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

BY NC SA

Andrea, Helene & Lauren

Montréal, September 2017

Photo: private

BY NC SA

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.

BY NC SA

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?

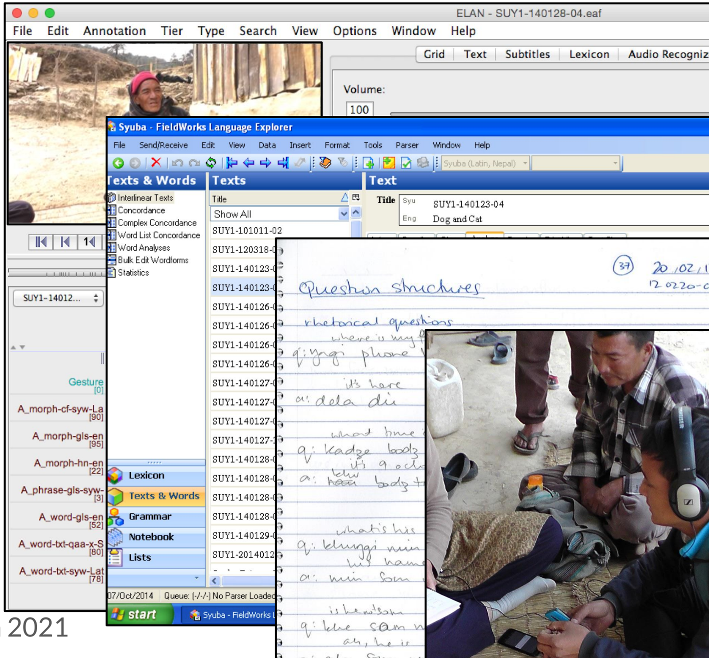
Data in language work means all samples of language:

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**.



BY NC SA

What do we mean by “citing” data?

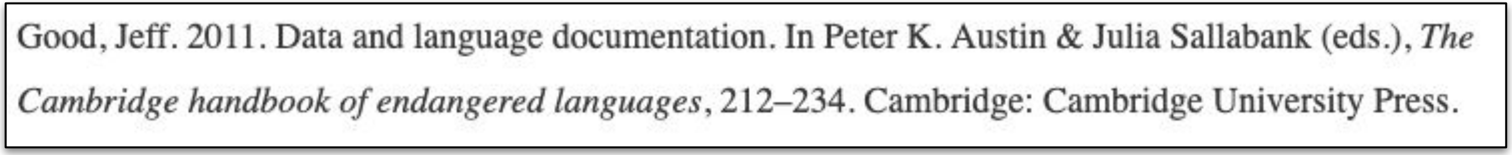
People are speakers, learners, elders, youth, teachers, helpers, parents, grandparents, researchers, culture-bearers, translators, poets, authors and more.

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

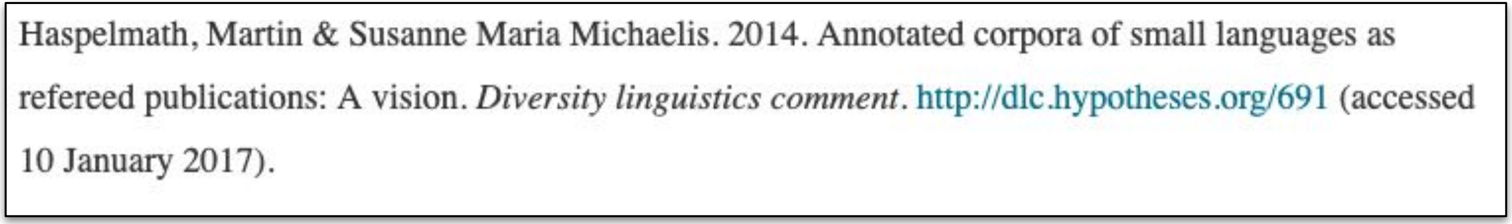
Citation is one way to do that.

We do this for publications **All. The. Time.**

Typical citations (of publications, not data)



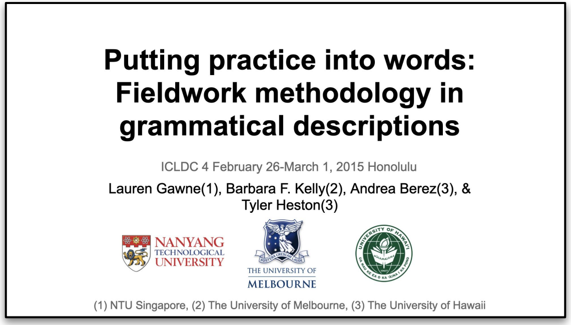
Good, Jeff. 2011. Data and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 212–234. Cambridge: Cambridge University Press.



