain points for scholars.
- Drafted recommendations and opened for commented by community.
- Solicited comments from linguistics journal editors, society committees

WHAT WE LEARNED

1. It's important to invite different sectors of the community at different stages
2. We identified barriers in our workshops, namely lack of:
   a. Education about citation
   b. Standards
   c. Incentive
3. It's important to make the researchers fully understand why they should engage in this
4. Explain the problem and solution carefully
   a. Austin Principles
   b. Our citation guidelines document
   c. A more in-depth chapter soon to appear in a handbook

WHAT YOU CAN DO

- Model good data citation practices for your field.
- Send the recommendations to linguists you know.
- Endorse the Austin Principles of Data Citation in Linguistics.

When developing new guidelines
for an audience of researchers:
invite the community,
identify barriers,
explain clearly.

For a copy of this poster and materials:

bit.ly/LDIGRDA14

Research Data Alliance Plenary 14 | Helsinki, Finland | 23-25 October 2019

TIMELINE

2013-14

- A survey of data citation practices in 9 linguistics journals over a ten-year period shows that we do not have a culture of citing data in our publications: Berez-Kroeker, Gawne, Kelly & Heston. 2017. *A survey of current reproducibility practices in linguistics journals, 2003-2012.* https://sites.google.com/a/hawaii.edu/data-citation/survey

2014

- Initial funding secured: US$137,657 from the US National Science Foundation, *Developing Standards for Data Citation and Attribution for Reproducible Research in Linguistics*

2015

- Workshop 1: Boulder, Colorado. 26 international linguists & data scientists identified barriers to better data citation in linguistics.

2016

- Workshop 2: Austin, Texas. 34 participants worked with Task Forces to strategize future steps: "Principles & Guidelines" "Citation Formats & Stylesheets" "Attribution for Academic Credit" "Education & Outreach"

2017

- Panel and poster session at the Linguistic Society of America Annual Meeting
- Workshop 3: Austin, Texas. 26 participants gathered to draft a position paper on data citation in linguistics.
- Formation of RDA Linguistic Data Interest Group (LDIG)
- RDA Plenary 10 in Montreal: Drafted the Austin Principles of Data Citation in Linguistics
- Additional funding secured from National Science Foundation
- Workshop on writing data management plans at Linguistic Summer Institute

2018

- RDA Plenary 11 in Berlin: Group session to discuss citation standards recommended by main stakeholders
- Contracted with MIT Press Open to develop the *MIT Open Handbook of Linguistic Data Management*
- Position paper published: Berez-Kroeker et al. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1): 1–18. doi.org/10.1515/ling-2017-0032 (Currently the most downloaded article in *Linguistics*).
- Pre-conference workshop on writing Data Management Plans at the Linguistic Society of America Annual Meeting

2019

- RDA Plenary 13 in Philadelphia: Group session to draft the Recommendations for Data Citation in Linguistics
- LDIG continues to draft standards, shares with community
- "Data Summer Camp" at the Linguistic Summer Institute
- RDM session at Ecole d'été: Phonologie de corpus (Lausanne)
- RDA Plenary 14 in Helsinki: Poster about work and efforts so far

NEXT STEPS

- Finalize recommendations & submit as RDA Supporting Output.
- Ask publications & societies to adopt or endorse.
- Education and outreach.