Data Transparency and Citation in Gesture

Lauren Gawne,[1] Chelsea Krajcik,[2] Helene N. Andreassen,[3] Andrea L. Berez-Kroeker,[4] Barbara F. Kelly[5]

1 La Trobe University, 2 SOAS, University of London, 3 UiT The Arctic University of Norway, 4 University of Hawai'i at Mānoa, 5 University of Melbourne

Background

Gesture Studies has a strong history of research that spans multiple fields, but there is still not a robust culture of valuing reproducibility.

Reproducibility provides benefits including:

- Accountability in research by facilitating access to the underlying data and methods ensuring that other researchers may also reach the same conclusions (Gezelter 2014).
- Raised professional valuation of developing corpora that can be reused (Haspelmath & Michaelis 2014; Margetts et al. 2016; Berez-Kroeker et al. 2018).

Skubisz's (2017) survey of data coding and terminological definitions in GESTURE demonstrated that these key features of research are often underspecified in articles published in the journal to date.

Data Transparency in linguistics

- **Language Description** 100 grammars (2003-2012) (Gawne et al. 2017a). Vast majority did not provide citations to underlying data.
- **Linguistic Typology** 50 articles from 5 years of *Linguistic Typology* (Gawne et al. 2017b). Low frequency of authors citing own data.
- **10 leading linguistics journals** 270 articles (2003-2012) (Berez-Kroeker et al. 2017). Different subfields have different strengths in methods descriptions and data citation.

Survey of Data Citation in *GESTURE*

- 5 years of research articles in GESTURE (2012-2017 vol. 12.1-16.1)
- Total of 56 articles
- Discussion and introductory articles omitted
- Based on methods from previous surveys (above)

We seek to understand how transparent each article is in regard to:

- Citation of data to a source that would allow the reader to analyse the data for themselves
- The type of data, and what languages are the target of the anslysis

Source of Data

Researchers draw on data from a variety of sources, but mostly collect their own data. (n=60, multiple sources were counted for some papers)

PUBD published	15
OWN author's own data	42
UNK unknown source	2
ARCH archived	1

Location of Data

Stating data location increases opportunity for reproducibility. Many articles represent the only location of the data, or a summary. (n=57)

UNK unknown	23
HEREsummary a summary of the data is given in the paper	21
PUBD in another publication (the author's or someone else's)	8
HERE the article contains the data, and is its own main source	3
ONL website or other non-archive internet storage	2

Data Citation Conventions

Data citation directs the reader back to the specific source of the data. Sources could be datasets (publicly accessible or private), published texts (e.g. Bible translations), or other academic publications. (n=53)

NONE no citation convention	32
NAME name of speaker or text	11
STD Standard citation to published source	5
NUM numbered in order of original recordings or discussion	3
EXPL an explained citation code that links back to materials	1
URL a weblink to the location of the data online	1

Most Common Data Types

Perhaps unsurprising, given the diversity of work in the field, there is diversity in the types of data surveyed. (n=56)

EXPER experimental data	19
CONVO conversation data	13
TASK task-based data	10
OTHER other data types	8
MULTI multiple data types (e.g. task & convo)	4
ELICIT elicited data	1
NARR narrative data	1

Discussion

This survey demonstrates that we need a more robust culture of data accountability in gesture research. Researchers are mostly drawing on their own data, but are not stating the location of their data, and are not providing citation of individual examples.

GESTURE has recently adopted the standards of the Center for Open Science, which requires thorough description of methods and analyses, plus presentations of data in online data repositories.

As a research community we need to foster a culture of valuing research reproducibility.