Haspelmath, Martin & Susanne Maria Michaelis. 2014. Annotated corpora of small languages as refereed publications: A vision. *Diversity linguistics comment*. http://dlc.hypotheses.org/691 (accessed 10 January 2017).

(Acknowledgment given to **author** and **editors**)

Hmmm.... we've been here before



Putting practice into words: Fieldwork methodology in grammatical descriptions

ICLDC 4 February 26-March 1, 2015 Honolulu

Lauren Gawne(1), Barbara F. Kelly(2), Andrea Berez(3), & Tyler Heston(3)

(1) NTU Singapore, (2) The University of Melbourne, (3) The University of Hawaii

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*)

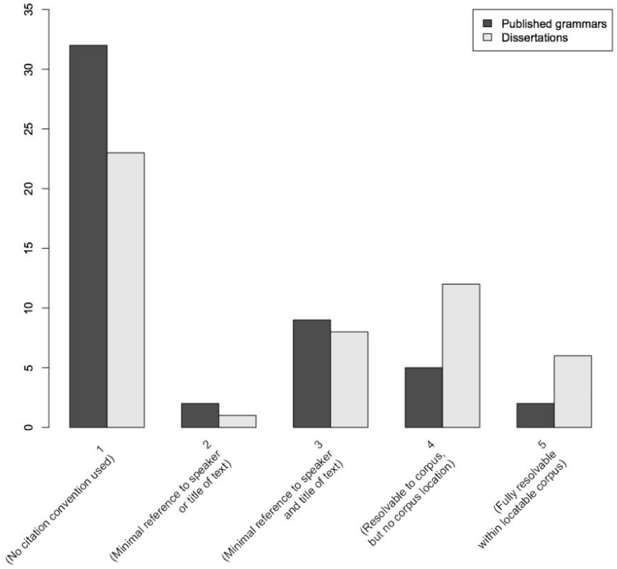
BY NC SA

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.



Gawne et al. 2017

Many roles for people in language work

Author

Editor

Speaker

Signer

Translator

Recorder

Data inputter

Consultant

Illustrator

Participant

Interviewer

Compiler

Participant

Depositor

Developer

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!)

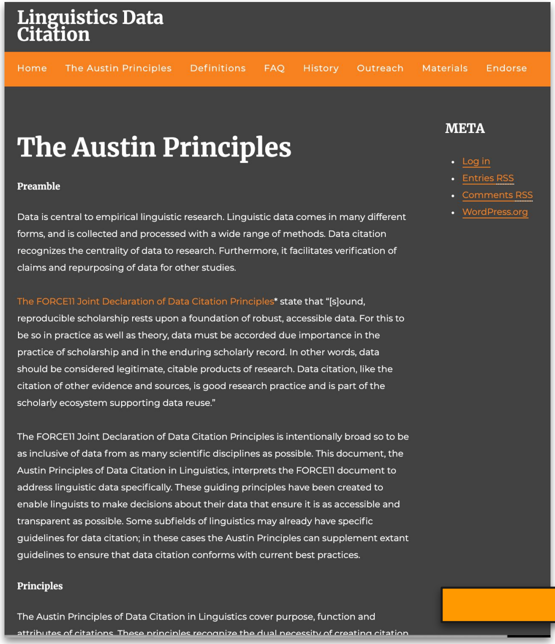
Two key publications:

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



Linguistics Data Citation

The Austin Principles

Preamble

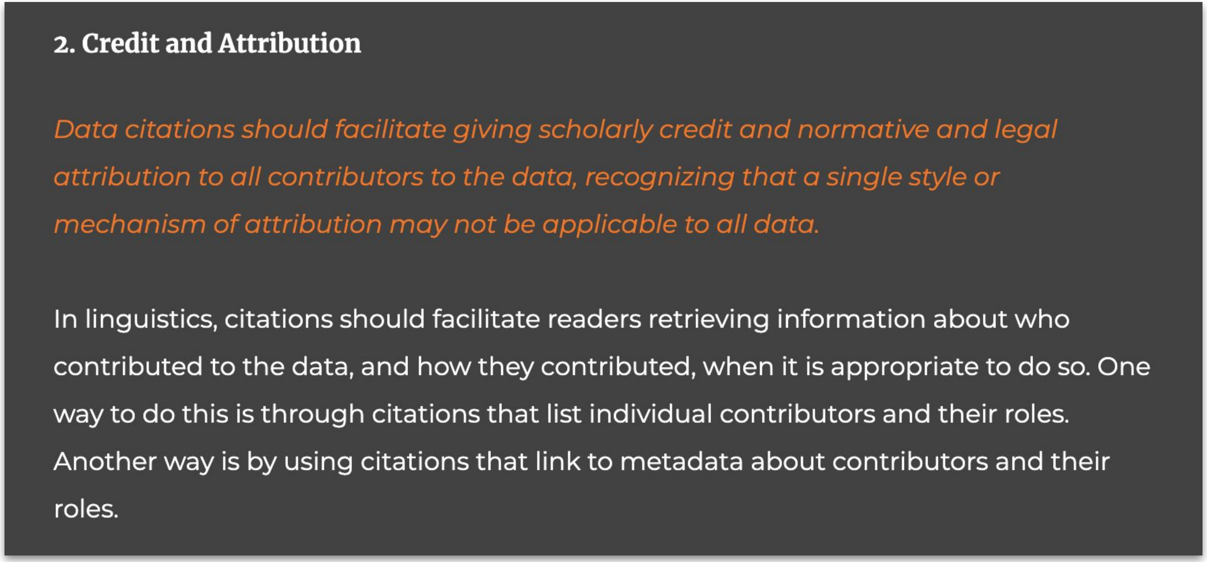
Data is central to empirical linguistic research. Linguistic data comes in many different forms, and is collected and processed with a wide range of methods. Data citation recognizes the centrality of data to research. Furthermore, it facilitates verification of claims and repurposing of data for other studies.

