

Data Citation in Linguistics: Looking forward to new standards

Lauren Gawne, Andrea L. Berez-Kroeker, Helene N. Andreassen
La Trobe University, University of Hawai'i Manoa, UiT The Arctic University of Norway

The 20th International Congress of Linguists
Cape Town International Convention Centre
July 2-6 2018

These slides available: bit.ly/lingdata-icl20

What is linguistic data?

- **All levels of language:**

- from recordings of individual words to hours of audio-video
- from a single sentence to hours of conversation
- from an individual speaker to a whole community
- from any of the world's 7000 languages

- **All types of data:**

Audio recordings (stories, conversations, interviews, elicitation), video recordings, xml annotations, transcripts, 'glossed' text, dictionaries, experimental data (eye tracking data, reaction time data etc.), introspection, tagged corpora, spectrograms, sonograms, GPS data...

- **All locations:**

Archived with institutional repository, DELAMAN, on personal hard drives, in shoeboxes under the bed...

Replication in Science

Good scientific research is **replicable**

Recreate a controlled study > New data > [Dis]confirm previous results

Some studies can't be truly replicated (e.g. behavioral research)

Aim for **reproducible** research instead

Reuse of another's data > same or different conclusions

Reproducibility in Linguistics

Linguistics increasingly values reproducibility

“When I began my term as editor [...] I did not expect that these cases would occur frequently – so frequently, in fact, that the assumption that the data in accepted papers is reliable began to look questionable” (Thomason 1994: 409)

“[Language] documentation [...] will ensure that the collection and presentation of primary data receive the theoretical and practical attention they deserve.” (Himmelman 1998:164)

“Linguistic data are the very building blocks of our field [...] our field needs to accept responsibility for the proper documentation, preservation, attribution, and citation of these assets.” (Berez-Kroeker et al. 2018)

Planning the future of linguistic data

Linguistics Data Interest Group (LDIG) of the Research Data Alliance

“...objective is to contribute to a positive culture of linguistic data management and transparency in ways that are in keeping with what is happening in the larger digital data management community.” ([LDIG charter](#))

Research Data Alliance:

“builds the social and technical bridges that enable open sharing of data.” ([website](#))

What is the state of the art?

How do linguists link research publications back to the underlying data?

- Where does our data come from?
- What kind of data are we using?
- Where is the data now?
- Are we citing our examples? If so, how?

Our study

We examined:

- 271 journal articles from 9 journals
Range of areal foci, linguistic subfields, theoretical persuasions
- Published 2003-2012
5 years after Himmelmann 1998
“[Language] documentation [...] will ensure that the collection and presentation of primary data receive the theoretical and practical attention they deserve.” (p. 164)

We also looked at 100 descriptive grammars, see Gawne, et al. (2017)

For a more detailed write up of this current study see Berez et al. (2017)

Journal	No. articles included	Abbreviation
International Journal of American Linguistics	33	IJAL
Journal of African Languages and Linguistics	29	JALL
Journal of Sociolinguistics	33	JS
Language	33	LANG
Linguistics of the Tibeto-Burman Area	18	LTBA
Natural Language and Linguistic Theory	32	NLLT
Oceanic Linguistics	33	OL
Studies in Language	30	SL
Studies in Second Language Acquisition	30	S2LA

Data Coding

1. Source of data
2. Data genre analyzed (linguistic genre)
3. Where data is now
4. Citation conventions used to reference data, if any

