A Survey of Current Reproducibility Practices in Linguistics Publications

Lauren Gawne,* Andrea Berez-Kroeker,** Barbara Kelly,*** Tyler Heston**

*School of Oriental & African Studies, **University of Hawai'i at Manoa, ***University of Melbourne

Abstract

This project considers the role of reproducibility in increasing verification and accountability in linguistic research. An analysis of over 370 journal articles, dissertations, and grammars from a ten-year span is taken as a sample of current practices in the field. These are critiqued on the basis of transparency of data source, data collection methods, analysis, and storage. While we find examples of transparent reporting, much of the surveyed research does not include key metadata, methodological information, or citations that are resolvable to the data on which the analyses are based. This has implications for reproducibility and hence accountability, hallmarks of social science research which are currently under-represented in linguistic research.

Rationale

Reproducible research aims to provide scientific accountability by facilitating access for other researchers to the data upon which research conclusions are based. The benefit of reproducibility is evident in cases where faithfully recreating the research conditions is impossible.

Because linguistics is a social science dealing with observations of complex behavior, it is a field that would seem to lend itself to the kind of scientific rigor that reproducibility provides; however, until now there has been little discipline-wide discussion of how we might implement reproducibility, nor has there even been a widespread identification of a need to do so. The factors contributing to the selection of one inflected form over another in spontaneous conversation by a speaker of language X are difficult to control for or even observe. Even in a prepared elicitation session or a grammaticality judgment task—a controlled setting for linguistic observation much like that of the chimpanzee tool-selection study—researchers cannot conceivably control for every possible variable leading to an utterance or judgment, including previous experience.

In Bird and Simons's seminal 2003 article on portability for linguistic data in the digital age, the authors present at least four domains of data management that directly support reproducible research as it is understood here: *citation, discovery, access,* and *preservation*. Of particular interest is *citation*. Bird and Simons (2003:572) value a robust citation practice: '[w]e value the ability of users of a resource to give credit to its creators, as well as learn the provenance of the sources on which it is based'. Moreover, proper citations should be resolvable to digital data in a manner that is persistent regardless of location, citable to a particular version of that data, and appropriately granular. This of course presumes that the data themselves are also properly preserved, discoverable, and accessible.

While the Bird & Simons 2003 position paper has been instrumental in defining the field's values toward digital data – its stated aim is to build consensus around broad principles of best practices – it stops short of providing actual guidelines for implementing those practices. Instead, the authors ask the linguistics community to engage in discussion, '[to] lead to deeper understanding of the problems with current practice ... and to greater clarity about the community's values' (Bird & Simons 2003:558). We agree.

What, then, would be needed to make linguistics a more reproducible scientific endeavor? We maintain that prioritizing transparency in two primary realms would be required:

1. Transparency about source data: authors making linguistic claims should aim to be transparent about how the data upon which those claims are based could be accessed by a reader wishing to do so:
   a. Transparency about what data have been used (published data, introspective data, corpus data, elicited data, testing data, etc.),
   b. Transparency about where data can be found (in publications, in archives, in field notes in a personal collection, online, etc.),
   c. Transparency about how to locate relevant subparts of a data set (page numbers, corpus line numbers, offset time-codes of starting and ending points in an audio or video recording, etc.).
2. Transparency about methods of data collection and analysis: authors making linguistic claims should aim to be transparent about the methods used to obtain and analyze the data upon which those claims are based:
   a. A description of the conditions under which data were collected (where were data collected and for how long, what genre of speech data was collected, etc.),
   b. Access to information about the apparatus used for data collection and analysis (hardware used for collecting data, software used to analyze data, analytical or theoretical frameworks, questionnaires and other stimulus tools, whether data were elicited, etc.),
   c. Information about who participated in the providing of data (demographic information, language community information, etc.) Although we strongly advocate for granular citation to recoverable source data, we also acknowledge that this must be done with sensitivity toward the communities whose language is being recorded (see Chelliah & de Reuse 2011:147-151).

Our study

Based on our previous study of transparency in descriptive grammars (Gawne et al., Subm.), we surveyed current practices in linguistics journal publication with regard to the two parameters for transparency described above.