The FORCE11 Joint Declaration of Data Citation Principles* state that "[s]ound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse."

The FORCE11 Joint Declaration of Data Citation Principles is intentionally broad so to be as inclusive of data from as many scientific disciplines as possible. This document, the Austin Principles of Data Citation in Linguistics, interprets the FORCE11 document to address linguistic data specifically. These guiding principles have been created to enable linguists to make decisions about their data that ensure it is as accessible and transparent as possible. Some subfields of linguistics may already have specific guidelines for data citation; in these cases the Austin Principles can supplement extant guidelines to ensure that data citation conforms with current best practices.

Principles

The Austin Principles of Data Citation in Linguistics cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation



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.

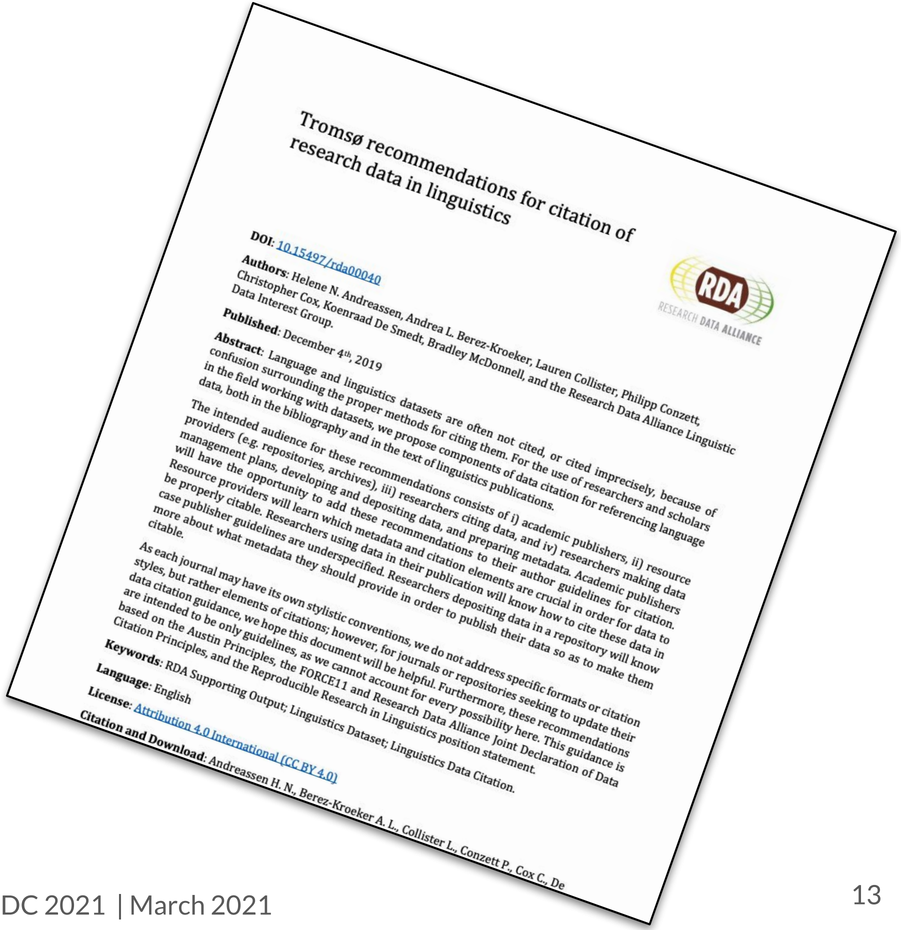
In linguistics, citations should facilitate readers retrieving information about who contributed to the data, and how they contributed, when it is appropriate to do so. One way to do this is through citations that list individual contributors and their roles. Another way is by using citations that link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

Model: FORCE11 Joint Declaration of Data Citation Principles

cc BY NC SA

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



Tromsø recommendations for citation of research data in linguistics

DOI: 10.15497/rda00040

Authors: Helene N. Andreassen, Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and the Research Data Alliance Linguistic Data Interest Group.