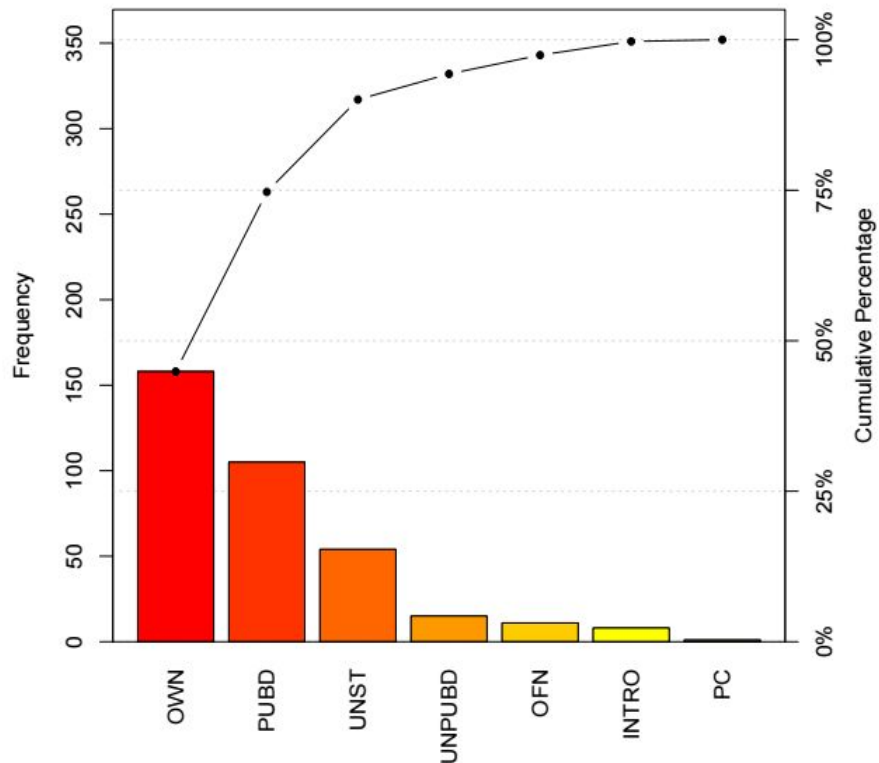
We also looked at whether the publication was presented with a clear research method

1. Source of data

- OWN: data collected by author
- PUBD: published data
- UNPUBD: unpublished data collected by someone other than the author (excluding fieldnotes)
- INTRO: introspection
- OFN: other person's fieldnotes
- UNST: source of data unstated
- NA: not applicable

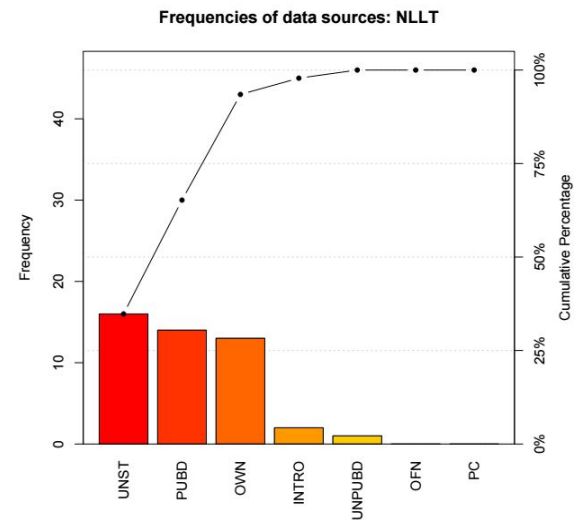
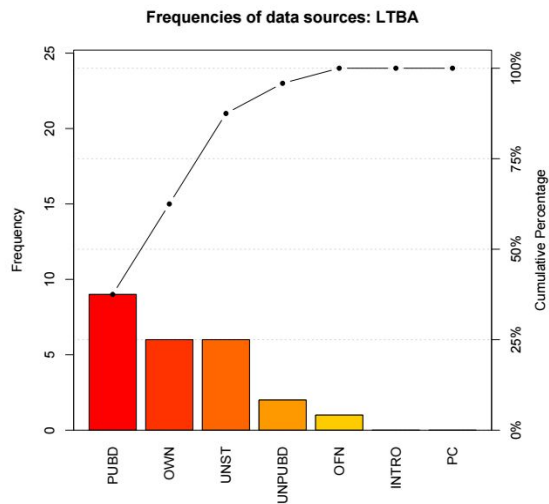
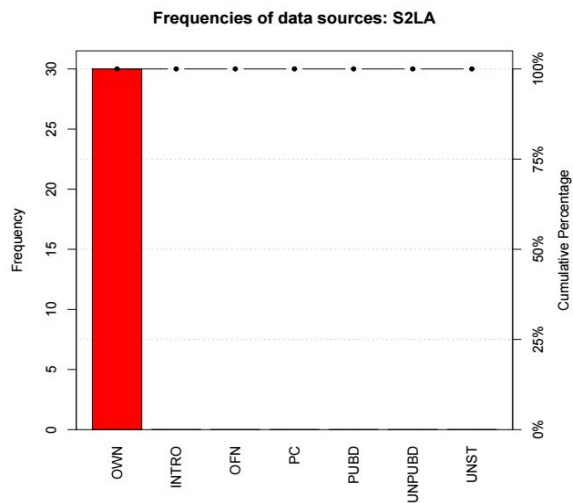
1. Source of data: all journals

Frequencies of data sources: All journals



- Most data come from authors' own research ~ 50%
- Followed by published data
- Followed by...unstated