We surveyed nine linguistics journals, aiming for broad coverage in a number of dimensions. These included four journals with areal foci, two targeted subfields divergent theoretical persuasions, and the top journal in the discipline. All items included in our survey span a ten-year period starting in 2003. Our data set includes 271 articles.

The table to the right shows the number of articles per journal included in our study, and the abbreviations used henceforth to refer to those journals.

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

Coding

Articles were coded for a range of variables, selected after exploratory coding of a subset of the articles. Variables are based in part on theoretical discussions in the language documentation literature (eg., Gippert et al. 2006, Bowern 2008, Lüpke 2010, Chelliah & de Reuse 2011, Thieberger 2012, Austin 2013, and Nakayama & Rice 2014).

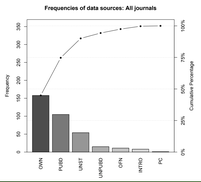
- *Description of data collection methods* (YES/NO): whether the author explicitly describes the methods used in data collection.
- *Information about participants in the study* (YES/NO): whether the author gives any demographic information about the people from whom data were collected.
- *Mention of data collection equipment* (YES/NO/NA): whether the author describes equipment (i.e., hardware)used in data collection.
- *Mention of data analysis tools or software* (YES/NO/NA). This variable tracks whether the author mentions the use of any tools that assist in analyzing data (e.g., Praat, R).
- *Mention of time spent collecting data* (YES/NO/NA).
- *Whether data have been archived* (YES/NO): whether the author mentions that data used in the article have been deposited in a repository with an institutional commitment to long-term preservation, cataloging, and access.
- *Source of data*. We coded for the source of the data used in each article; multiple sources were allowed. Sources include:
  INTRO: introspection
  NA: not applicable
  OFN: fieldnotes collected by someone other than the author
  OWN: data collected by the author
  PUBD: published
  UNST: not stated
  UNPUBD: unpublished data collected by someone other than the author
- *Where the data are now.* We coded for where the data are currently located, if stated by the author. Options included:
  ARCH: archived in an institutional repository, either digital or physical
  HERE: the article contains the data, and is its own main source
  HERESUMMARY: data are summarized in the article, using descriptive statistics, tables, graphs, or other presentation
  PUBD: published
  ONL: online (a website or other non-archive internet-based storage)
  UNST: not stated
- *Citation conventions used in numbered examples from the source data*. The use of numbered examples is a hallmark of linguistics writing, and these are usually drawn from collected or ad hoc data. We discovered a broad range of methods for citing numbered examples back to their sources. Sources could be data sets (both publically accessible and privately held), published texts such as Bible translations, or other academic publications. Note that articles with no numbered examples were coded as NA, and those with examples but no citation were coded as NONE. See paper for examples.

Results

The table below shows the results of all the binary (or binary + NA) variables:

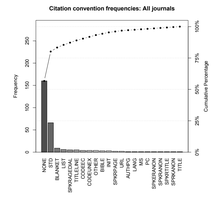
	Methods	Participants	Equipment	Tools	Time	Archived
IJAL	45.5	68.8	9.4	12.1	37.5	19.4
OL	15.2	39.4	0.0	6.1	45.5	3.0
JALL	21.4	58.6	7.1	10.7	50	0.0
LTBA	22.2	60.0	0.0	0.0	20.0	0.0
JS	54.5	93.1	34.5	33.3	77.8	0.0
S2LA	100.0	100.0	46.7	50.0	76.7	0.0
NLLT	15.6	32.1	6.9	6.9	6.9	0.0
SL	30.0	42.9	8.3	11.5	41.2	0.0
LANG	51.5	62.1	35.5	48.4	23.1	3.4

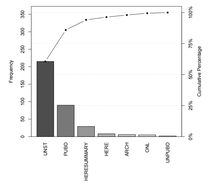
- *Methods:* S2LA authors are best by far at describing the research methods they employed. OL and NLLT authors described their methods the least (only 15% of the time).
- *Participants:* S2LA and JS authors described their participants well. NLLT authors described them poorly.
- *Equipment:* Not standard practice in all journals surveyed, but phonetics papers in every journal did this well.
- *Tools:* Authors in S2LA and LANG had the highest frequency of tool reporting (but still maximally only half).
- *Time:* Authors in experimental journals S2LA and JS best described time spent collecting data.
- *Archived:* Resoundingly poor performance across all journal authors.