Published: December 4th, 2019

Abstract: Language and linguistics datasets are often not cited, or cited imprecisely, because of confusion surrounding the proper methods for citing them. For the use of researchers and scholars in the field working with datasets, we propose components of data citation for referencing language data, both in the bibliography and in the text of linguistics publications.

The intended audience for these recommendations consists of i) academic publishers, ii) resource providers (e.g. repositories, archives), iii) researchers citing data, and iv) researchers making data management plans, developing and depositing data, and preparing metadata. Academic publishers will have the opportunity to add these recommendations to their author guidelines for citation. Resource providers will learn which metadata and citation elements are crucial in order for data to be properly citable. Researchers using data in their publication will know how to cite these data in case publisher guidelines are underspecified. Researchers depositing data in a repository will know more about what metadata they should provide in order to publish their data so as to make them citable.

As each journal may have its own stylistic conventions, we do not address specific formats or citation styles, but rather elements of citations; however, for journals or repositories seeking to update their data citation guidance, we hope this document will be helpful. Furthermore, these recommendations are intended to be only guidelines, as we cannot account for every possibility here. This guidance is based on the Austin Principles, the FORCE11 and Research Data Alliance Joint Declaration of Data Citation Principles, and the Reproducible Research in Linguistics position statement.

Keywords: RDA Supporting Output; Linguistics Dataset; Linguistics Data Citation.

Language: English

License: Attribution 4.0 International (CC BY 4.0)

Citation and Download: Andreassen H. N., Berez-Kroeker A. L., Collister L., Conzett P., Cox C., De

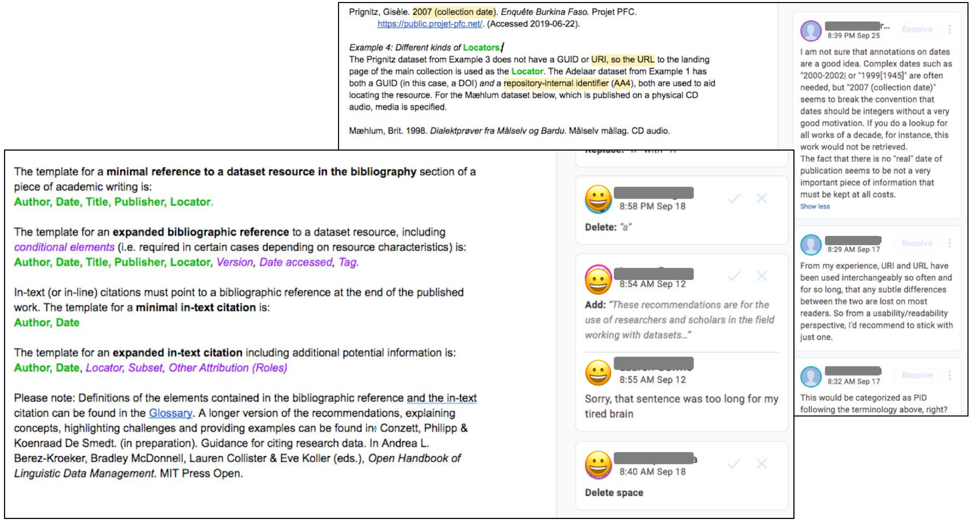
BY NC SA

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.



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 Målselv og Bardu*. Målselv mållag. CD audio.

[redacted] 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[1945]" 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 work would not be retrieved.
The fact that there is no "real" date of publication seems to be not a very important piece of information that must be kept at all costs.

[redacted] 8:29 AM Sep 17

From my experience, URI and URL have been used interchangeably so often and for so long, that any subtle differences between the two are lost on most readers. So from a usability/readability perspective, I'd recommend to stick with just one.

[redacted] 8:32 AM Sep 17

This would be categorized as PID following the terminology above, right?

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: Conzett, Philipp & Koenraad De Smedt. (in preparation). Guidance for citing research data. In Andrea L. Berez-Kroeker, Bradley McDonnell, Lauren Collister & Eve Koller (eds.), *Open Handbook of Linguistic Data Management*. MIT Press Open.