1. Source of data: variation in data

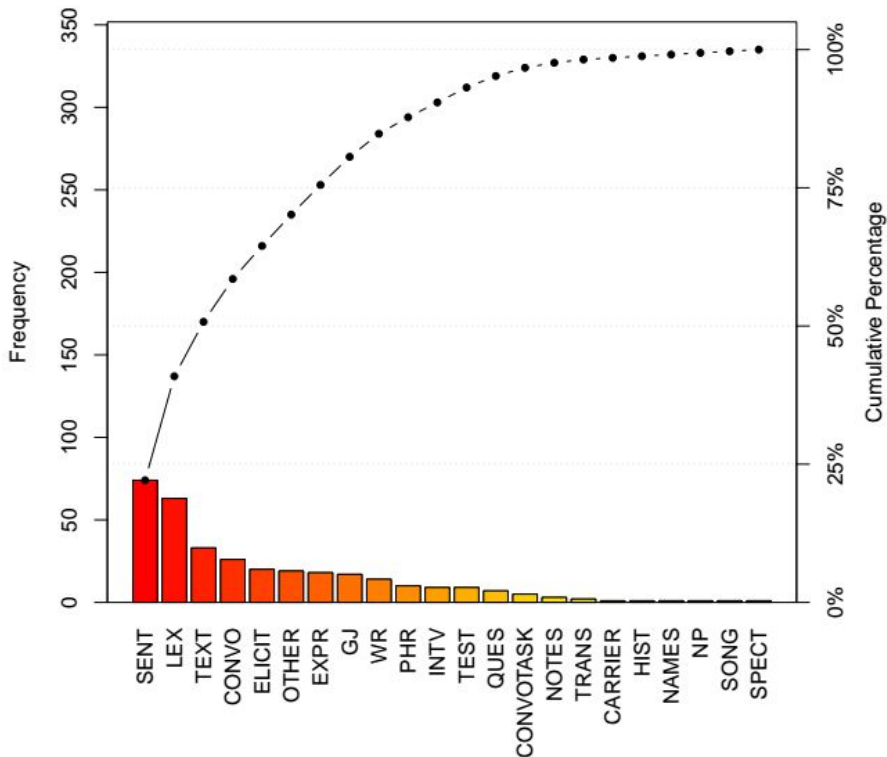


2. Data genre

- NOTES: own fieldnotes
- NP: noun phrases
- PHR: other phrases
- QUEST: questionnaires
- SENT: sentence data (broadly defined)
- SONG: songs
- SPECT: spectrograms
- TEXT: texts (broadly defined)
- TRANS: translation tasks (eg acquisition studies)
- TEST: tests in a school environment
- WR: written data (eg newspapers)
- CARRIER: data in a carrier sentence
- CONVO: conversational data (natural)
- CONVOTASK: conversational task (eg acquisition studies)
- ELICIT: elicitation
- EXPR: experimental
- GJ: grammaticality judgments
- HIST: historical data (eg correspondence sets)
- INTV: interviews
- LEX: lexical items/words
- NAMES: names
- OTHER: other

