Citation of Research Data for Linguistic Publications

To our participants

- Feel free to add thoughts and comments in our collaborative notes document.
- Help the organisers with their tracking report, and write your name in the collaborative notes document (alternatively: complete the paper attendance sheet that is circulated)
- All documents will remain available at our RDA13 session page
- Contact person for this meeting: helene.n.andreassen@uit.no

Citation of research data for linguistic publications

Linguistics Data Interest Group Meeting @RDA Plenary 13, Philadelphia
Wednesday April 3, 2019, 14.30-16.00 (Commonwealth C room)

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

Meeting agenda

14.30-14.40
Welcome & introduction to RDA Linguistics Data Interest Group (LDIG), including challenges and (planned) outputs

14.40-15.40
Group editing document on citation of research data

15.40-16.00
Next steps, including timeline

An editor’s thoughts on data collection and (re)use in linguistics

I have come to think of the **accuracy of data as a serious problem** for all linguistics journals, and for the field at large. I do not see any way to deal with it unless we constantly remind ourselves and our students of the importance of working to maintain the quality of the data that we use (p. 409)

[Keren Rice] notes that theories are often based on a **misunderstanding** of the primary sources or on an **inappropriately restricted** subset of the data available in the primary sources (p. 412)

it is vital for all authors to ensure '**clarity and replicability of the chain of evidence**' so that it will be as easy as possible for other scholars 'to evaluate the solidity of the various steps in the chain, and then to replicate and extend the work the claim is based on, if they choose to' (Mark Liberman, via email, 1993) (p. 410)

(Thomason, 1994, bolding ours)

Rationale for citation standards in linguistics

Position statements

1. Berez-Kroeker et al. *Reproducible research in linguistics: A position statement on data citation and attribution in our field.* Linguistics 56:1, 2017 (https://doi.org/10.1515/ling-2017-0032)

2. Ted Pedersen: *Empiricism Is Not a Matter of Faith*. Computational Linguistics 34:3, 2008 — *The sad tale of the Zigglebottom tagger*

Estimates in chemistry and medicine: only 20–25% of research can be replicated.

Linguistics Data Interest Group (LDIG)

Endorsed in 2017, co-chaired by Helene N. Andreassen (UiT), Andrea Berez-Kroeker (U. Hawai‘i at Manoa) & Lauren Gawne (La Trobe)

Main objectives

1. Development and adoption of common principles and guidelines for data citation and attribution (researchers, professional organizations, academic publishers, archives).

2. Education and outreach efforts (practical training and awareness of principles/sociological change).

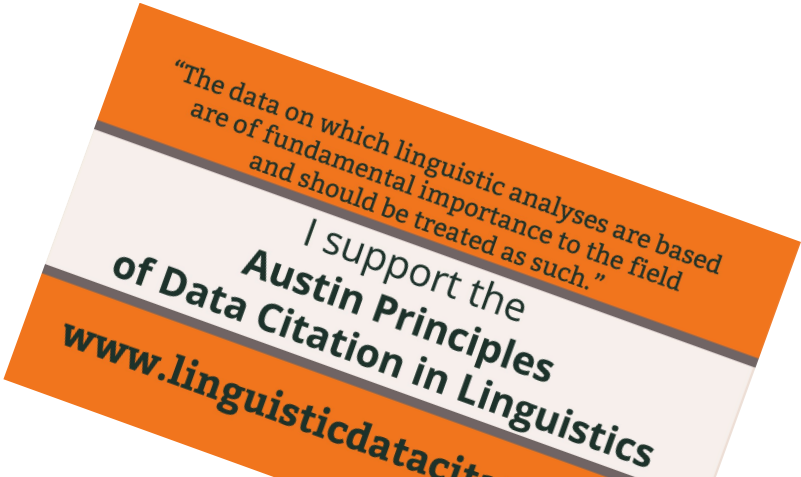
3. Greater attribution of linguistic data set preparation within the linguistics profession (value “data work” as scholarly output at all career stages).