[redacted] 8:58 PM Sep 18

Delete: *"a"*

[redacted] 8:54 AM Sep 12

Add: *"These recommendations are for the use of researchers and scholars in the field working with datasets..."*

[redacted] 8:55 AM Sep 12

Sorry, that sentence was too long for my tired brain

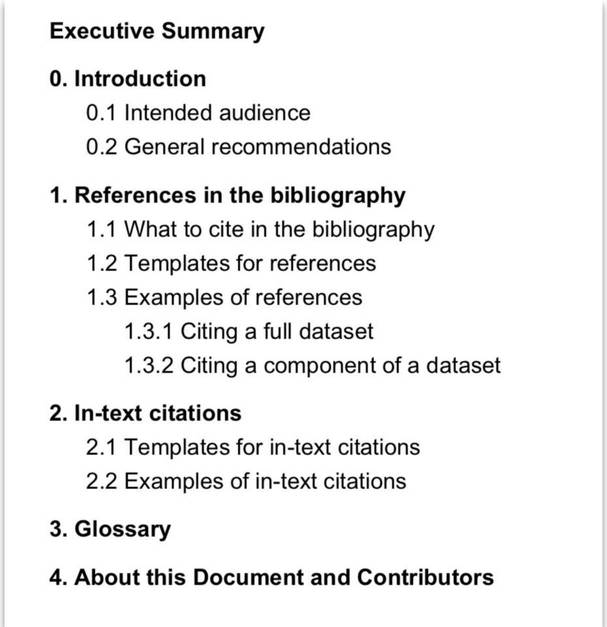
[redacted] 8:40 AM Sep 18

Delete space

The Tromsø recommendations - Outline

Includes:

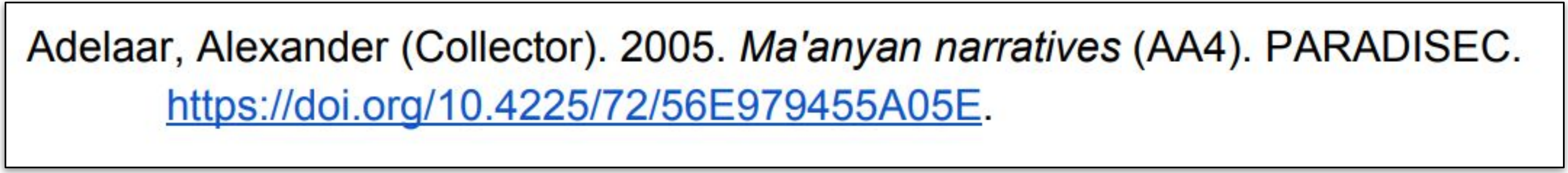
- Recommendations for in-text citation and bibliographic reference
- Full data set and specific example citation
- Examples of citation using real data
- Flexibility to fit with journal style guidelines



Of note here: Roles

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

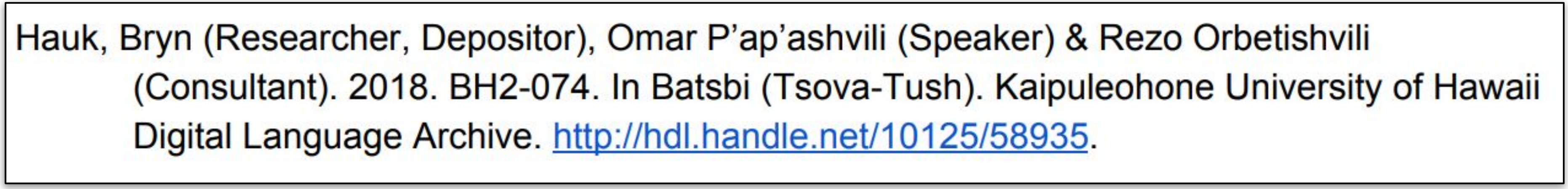
For example: the data **collector:**



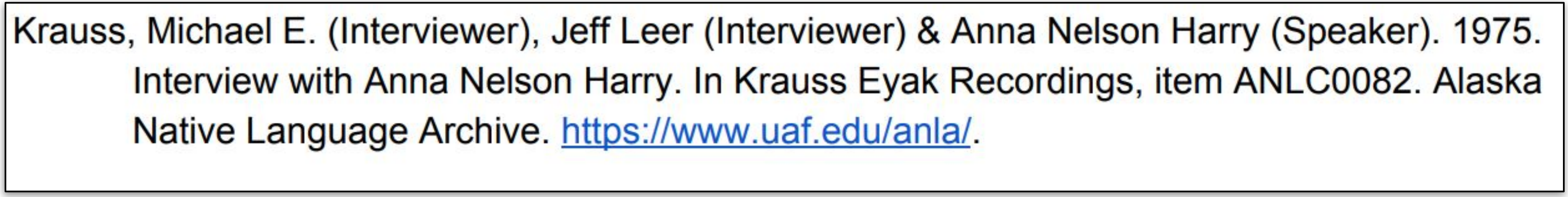
Adelaar, Alexander (Collector). 2005. *Ma'anyan narratives* (AA4). PARADISEC. https://doi.org/10.4225/72/56E979455A05E.