2. Data genre: all journals

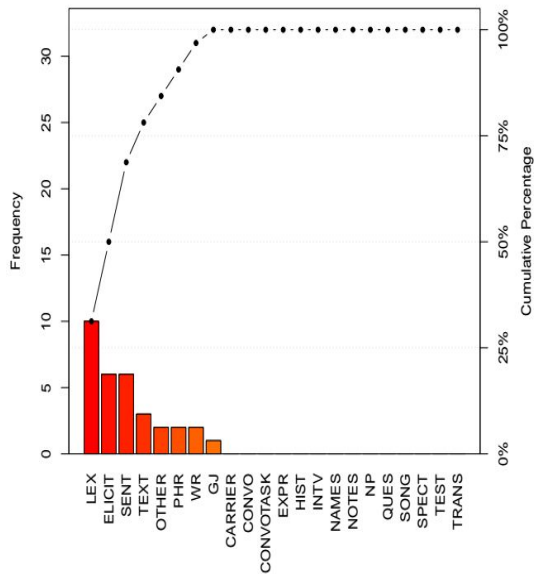
Data genre frequencies: All journals



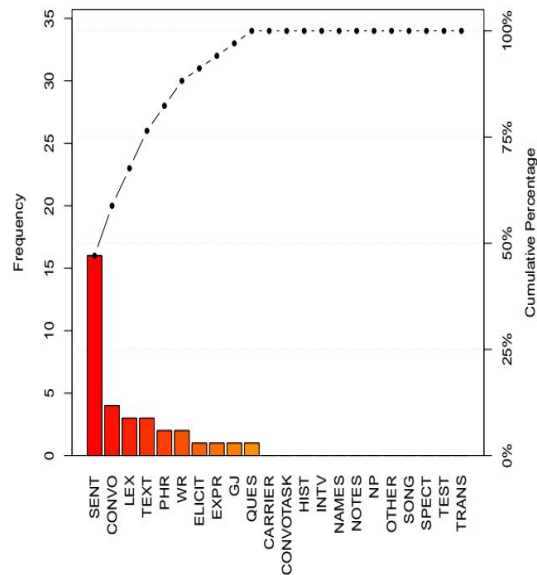
- Sentences
- Lexical items/words
- Texts

2. Data genre: individual journals

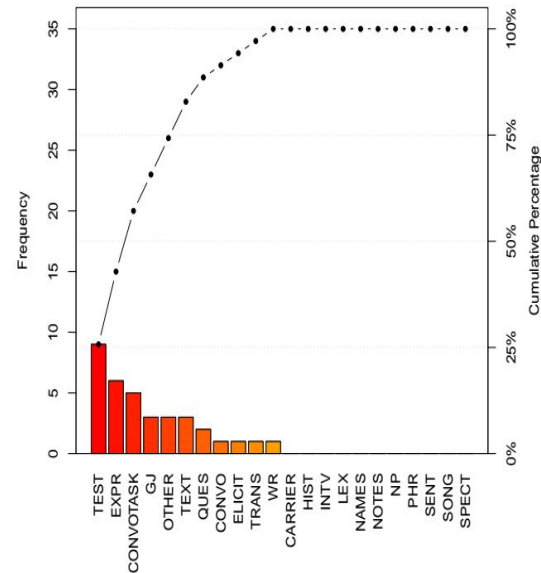
Data genre frequencies: JALL



Data genre frequencies: SL



Data genre frequencies: S2LA

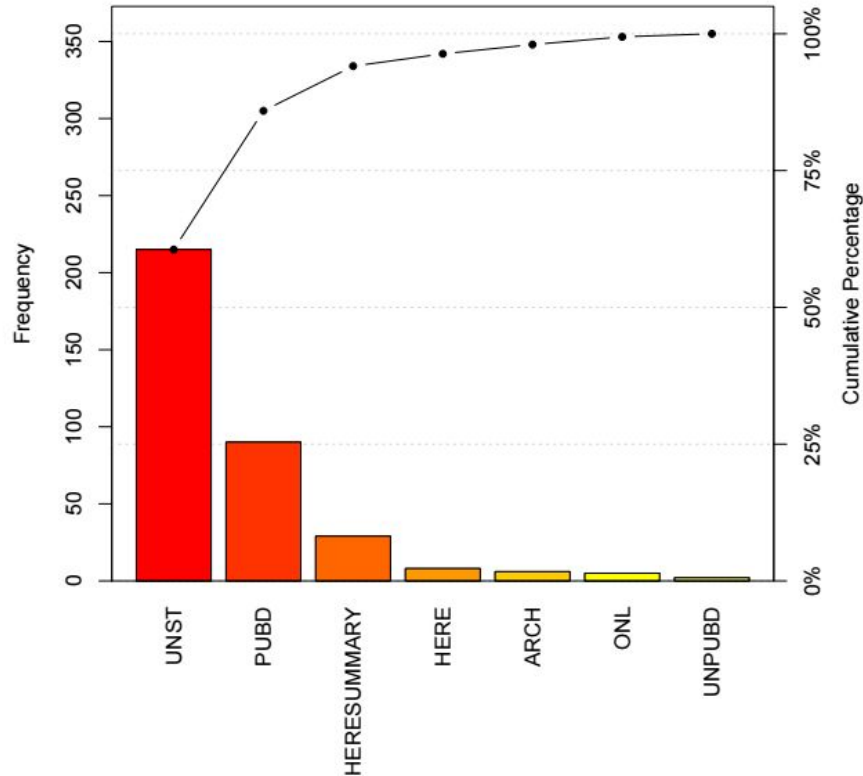


3. Where the data is now

- ARCH: archived in institutional repository
- PUBD: published
- HERE: article contains the primary data
- HERESUMMARY: data summarized in the article (stats, graphs, tables)
- ONL: online (website or other non-archive)
- UNST: location of data not stated

3. Where the data is now: all journals

Where the data can be found: All journals



- Mostly we don't know!
- "Published" a distant 2nd

4. Citation conventions used in examples

We found ~18 ways that people use some kind of formalised convention when referring to their data

You can see all of them, including examples at bit.ly/DataCitationSOTA

4. Citation conventions used in examples

SPKRAGEDIAL: citation appears as speaker's name + other demographic info

[T]here are times when I get stuck, and probably all my grammar is wrong, but I can - yeah, I can manage.