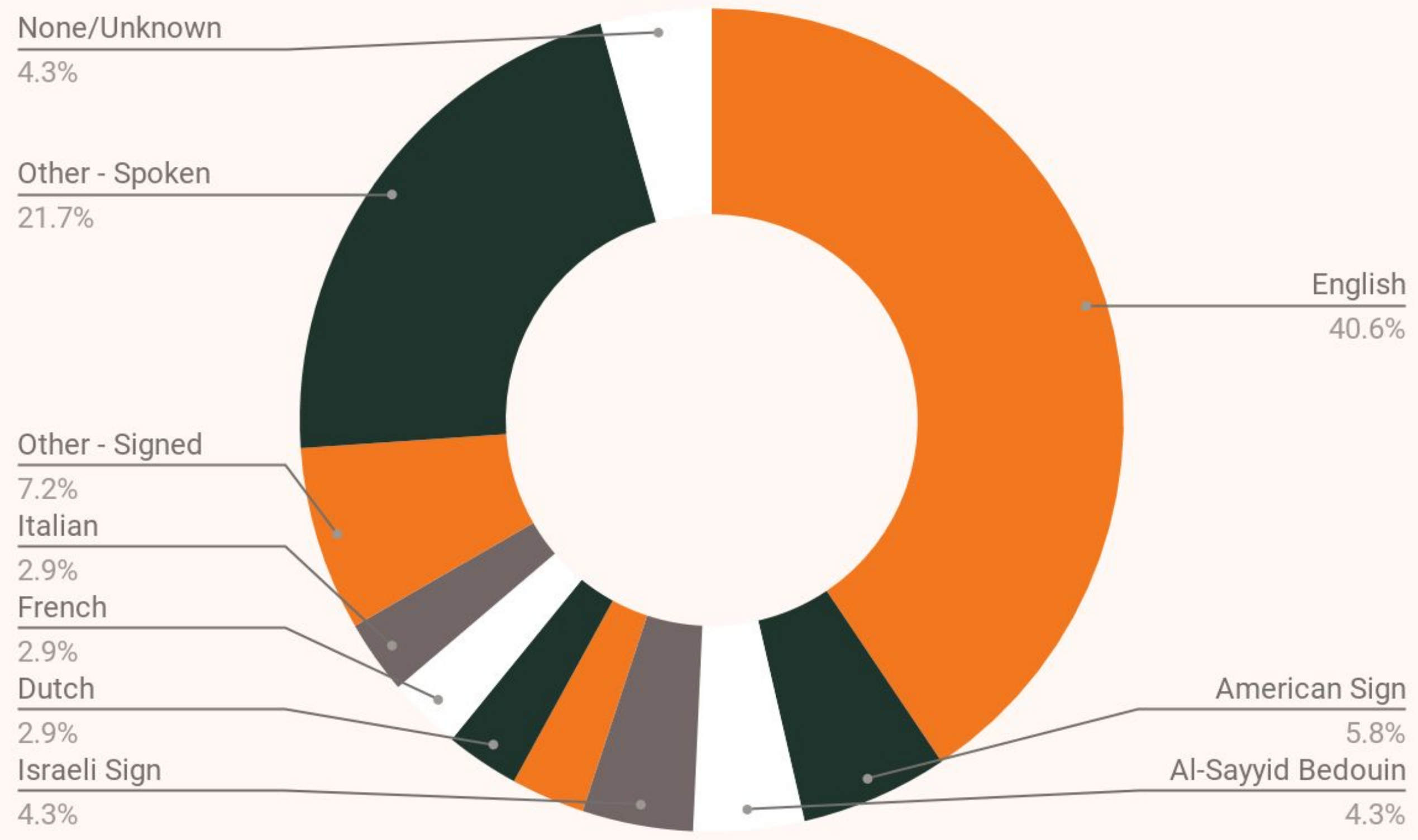
Challenges ahead:

- Managing video data that includes identifying footage of individuals (Green et al. 2013)
- Data citation methods that reflect the granularity of citation and formatting (c.f. Ball & Duke 2015)
- Training and support in data management for researchers

Languages included

We additionally coded the papers for the languages they include.

- 66 language mentions over 53/56 paper
- 28 distinct languages
- Gesture research is still overly focused on English



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)

Contact: l.gawne@latrobe.edu.au (Gawne) **Austin Principles QR link:**

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. 2017a. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11: 157-189.

Gawne, L., A.L. Berez-Kroeker & H.N. Andreassen. 2017b. Data Citation in Linguistic Typology: Working Towards a Data Citation Standard in Linguistics (poster). Association for Linguistic Typology 12. Canberra: December 11-15.

Gezelter, D. 2009. Being scientific: Falsifiability, verifiability, empirical tests, & reproducibility. *The Open Science 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

Green, J., G. Woods, & B. Foley. 2011. Looking at language: Appropriate design for sign language resources in remote Australian Indigenous communities. In N. Thieberger, L. Barwick, R. Billington & J. Vaughan. S*ustainable data from digital research: Humanities perspectives on digital scholarship*, 66-89.

Haviland, J.B. 2013. Introduction: Where does "Where do nouns come from?" come from? *Gesture* 13(3), 245-252.

Margetts, A., N. Thieberger, S. Morey & S. Musgrave. 2016. Assessing annotated corpora as research output. Australian Journal of Linguistics 36(1). 1-21.

8th Conference of the International Society for Gesture Studies, Cape Town, July 4-8, 2018