BY NC SA

Also, **researchers, depositors, speakers, consultants, interviewers**…

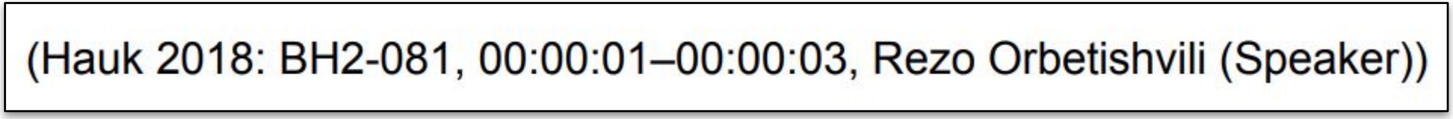


Hauk, Bryn (Researcher, Depositor), Omar P’ap’ashvili (Speaker) & Rezo Orbetishvili (Consultant). 2018. BH2-074. In Batsbi (Tsova-Tush). Kaipuleohone University of Hawaii Digital Language Archive. http://hdl.handle.net/10125/58935.



Krauss, Michael E. (Interviewer), Jeff Leer (Interviewer) & Anna Nelson Harry (Speaker). 1975. Interview with Anna Nelson Harry. In Krauss Eyak Recordings, item ANLC0082. Alaska Native Language Archive. https://www.uaf.edu/anla/.

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.

BY NC SA

Publishers: Make citation an expectation

The Trømso Recommendations can be adopted by any journal or publication

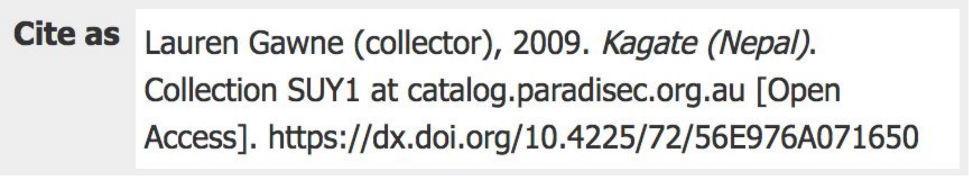
Example: The Australian Journal of Linguistics guidelines include the Generic Style Rules for Linguistics, the Leipzig Glossing Rules, the Austin Principles & the Tromsø Recommendations:

“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.



Cite as Lauren Gawne (collector), 2009. *Kagate (Nepal).* Collection SUY1 at catalog.paradisec.org.au [Open Access]. 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.

BY NC SA

References

- Andreassen, H. N. (2020). The Tromsø Recommendations for Citation of Research Data in Linguistics: Collaboration illustrated. *Ravnetrykk* 39: 114-122.
- Andreassen, H. N., Berez-Kroeker, A. L., Collister, L., Conzett, P., Cox, C., Smedt, K.D., McDonnell, B. and and the Research Data Alliance Linguistic Data Interest Group. (2019). Tromsø recommendations for citation of research data in linguistics (Version 1). R*esearch Data Alliance*. DOI: 10.15497/RDA00040
- Berez-Kroeker, A.L., L. Gawne, S. Kung, B.F. Kelly, T. Heston, G. Holton, P. Pulsifer, D. Beaver, S. Chelliah, S. Dubinsky, R.P. Meier, N. Thieberger, K. Rice & A. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1): 1–18.
- Berez-Kroeker, A.L., H.N. Andreassen, L. Gawne, G. Holton, S. Smythe Kung, P. Pulsifer, L.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/
- 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
- Gawne, L., B.F. Kelly, A.L. Berez- Kroeker & T. Heston. 2015. Putting practice into words: Fieldwork methodology in grammatical descriptions. 4th International Conference on Language Documentation and Conservation. Hawaii: February 26-March 1.
- Gawne, L., B.F. Kelly, A.L. Berez- Kroeker & T. Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. Language Documentation & Conservation 11: 157-189.
- Leonard, W.Y. (2017. Producing language reclamation by decolonising 'language'. In W. Y. Leonard & H. De Korne (Eds), *Language Documentation and Description*, vol 14. London: EL Publishing. pp. 15-36.
- Pawley, A. 2014. Grammar writing from a dissertation advisor's perspective. In T. Nakayama & K. Rice *The Art and Practice of Grammar Writing*, 7-23. University of Hawaii Press.