(Rita, f27)

(example from JS, Chand 2011:17)

4. Citation conventions used in examples

TITLE: citation appears as the title of the story or conversation it was taken from

[...]

83 *kyoo desho?*

today COP

'The day when they cook sukiyaki is tomorrow, and the day when they bring something [to us] is today, right?

(Broccoli)

(example from SL, Takara 2012: 95)

4. Citation conventions used in examples

CODEEC: citation is a code that is explained by author

So the buggies [bugíz] came out. **[BN T3P12]**

(endnote explains “[t]he code [BP T3P12] means speaker BN, tape 3, transcription page 12.”)

(example from JS, Brown 2003:21, note 9)

4. Citation conventions used in examples

MS: citation appears as standard reference to unpublished manuscript.

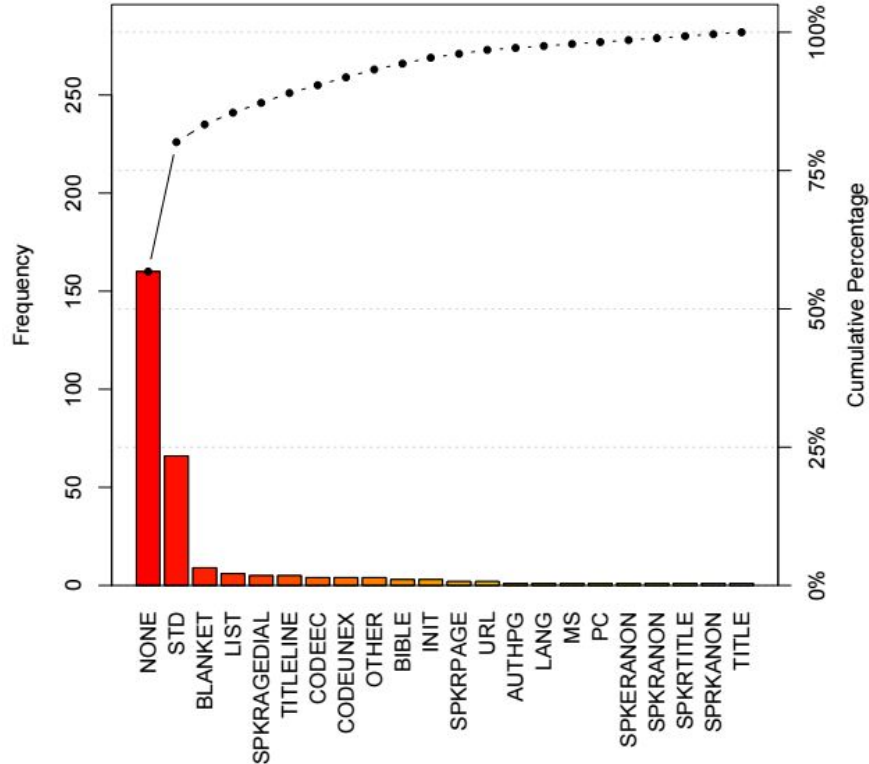
NONE: author did not include any form of citation

NA: article did not contain numbered examples

OTHER: other practice not easily classifiable here

4. Citation conventions used in examples: all

Citation convention frequencies: All journals



- No citation is most common
- “Standard” is a distant 2nd

What is the state of the art?

- Where does our data come from?
Mostly from authors' own research, and published
- What kind of data are we using?
Incredibly diverse range, but still heavy focus on sentence and word
- Where is the data now?
For over 50% of data we don't know, 25% is from existing publications
- Are we citing our examples? If so, how?
Overwhelmingly no. Published data cited using existing standards

LDIG Aims

- Development and adoption of common principles and guidelines for data citation and attribution
 - researchers, professional organizations, academic publishers, archives
- Education and outreach efforts
 - practical training and awareness of principles/sociological change
- 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](#))

Austin Principles of Data Citation in Linguistics

“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.”

