Data Citation in Linguistic Typology

Working Towards a Data Citation Standard in Linguistics

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

Abstract

Linguistic typology relies on high quality documentation and description work. As a field, linguistics has not created clear benchmarks for evaluating the creation, curation, and sharing of data sets. We present a survey of data citation in 5 years of *Linguistic Typology*, and argue that a data citation policy is needed.

Rationale

Typologists rely on the descriptive work of others, often with very little opportunity to evaluate for themselves the claims being made. Where citations are provided, the connections to the data sets are usually only vaguely identified.

The disconnect between linguistics publications and their supporting data results in much linguistic research being unreproducible, either in principle or in practice.

Previous surveys:

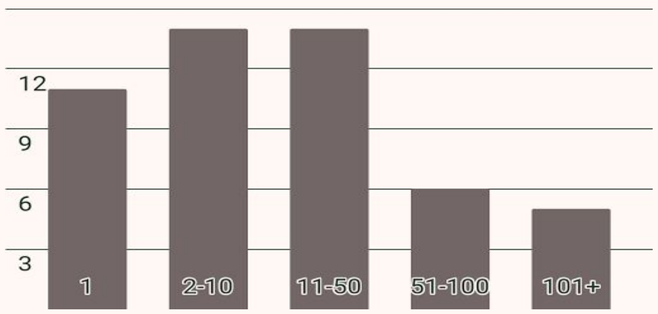
- **100 grammars** published 2003-2012 (Gawne et al., 2017). Vast majority did not provide citations to underlying data (which has serious implications for typology).
- **270 articles** from 10 journals (2003-2012) (Berez-Kroeker et al. 2017). Different subfields have different strengths in methods descriptions and data citation.

5 Year Survey of Data Citation in *Ling Typ*

- 5 years of research articles in *Linguistic Typology* (2012-2017 vol. 16.3-21.2)
- Total of 50 articles
- Discussion articles omitted (e.g. vol. 20.3, 2016)
- Based on methods from previous surveys (above)

Number of languages per article

Linguistic Typology attracts both 'big typology' (drawing extensively on primary literature) and articles focusing on individual languages.



Source of Data

Researchers draw on data from a variety of sources; multiple sources were counted for some papers.

PUBD published	45
OWN author's own data	17
CORP public corpus	5
OTHER other source of data	3
UNPUBD unpublished data from someone other than the author	2
UNK unknown source	1

Location of Data

Stating data location increases opportunity for reproducibility.

PUBD in another publication	43
HERE the article contains the data, and is its own main source	9
UNK unknown	7
ONL website or other non-archive internet storage	5
ARCH archived in an institutional repository, digital or physical	1

For papers based on authors own data we coded for whether the author mentions that data in the article have been deposited in a repository with an institutional commitment to long-term preservation, cataloging, and access. **Only 1 of the 17 papers mentions archiving.**

Data Citation Conventions

Data citation directs the reader back to the specific source of the data. Sources could be datasets (publically accessible or privately held), published texts (e.g. Bible translations), or other academic publications.

NONE no citation convention	21
URL a weblink to the location of the data online	1
STD Standard citation to published source	42
EXPL an explained citation code that links back to corpus	3
UNEXP unexplained code, or unclear how it links back to corpus	3
NAME name of speaker or text	1

Most Common Data Types

Compared to other journals, *Linguistic Typology* draws very strongly on sentence-level data.

Sentence	34
Lexical	23
Morphological	11
Phonetic	2
Phrasal	1

Discussion

This survey demonstrates that we need a more robust culture of providing accountability in research. Valuing reproducibility means facilitating access to the data and methods ensuring that other researchers may also reach the same conclusions.

Benefits of conventionalised research data citation:

- Enhancing the accessibility and transparency of research in general (Gezeltzer 2009; Boulton 2014)
- Raising the professional valuation of descriptive work (Haspelmath & Michaelis 2014; Margetts et al. 2016, Berez-Kroeker et al. 2018)

Challenges ahead:

- Data citation methods that reflect the granularity of citation and formatting (c.f. Ball & Duke 2015)
- Providing training and support in data management and citation for researchers

The Austin Principles of Data Citation in Linguistics

The "Austin Principles" interprets the FORCE11 Joint Declaration of Data Citation Principles to address linguistic data specifically. These guiding principles have been created to enable YOU to make decisions about your data to ensure it is as accessible and transparent as possible.

www.linguisticsdatacitation.org

Help shape the future of data citation

Join the RDA Linguistics Data Interest Group (LDIG)

Aims of the Research Data Alliance (RDA) LDIG:

1. Development and adoption of common principles and guidelines for data citation and attribution
2. Education and outreach to on the principles of reproducible research and the value of data citation and management
3. Efforts to ensure greater attribution of linguistic data set preparation within the linguistics profession.

QR link to the Austin Principles:

Contact us: l.gawne@latrobe.edu.au (Gawne), andrea.berez@hawaii.edu (Berez-Kroeker), helene.n.andreassen@uit.no (Andreassen),

References
Ball, A. & Duke, M. 2015. 'How to Cite Datasets and Link to Publications'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides
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., 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).
Boulton, G. 2014. Open data and the future of science. Paper presented at the *9th Munin Conference on Scholarly Publishing*, Tromsø, 26-27 Nov. 2014. http://septentrio.uit.no/index.php/SCS/issue/view/265
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, & reproducibility. *The OpenScience project*. Online: http://www.openscience.org/blog/?p=312.
Haspelmath, Martin & Michaelis, Susanne Maria. 2014. Annotated corpora of small languages as refereed publications: A vision. Diversity linguistics comment. Online: http://dlc.hypotheses.org/691
Margetts, A., N. Thieberger, S. Morey & S. Musgrave. 2016. Assessing annotated corpora as research output. Australian Journal of Linguistics 36(1). 1-21.

12th Conference of the Association for Linguistic Typology (ALT), Canberra, Australia. 12-14 December 2017.