Acknowledgments

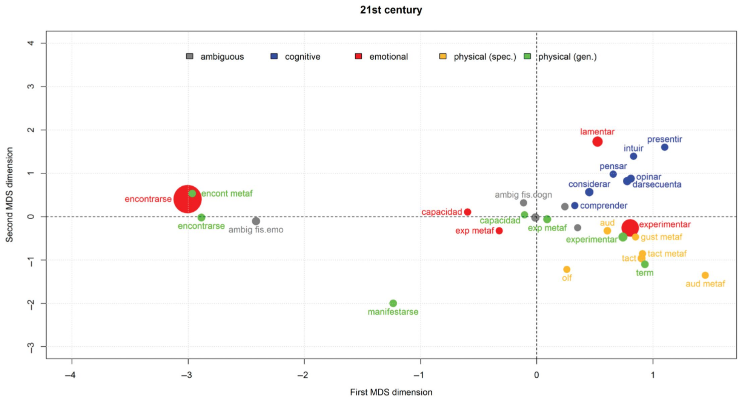
Many thanks to the participants in the Data Citation and Attribution project, the Data Science for All of Linguistics project, members of the RDA LDIG, and attendees at our previous workshops, courses and presentations for fruitful discussion.

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

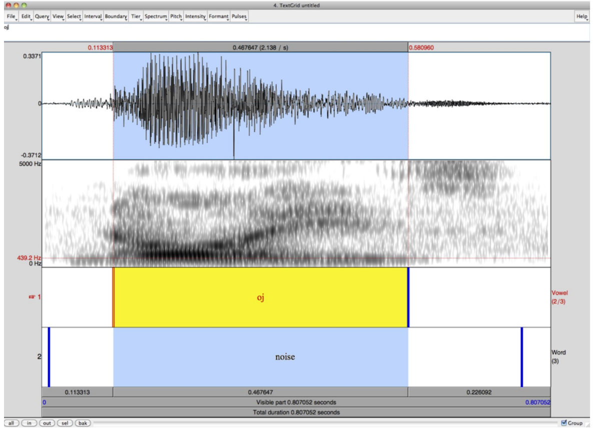
BY NC SA

Background:
What is linguistic data?



Jansegers & Gries
2017:10

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



Styler
2017:54

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

Payne 1997:253

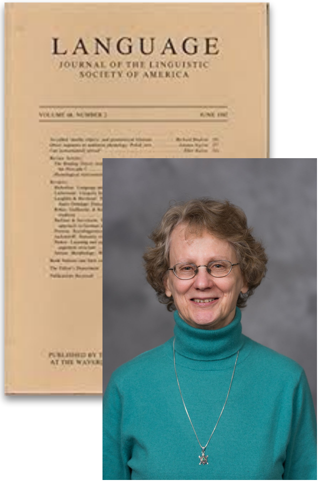
Background: A long-noticed problem

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.



BY NC SA

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)

Documentary and descriptive linguistics*

NIKOLAUS P. HIMMELMANN

Abstract

*Much of the work that is labeled “descriptive” within linguistics comprises
two activities, the collection of primary data and a (low-level) analysis of
these data. These are indeed two separate activities as shown by the fact
that the methods employed in each activity differ substantially. To date,
the field concerned with the first activity — called “documentary linguistics”
here — has received very little attention from linguists. It is proposed that
documentary linguistics be con
linguistic inquiry and practice
descriptive framework. A forma
to language descriptions) is pr
and theoretical issues connect
include the rights of the indiv
language documentation (sectio
the data to be included in a docu
of the quality of such data (sec*

1. Distinguishing description a

This article presents some ref
linguistics. My concern is the a
little-known or previously unre
are endangered, and the preser
by the recent surge of interest
Robins and Uhlenbeck 1991;
for descriptive work on these
The task of recording a little
the first being the collection,
data and the second a low-lev

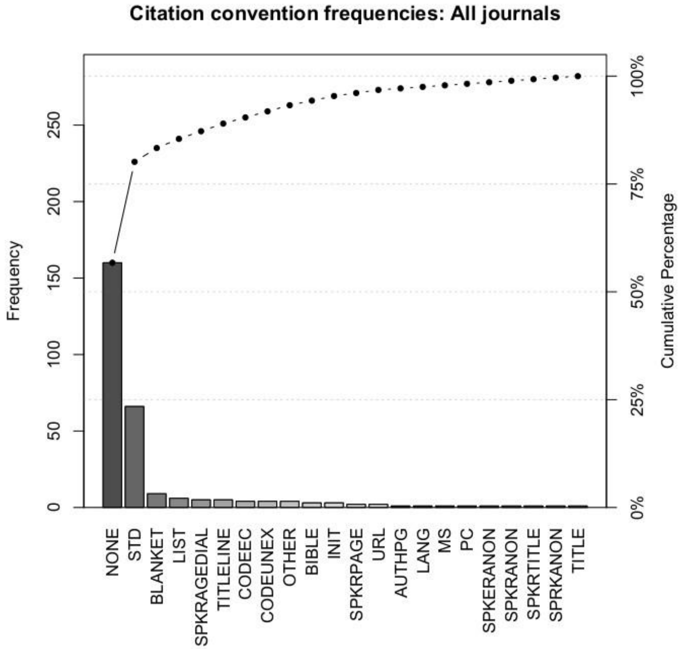
Linguistics 36 (1998), 161–195

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.

