



Data citation in linguistics publications: A scholar-led, community-based initiative

Helene N. Andreassen, UiT The Arctic University of Norway
Andrea Berez-Kroeker, University of Hawaii at Manoa
Lauren Collister, University of Pittsburgh
Phillip Conzett, UiT The Arctic University of Norway
Christopher Cox, Carleton University
Koenraad De Smedt, University of Bergen
Lauren Gawne, La Trobe University
Bradley McDonnell, University of Hawaii at Manoa





Overview of presentation

Background: Linguistics and linguistic data

Our project: Network building, deliverables and outreach activities

What we have learned

Background: What is linguistics?



Broadly: The study of human language.

Better:

Data-driven social science in which inferences about cognition and social structure are drawn from **observations of language use**.

Primary data underlying the research:

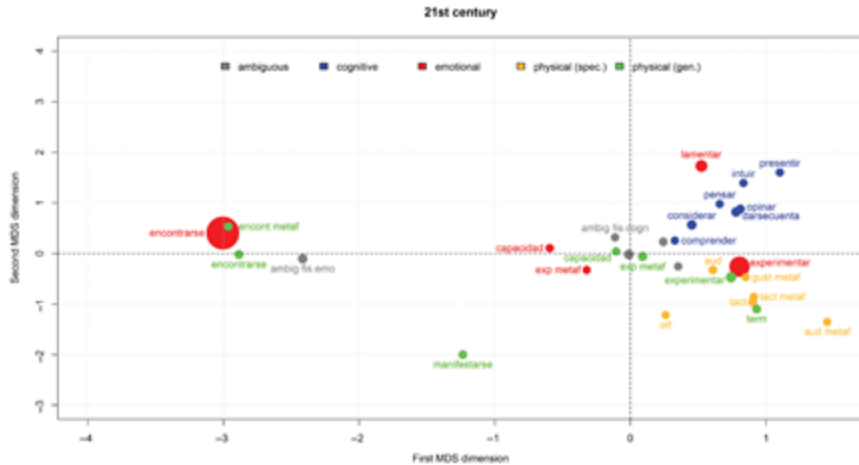
Records of language

Annotations of those records

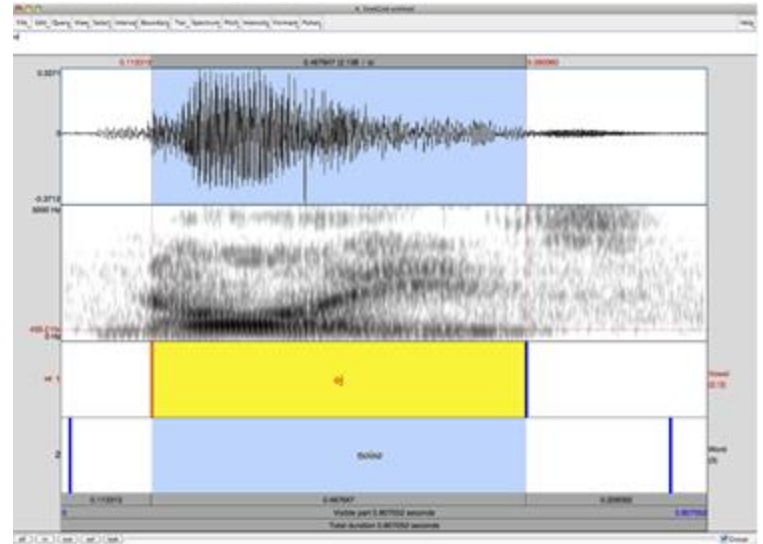


Ak'a-ggem ayag-llru-uq already-INFER leave-PAST-3 “It seems he already left.”

Payne 1997:253



Jansegers & Gries 2017:10



Styler 2017:54

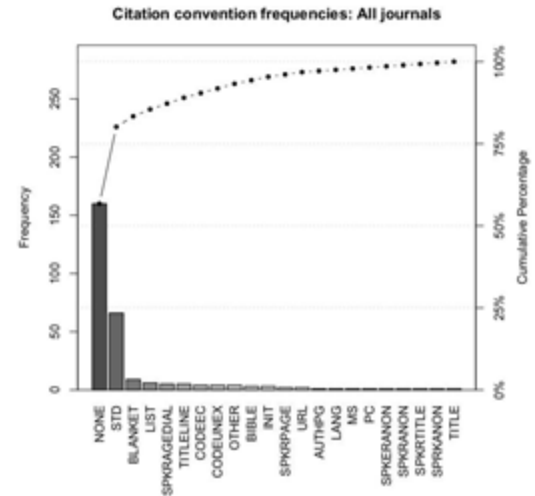
Background: Linguists don't cite data (much)

But!

