Using the Tromsø Recommendations to cite data in language work

Helene N. Andreassen (@n_andreassen), Andrea L. Berez-Kroeker, Lauren Gawne (@superlinguo), Philipp Conzett (@PhilippConzett), Koenraad De Smedt, Christopher Cox, Lauren B. Collister (@parnopaeus)

bit.ly/TRecsICLDC21 #lingdata
Andrea, Helene & Lauren

Montréal, September 2017

Photo: private
Overview: The Tromsø Recommendations For Citation of Research Data in Linguistics

Language and linguistics datasets are often not cited well.

The people involved in creating language data are not receiving proper recognition.

The Tromsø Recommendations detail how to cite data - and people! - in language work.
What do we mean by “language work”?

Leonard 2017:

“an umbrella expression to include language documentation, description, teaching, advocacy, and resource development” (2017:16)

...everything that ICLDC participants do!
What do we mean by “data” in language work?

**Data** in language work means all samples of language:

- Recordings and written language
- Words, sentences, verbal art, storytelling, oration, song, etc.

**People** create and contribute to language data.

Data in language work is **precious** because language is about **people**.
What do we mean by “citing” data?

People are speakers, learners, elders, youth, teachers, helpers, parents, grandparents, researchers, culture-bearers, translators, poets, authors and more.

People deserve to be thanked - credited - acknowledged for their contributions in language work.

Citation is one way to do that.

We do this for publications All. The. Time.
Typical citations (of publications, not data)


(Acknowledgment given to author and editors)
Hmmm…. we’ve been here before

ICLDC 2015: language documenters aren’t that great at citing data.

(Gawne et al. 2015, Gawne et al. 2017)

Which means we aren’t giving proper credit to the people involved in language work.

Why not?

Linguistics doesn’t have a history of requiring the citation of data.

(even though we do have a history of requiring the citation of publications)
Documenters don’t really do this for data

In 2015/2017 we were concerned about the effects of this on our science.

At this ICLDC today we are concerned about the effects of this on people who are being left out.

People serve a lot of roles in language work.
Many roles for people in language work

Author  Translator  Illustrator  Participant
Editor  Recorder  Participant  Depositor
Speaker  Data inputter  Interviewer  Developer
Signer  Consultant  Compiler  Sponsor

So why aren’t we giving people credit for these roles through proper citation?
Because we don't know how!

Part of the problem is that we don’t know how or why to cite data (Berez-Kroeker et al. 2018).

2017: Research Data Alliance Linguistic Data Interest Group (link below!)

Two key publications:

For the WHY: The Austin Principles of Data Citation (link below!)

For the HOW: Tromsø Recommendations for Citation of Research Data in Linguistics (link below!)
Austin Principles of Data Citation in Linguistics

www.linguisticsdatacitation.org

2. Credit and Attribution

Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

In linguistics, citations should facilitate readers retrieving information about who contributed to the data, and how they contributed, when it is appropriate to do so. One way to do this is through citations that list individual contributors and their roles. Another way is by using citations that link to metadata about contributors and their roles.

Model: FORCE11 Joint Declaration of Data Citation Principles
The Tromsø recs

- Minimal and expanded templates for in-text citations & bibliographic references
- Explanation of elements in clear terms
- Examples from real linguistic data
- Highlights issues that are important to linguistic data
The Tromsø recommendations for citation of research data in linguistics

Developed through asynchronous meetings of the LDIG, plus invited input from VIPs

**Aim:** Practical and concise advice for data citation, with consideration of the variety of linguistic data

**Intended audience:** Editors of linguistic publications, researchers, and repositories.
The Tromsø recommendations - Outline

Includes:

● Recommendations for in-text citation and bibliographic reference
● Full data set and specific example citation
● Examples of citation using real data
● Flexibility to fit with journal style guidelines

Executive Summary

0. Introduction
  0.1 Intended audience
  0.2 General recommendations

1. References in the bibliography
  1.1 What to cite in the bibliography
  1.2 Templates for references
  1.3 Examples of references
    1.3.1 Citing a full dataset
    1.3.2 Citing a component of a dataset

2. In-text citations
  2.1 Templates for in-text citations
  2.2 Examples of in-text citations

3. Glossary

4. About this Document and Contributors
Of note here: Roles

The T-Recs allow you to give credit to many people and explain the role they played.

For example: the data collector:

[https://doi.org/10.4225/72/56E979455A05E](https://doi.org/10.4225/72/56E979455A05E).
Also, researchers, depositors, speakers, consultants, interviewers...


Even for in-line citations:

(Hauk 2018: BH2-081, 00:00:01–00:00:03, Rezo Orbetishvili (Speaker))
There are many lists of Roles you can use

Some standard lists of contributor roles include

CASRAI

DataCite

OLAC role vocabulary
Next: Data citation in your work

Working towards normalising the practice of citing linguistic data

Relevant to everyone who works with linguistic data
Language workers: Cite your data

Build data citation into projects from the beginning

Data citation is distinct from, but closely related to, making underlying data available. Citation co-exists with ethical approaches to archiving and access

Cite other people’s linguistic data if you use it in your work
Supervisors & project managers: Encourage best practice

Introduce students to best practice in the field (cf. Pawley 2014)

Example: Data citation and archiving have been expectations at University of Hawaii since Fall 2013. Included in PhD student handbook.
Publishers: Make citation an expectation

The Trømso Recommendations can be adopted by any journal or publication

**Example:** The [Australian Journal of Linguistics](https://www.publish.csiro.au/journals/alql) guidelines include the Generic Style Rules for Linguistics, the Leipzig Glossing Rules, the Austin Principles & the Trømsø Recommendations:

“For research based on original fieldwork or archival documentary materials, authors must provide the sources and provenance of data, as well as the methods used to collect it, including the time period and locations in which fieldwork was conducted”.
Data managers: Encourage citation

Data managers can provide training and support to encourage citation.

**Example:** PARADISEC provides a “cite as” field on all pages of the archive, giving a formatted citation to the relevant level of granularity.

![Citation Example](https://dx.doi.org/10.4225/72/56E976A071650)
Normalising data citation in language work

Language documentation and reclamation is about people. People make language records.

Proper citation of linguistic records (data) gives credit to everyone involved.

The Tromsø Recommendations provide practical examples for how to cite linguistic data.
References

- Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk
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.

Slides: [bit.ly/TRecsICLDC21](bit.ly/TRecsICLDC21)
Background: What is linguistic data?

http://ase.tufts.edu/psychology/psycholinglab/asl-lex/visualization.html

Jansegers & Gries
2017:10

Payne 1997:253

Styler
2017:54

Ak’a-ggem ayag-lrulu-uq already-INFER leave-PAST-3 “It seems he already left.”
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.
Overview of presentation

Background: Citing linguistic data

Outline: The main features of the Tromsø Recommendations

Next: Building support for data citation in your work and community
Background: A long-noticed problem

“It is simply a feature of a scientific enterprise to make one’s primary data accessible to further scrutiny”

(Himmelmann 1998: 165)

See also Gawne & Berez-Kroeker (2018)
Background: Linguists don't cite data (much)

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 set, making reproducible research very hard.
Background: LDIG

Related LDIG 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.
Background: LDIG

2017: Research Data Alliance Linguistic Data Interest Group founded

First publication: The Austin Principles of Data Citation
Explains the importance of why to cite data, but not how.

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.
Background References