DE GRUYTER MOUTON Linguistics 2018; 56(1): 1–18

Andrea L. Berez-Kroeker*, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice and Anthony C. Woodbury

Reproducible research in linguistics: A position statement on data citation and attribution in our field

https://doi.org/10.1515/ling-2017-0032

Abstract: This paper is a position statement on reproducible research in linguistics, including data citation and attribution, that represents the collective views of some 41 colleagues. Reproducibility can play a key role in increasing

***Corresponding author: Andrea L. Berez-Kroeker,** Department of Linguistics, University of Hawai'i at Mānoa, 1890 East West Road, Moore 569, Honolulu, HI 96822, USA, E-mail: andrea.berez@hawaii.edu
Lauren Gawne, Department of Languages and Linguistics, SOAS University of London, London WC1H 0XG, UK; La Trobe University, Melbourne, VIC 3086, Australia, E-mail: l.gawne@latrobe.edu.au
Susan Smythe Kung, Archive of the Indigenous Languages of Latin America, University of Texas at Austin, Austin, TX 78712, USA, E-mail: skung@austin.utexas.edu
Barbara F. Kelly, Department of Languages and Linguistics, The University of Melbourne, Parkville, VIC 3010, Australia, E-mail: b.kelly@unimelb.edu.au
Tyler Heston, Payap University, Chiang Mai 50000, Thailand, E-mail: tylerheston@earthlink.net
Gary Holton, Department of Linguistics, University of Hawai'i at Mānoa, 1890 East West Road, Moore 569, Honolulu, HI 96822, USA, E-mail: holton@hawaii.edu
Peter Pulsifer, National Snow and Ice Data Center, Boulder, CO 80303, USA, E-mail: pulsifer@nsidc.org
David I. Beaver, Department of Linguistics, University of Texas at Austin, Austin, TX 78712, USA, E-mail: dib@utexas.edu
Shobhana Chelliah, Department of Linguistics, University of North Texas, Denton, TX 76203,

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.

BY NC SA

Background References

- Andreassen, H. N., Berez-Kroeker, A. L., Collister, L., Conzett, P., Cox, C., Smedt, K.D., McDonnell, B. and and the Research Data Alliance Linguistic Data Interest Group. (2019). Tromsø recommendations for citation of research data in linguistics (Version 1). R*esearch Data Alliance*. DOI: 10.15497/RDA00040
- Berez-Kroeker, A.L., H.N. Andreassen, L. Gawne, G. Holton, S. Smythe Kung, P. Pulsifer, L.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/
- Berez-Kroeker, A.L., L. Gawne, B.F. Kelly & T. Heston. 2017. *A survey of current reproducibility practices in linguistics journals, 2003-2012.* link.
- Berez-Kroeker, A.L., L. Gawne, S. Kung, B.F. Kelly, T. Heston, G. Holton, P. Pulsifer, D. Beaver, S. Chelliah, S. Dubinsky, R.P. Meier, N. Thieberger, K. Rice & A. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1): 1–18.
- Berez-Kroeker, A.L., B. McDonnell, E. Koller & L.B. Collister. In prep. *The Open Handbook of Linguistic Data Management*. MIT Press Open.
- Gawne, L. & A.L. Berez-Kroeker. 2018. Reflections on reproducible research. In B. McDonnell, A.L. Berez-Kroeker & G. Holton. (Eds.) *Reflections on Language Documentation 20 Years after Himmelmann 1998*, pp. 22–32. University of Hawaiʻi Press.
- Himmelmann, N. P. (1998). Documentary and descriptive linguistics. Linguistics, 36, 161-196.
- Jansegers, M. & S.Th. Gries. 2017. Towards a dynamic behavioral profile: A diachronic study of polysemous *sentir* in Spanish. *Corpus Linguistics and Linguistic Theory* 1-43.
- Payne, T. E. 1997. *Describing Morphosyntax: A Guide for Field Linguists*. Cambridge University Press.
- Styler, W. 2017. *Using Praat for Linguistic Research*, version 1.8.1. Online: http://savethevowels.org/praat. Accessed 20 November 2019.
- Thomason, S. 1994. The editor’s department. *Language* 70: 409-413.

BY NC SA