Data in publications **don't generally have citations**

(cf Berez-Kroeker et al. 2017)

If they do, citation only vaguely linked to the actual data making reproducible research very hard.





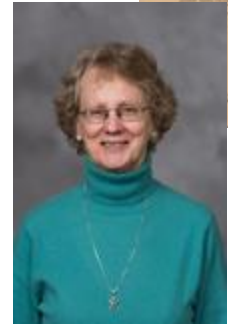
Background: A long-noticed problem

1994: Editor of *Language*, top journal in the field

Found many cases where use of data was problematic

“...so frequently, in fact, that the assumption that the **data in accepted papers is reliable** began to look questionable” (Thomason 1994:409)

Exhortation to use data carefully,
Describe and cite sources well,
Say how data was collected.





Background: Why now?

Language data are precious:

Captures world-views

Captures cultures at a certain point in time, and their contact over time with each other

Captures cognitive capacities and variation (grammar, acoustic properties)



Background: Why now?

Technological infrastructure exists for
Creating vastly more data
Properly archiving and citing data

Precious language data are useless unless we archive and **cite them** according to best practices.



Network building, Phase 1



“Data Citation and Attribution in Linguistics” grant (NSF since 2015) <https://sites.google.com/a/hawaii.edu/data-citation/>

3 multi-day workshops with 40+ participants

Identified three barriers to citation of linguistic data:

No understanding of why (no culture of doing it)

No rewards for it

No guidelines or formats to follow





Network building, Phase 2



Research Data Alliance (RDA):

“RDA provides a **neutral space** where its members can come together to **develop and adopt infrastructure** that promotes **data-sharing and data-driven research.**” (<https://www.rd-alliance.org/about-rda>)

Interest Groups, Working Groups, national groups.
Adopts recommendations and outputs.

Linguistic Data Interest Group (LDIG): network of ~100 international linguists.

Work at Plenaries and virtually to develop deliverables specifically aimed at linguistic researchers.

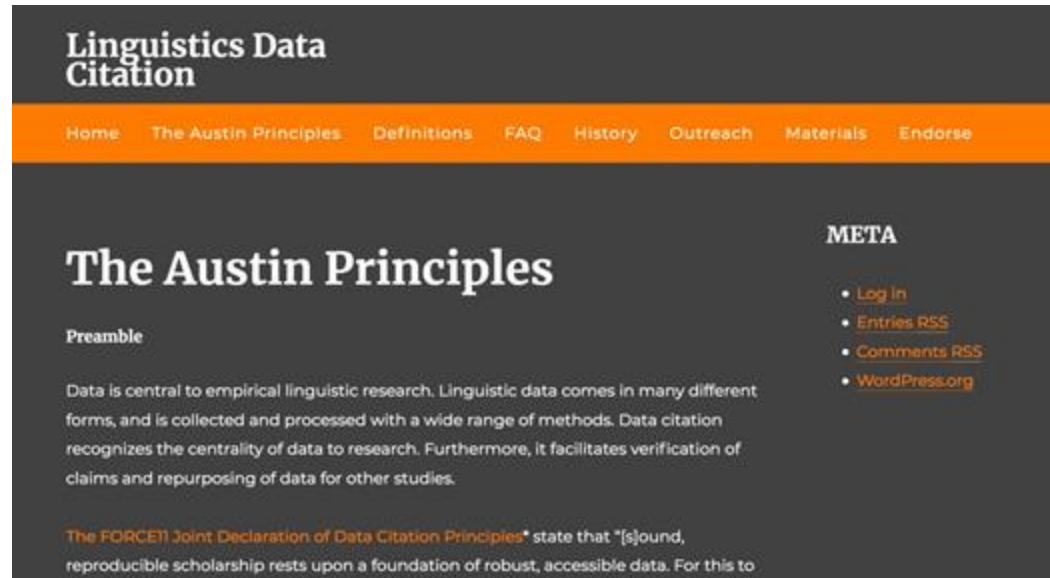
Deliverables

The Austin Principles of Data Citation in Linguistics (2018)

Set of guidelines to help linguists to make informed decisions regarding the accessibility and transparency of their research data.