“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.”

linguisticsdatacitation.org

Goal: Encourage and improve visibility and retrievability of research data
Based on FORCE 11 Joint Declaration of Data Citation Principles

Encouraging students

We can introduce students to good practice (cf. Pawley 2014)

Tromsø data management plan required for all research, including open publication of data if possible (all employees, including PhD students)

At University of Hawaii major change to PhD Handbook of Requirements (since Fall 2013):

- Students whose theses are based on fieldwork are required to properly archive their data
- Archiving plans part of the Dissertation Proposal. Only accepted DELAMAN archives may be used.
- Students required to submit proof of deposit to the committee before the dissertation can be approved.
- Descriptive theses must cite resolvable resources.

Encouraging colleagues

- [Endorse](#) the Austin Principles
- Encourage your organisations and publishers to endorse the Austin Principles
- Use peer review process to encourage colleagues to give more information about their research
- Build expectations about data transparency into research planning and funding
- Get involved in the LDIG

Next steps

- Guidelines for formatting data citations (for creators and users of data, humans and machines)
- Inclusion of guidelines in style guides
- Continued education about the need to consider the centrality of data to linguistic analysis

References

- Blevins, Juliette. 2005. Origins of northern Costanoan jak:en 'six': A reconsideration of senary counting in Utian. *IJAL* 71(1): 87-101.
- Berez-Kroeker, A.L., L. Gawne, B.F. Kelly & T. 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, A.L., H.N. Andreassen, L. Gawne, G. Holton, S.S. Kung, P. Pulsifer, L.B. Collister, The Data Citation and Attribution in Linguistics Group & the Linguistics Data Interest Group. 2017. The Austin Principles of Data Citation in Linguistics (Version 0.1).
- Berez-Kroeker, A.L., L. Gawne, S. Kung, B.F. Kelly, T. Heston, G. Holton, P. Pulsifer, D. Beaver, S. Chelliah, S. Dubinsky, R. 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): 1-17
- Brown, Becky. 2003. Code convergent borrowing in Louisiana French. *J Socio* 7(1): 3-23.
- Chand, Vineeta. 2011. Elite positionings toward Hindi: Language policies, political stances and language competence in India. *J Socio* 15(1): 6-35.
- Crowley, Terry. 2007. *Field linguistics: a beginner's guide*. Edited by Nicholas Thieberger, Oxford linguistics. Oxford: Oxford University Press.
- 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.
- Gezelter, D. 2009. Being scientific: Falsifiability, verifiability, empirical tests, and reproducibility. The OpenScience project. Online: <http://www.openscience.org/blog/?p=312>.
- Guérin, V. M. (2008). *Discovering Mavea: Grammar, texts, and lexicon*. PhD dissertation. Honolulu: University of Hawaii.
- Haspelmath, Martin, and Susanne Maria Michaelis. 2014. Annotated corpora of small languages as refereed publications: a vision. *Diversity linguistics comment*. Online: <http://dlc.hypotheses.org/691>.
- Himmelman, Nikolaus P. 1998. "Documentary and descriptive linguistics." *Linguistics* no. 36:161–195.
- Pawley, Andrew. 2014. "Grammar writing from a dissertation advisor's perspective." In *The Art and Practice of Grammar Writing*, edited by Toshihide Nakayama and Keren Rice, 7-23. Honolulu: University of Hawaii Press.
- Takara, Nobutaka. 2012. The weight of head nouns in noun-modifying constructions in conversational Japanese. *SiL* 36(1): 33-72.
- Thieberger, Nicholas. 2009. "Steps toward a grammar embedded in data" In Patricia Epps and Alexandre Arkipov (eds.) *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*, 389-408. Berlin; New York, NY: Mouton de Gruyter Mouton.
- Thomason, Sarah. 1994. The editor's department. *Language* 70. 409-423

Acknowledgements

We would like to thank the Research Data Alliance for their support of the Linguistics Data Interest Group. This work has also been supported by the National Science Foundation project Developing Standards for Data Citation and Attribution for Reproducible Research in Linguistics (NSF 1447886, PIs Berez-Kroeker, Holton, Kung, Pulsifer).

Lauren Gawne would like to thank La Trobe University. Funding for travel from the David Myers Research Fellowship, and the Humanities and Social Sciences Internal Research Grant Scheme.

Helene N. Andreassen would like to thank UiT The Arctic University of Norway. Andrea L. Berez-Kroeker would like to thank the University of Hawai'i.

These slides are available at
bit.ly/lingdata-icl20

linguisticsdatacitation.org

l.gawne@latrobe.edu.au
lingdata@hawaii.edu