From the LDIG charter

LDIG outputs

Output 1: Austin Principles of Data Citation in Linguistics

- Set of guidelines that enable linguists to make informed decisions regarding the accessibility and transparency of their research data.
- Based on the FORCE11 Joint Declaration of Data Citation Principles.
- Available at linguisticsdatacitation.org (Berez-Kroeker et al., 2017a).



LDIG outputs

Output 2: Recommendations for data citation in linguistics

1. A short version containing citation formats and examples, relevant to all subfields of linguistics, all types of linguistic data.

   To be shared with editors and stylesheet curators for inclusion in existing stylesheets (Unified Style Sheet for Linguistic Journals, by Joseph et al., 2007, and the Generic Style Rules, by Haspelmath, 2014).

   To be shared with researchers needing guidelines on how to cite their own and other people's data.

   To be shared with data providers who host citable linguistic data, encourage adoption of metadata standards.

LDIG outputs

Output 2: Recommendations for data citation in linguistics

2. A long version containing explanations and examples of citation format elements.

   Designed for a general linguistics audience to teach them more about why and how to cite data.

3. Chapter on citation in the MIT Open Handbook of Linguistic Data Management.

   Forthcoming 2020, eds. Berez-Kroeker, McDonnell, Koller & Collister

The short version: The draft

https://docs.google.com/document/d/1_r2D9ReMTin3_qUlfEO48z5X54otTX33m9Sujso68tI/edit#

Thanks to the LDIG members who have already shared their thoughts with us in writing!

Write down your questions and comments while we go through this. There will be plenty of time for Q&A and discussion afterwards!

Discussion

1. Bibliographic / references list section: Discussion

   Your questions
   Feedback

2. In-text citation style: Your input needed!

   1-2-4 style - borrowed from BoF Content Mining session (thank you!)

   1: 2 minutes to look at the document on your own and note your questions/ideas
   2: 3 minutes to share with another person and discuss questions/ideas/feedback you had
   4: 5 minutes to discuss in small groups the questions, with a report out

Next steps

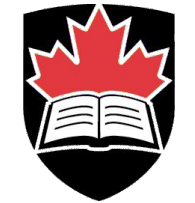
- Completion of written documents, including feedback from community.
- RDA official endorsement to assure quality
- Publication
  - Short version on GitHub or other website
  - Long version on https://site.uit.no/linguisticsdatacitation/

Next steps

- Integration into stylesheets & dissemination (publishers, researchers, data providers)
- To make this happen: 1) timeline, 2) collaboration & division of tasks

    - Mailing-lists and networks, open webinars, training sessions
    - Package of Citation Guidelines and Austin Principles sent to Linguistic Society of America for endorsement (May 2019?)
    - Session at Linguistic Society of America meeting in January 2020
    - Chapter in Open Handbook of Linguistic Data Management (forthcoming 2020)

Thanks to our institutions and sponsors!



Carleton UNIVERSITY

UiT / THE ARCTIC UNIVERSITY OF NORWAY

LA TROBE UNIVERSITY

CLARIN

Common Language Resources and Technology Infrastructure

ands

AUSTRALIAN NATIONAL DATA SERVICE

References

Berez-Kroeker, A. L., Andreassen, H. N., Gawne, L., Holton, G., Kung, S. S., Pulsifer, P., Collister, L. B., The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. (2017a). *Draft: The Austin Principles of Data Citation in Linguistics (Version 0.1).* http://site.uit.no/linguisticsdatacitation/austinprinciples/ Accessed 19.03.2018.

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., . . . Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics, 56*(1), 1-18. https://doi.org/10.1515/ling-2017-0032

Haspelmath, M. (2014). *The Generic Style Rules for Linguistics.* https://doi.org/10.5281/zenodo.253501

Joseph, B. et al. (2007). *Unified Style Sheet for Linguistics*. https://www.linguisticsociety.org/resource/unified-style-sheet

Pedersen, T. (2008). Empiricism is not a matter of faith. Computational Linguistics, 34(3).

Thomason, S. G. (1994). The Editor's Department. *Language, 70*(2), 409-413.