Based on the FORCE11 Joint Declaration of Data Citation Principles.

www.linguisticsdatacitation.org



The screenshot shows the website for Linguistics Data Citation. The header includes the title "Linguistics Data Citation" and a navigation menu with links for Home, The Austin Principles, Definitions, FAQ, History, Outreach, Materials, and Endorse. The main content area features the title "The Austin Principles" and a section for the Preamble, which states: "Data is central to empirical linguistic research. Linguistic data comes in many different forms, and is collected and processed with a wide range of methods. Data citation recognizes the centrality of data to research. Furthermore, it facilitates verification of claims and repurposing of data for other studies." Below this, a quote from the FORCE11 Joint Declaration of Data Citation Principles is visible. On the right side, there is a "META" section with links for Log In, Entries RSS, Comments RSS, and WordPress.org.



Deliverables

Publications

2018: Open access position paper on reproducibility in linguistics.

Most downloaded article of the journal.

To appear: The Open Handbook of Linguistic Data Management, MIT Press Open (Berez-Kroeker, McDonnell, Koller & Collister, eds.).

13 chapters on conceptual foundations of data management for linguistics and best practices. 50 short data management use cases. Appr. 90 authors from four continents.



Deliverables

Recommendations for citation of research data
in linguistics (working title)

Citation model for in-text citations and
bibliographic references, including commented
examples and elaborated definitions.

Intended audience: Editors of linguistic
publications, researchers, and repositories.



Deliverables

Engaging the community

Several asynchronous meetings with the LDIG community.

Invitation to comment sent to “VIPs” (editors, directors, presidents, and other linguists active in the Open Science movement).

Prignitz, Gisèle. 2007 (collection date). Enquête Burkina Faso. Projet PFC. <https://public.projet-pfc.net/>. (Accessed 2019-06-22).

Example 4: Different kinds of Locators/
The Prignitz dataset from Example 3 does not have a GUID or URI, so the URL to the landing page of the main collection is used as the **Locator**. The Adelaar dataset from Example 1 has both a GUID (in this case, a DOI) and a **repository-internal identifier** (AA4), both are used to aid locating the resource. For the Mæhlum dataset below, which is published on a physical CD audio, media is specified.

Mæhlum, Brit. 1998. *Dialektprøver fra Måselv og Bardu*. Måselv målag. CD audio.

The template for a **minimal reference to a dataset resource in the bibliography** section of a piece of academic writing is:
Author, Date, Title, Publisher, Locator.

The template for an **expanded bibliographic reference** to a dataset resource, including **conditional elements** (i.e. required in certain cases depending on resource characteristics) is:
Author, Date, Title, Publisher, Locator, Version, Date accessed, Tag.

In-text (or in-line) citations must point to a bibliographic reference at the end of the published work. The template for a **minimal in-text citation** is:
Author, Date

The template for an **expanded in-text citation** including additional potential information is:
Author, Date, Locator, Subset, Other Attribution (Roles)

Please note: Definitions of the elements contained in the bibliographic reference and the in-text citation can be found in the [Glossary](#). A longer version of the recommendations, explaining concepts, highlighting challenges and providing examples can be found in: Konzett, Philipp & Koenraad De Smedt. (in preparation). Guidance for citing research data. In Andrea L. Berez-Kroecker, Bradley McDonnell, Lauren Collister & Eve Koeller (eds.), *Open Handbook of Linguistic Data Management*. MIT Press Open.

8:39 PM Sep 25
I am not sure that annotations on dates are a good idea. Complex dates such as "2000-2002" or "1999[1943]" are often needed, but "2007 (collection date)" seems to break the convention that dates should be integers without a very good motivation. If you do a lookup for all works of a decade, for instance, this

not be retrieved.
if there is no "real" date of seems to be not a very piece of information that at all costs.

8:56 PM Sep 18
Delete: "r"

8:54 AM Sep 12
Add: "These recommendations are for the use of researchers and scholars in the field working with datasets..."

8:55 AM Sep 12
Sorry, that sentence was too long for my tired brain

8:40 AM Sep 18
Delete space

AM Sep 17
experience, URI and URL have interchangeable so often and that any subtle differences two are lost on most from a usability/readability , I'd recommend to stick with

AM Sep 17
be categorized as PID e terminology above, right?

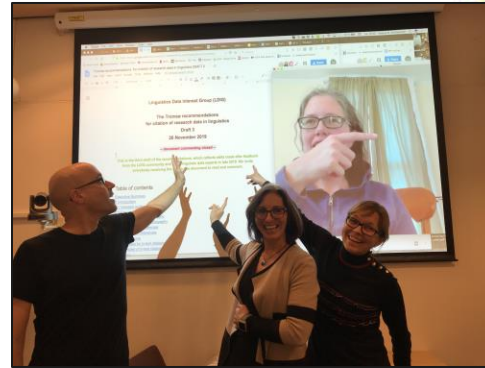
Deliverables

Calendar

November 2019: Revision and finalisation of recommendations.

January 2020: RDA community review.

February 2020: Published with doi as an RDA Supporting Output, outreach to intended audience.



Outreach activities

Impromptu talks at conferences, pubs, etc.

Short-form courses at gatherings of the Linguistic Society of America:

- Writing Data Management Plans
- Data Science for Reproducible Linguistics
- Data Summer Camp

PhD summer school on corpus phonology in Lausanne (Switzerland)

- Treatment of acoustic data from A to Z, including transparency of research and best practices of research data management.



Image: To Deposit or Not to Deposit, by Ainsley Seago, CC-BY. [dx.doi.org/10.1371/journal.pbio.1001779.g001](https://doi.org/10.1371/journal.pbio.1001779.g001)



What we have learned

Researchers experience barriers to data citation, such as the lack of

- Awareness
- Training
- Standards
- Incentives



What we have learned

It is important to invite different sectors of the community at all stages

- To identify practices and challenges
- To get feedback on ongoing work
- To implement deliverables in the research and publication process



What we have learned

In a busy world, it is challenging to fully engage the different sectors

- For many researchers, moving from good intentions to practice takes time.
- For many editors, other aspects of the publishing process are considered more pressing.
- For many repositories, the best practices of data citation are not reflected in the metadata and documentation guidelines.



What we have learned

Continuous outreach seems to move things (slowly) forward

- Concrete deliverables are key
- Right context, looking for opportunities for outreach
- Enough time for presentation, Q&A
- Getting the right people on board, trend-setters in the community



Acknowledgments

Many thanks to the participants in the Data Citation and Attribution project, the Data Science for All of Linguistics project, members of the RDA LDIG, and attendees at our previous workshops, courses and presentations for fruitful discussion.



This material is based upon work supported by the National Science Foundation under Grants No. 1447886 and 1745349. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.



References

- Berez-Kroeker, Andrea L., Lauren Gawne, Barbara F. Kelly & Tyler Heston. 2017. A survey of current reproducibility practices in linguistics journals, 2003-2012. <https://sites.google.com/a/hawaii.edu/data-citation/survey>.
- Berez-Kroeker, Andrea L., Lauren Gawne, Susan Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1): 1–18.
- Berez-Kroeker, Andrea L., Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. 2018. The Austin Principles of Data Citation in Linguistics. Version 1.0. <http://site.uit.no/linguisticsdatacitation/austinprinciples/> Accessed 26 Nov. 2019.
- Berez-Kroeker, Andrea L., Bradley McDonnell, Eve Koller & Lauren Collister. In prep. The Open Handbook of Linguistic Data Management. Cambridge, MA: MIT Press Open.
- Jansegers, Marlies & Stefan Th. Gries. 2017. Towards a dynamic behavioral profile: A diachronic study of polysemous sentir in Spanish. *Corpus Linguistics and Linguistic Theory* 1-43.
- Payne, Thomas E. 1997. *Describing Morphosyntax: A Guide for Field Linguists*. Cambridge: Cambridge University Press.
- Styler, Will. 2017. Using Praat for Linguistic Research, version 1.8.1. Online: <http://savethevowels.org/praat>. Accessed 20 November 2019.
- Thomason, Sarah. 1994. The editor's department. *Language* 70: 409-413.