Source of data. Not surprisingly, across all journals, authors show a strong preference for using data they collect themselves, followed by data coming from traditionally published sources.

Authors mentioned other data sources (unpublished data, introspective data) less frequently than simply not stating the source of data. See the paper for performance by authors in individual journals.





Data citation conventions. Overwhelmingly, journal authors *do not provide any citation for numbered examples whatsoever.*

Interestingly, this means that although most examples come from the author's own data, that data is rarely cited.

When examples are cited, the most common citation format is a "standard" (STD) format for published data of Author, Year: page, as in

(Everett and Kern 1967:162)

See the paper for performance by authors in individual journals.

Where the data are now. Over half the journal authors we surveyed do not explicitly state where the data upon which their study is based can be located. When a location is given, the location is a traditional publication.

See the paper for performance by authors in individual journals.

Discussion

This survey of our discipline reveals both good news and bad for the current state of reproducibility in linguistics.

First, the bad news: with regard to the metrics we examined here, linguistics publications as a whole have a lot of room for improvement if they are to provide the kind of transparency of data and methods that we advocate here.

We found that readers are implicitly asked to make assumptions about aspects of the research process: that data are collected in a felicitous manner, that data sets are locatable and verifiable, and that examples of linguistic phenomena are representative of the context(s) from which they are drawn.

Few among us advertise in our publications that we have taken responsibility for the longevity and accessibility of our data sets, which means that precious endangered language data can disappear, and expensive experiments may be recreated out of ignorance, rather than from a spirit of scientific reproducibility. In short, we are in danger of being a social science asking its audience to *take our word for it*.

But our study also reveals some very good news, which holds promise for linguistics becoming more transparent in the future. We found that in fact different subfields do have strengths in facets of research transparency, as represented by the publications we surveyed.

Practitioners in different subfields 'do transparency' differently well, and these practices could serve as models for an eventual amalgamated standard. For example, S2LA authors describe research methods exceptionally well—the strong experimental focus of the journal means that a methods section is a normalized expectation. Authors in S2LA, JS, and IJAL frequently provide information about research participants. Authors in S2LA and JS usually provide information about tools, hardware, and software, as do authors of phonetics papers across all journals.

Differences across subfields account for our findings: some journal authors omit the explication of some factors because they are generally *understood*, while others include them because of *tradition*. Claims about introspective data are generally understood to have been made by people with fluency, and historical-comparative data is understood to come from unpublished worldists and published dictionaries. Field linguists describe the speech community and their fieldwork conditions by tradition; phoneticians have a tradition of describing equipment.

And everyone includes standard citations of published material, which precisely illustrates our point: because there is a *disciplinary expectation* to cite published material correctly, and a *standard format* for doing so, all authors in all journals we surveyed do it consistently. Linguistics has no such expectations or recommendations for other factors we examined here.

Importantly, the field is not without disciplinary resolutions valuing aspects of research transparency. The Linguistic Society of America's *Resolution on Cyberinfrastructure* encourages linguists to 'make the full data sets available, subject to all relevant ethical and legal concerns'; the *Ethics Statement* urges linguists to 'carefully cite the original sources of ideas, descriptions, and data'. But what is lacking are discipline-wide guidelines for where to store data or how to cite it, as well as minimum standards for methodological accountability in publications. The *Unified Style Sheet for Linguistics* does not contain advice for citing or formatting references to data sets.

In other words, valuing transparency is one thing; implementing seems to be quite another.

If the field of linguistics decides to pursue the development and adoption of practical methods for increasing transparency, we need to make some collective decisions about our attitudes toward data collection and management. This discussion could be held within each subfield of linguistics, with subsequent larger conversations held across sub-disciplines to ensure that the bar we set for ourselves reflects the working realties of all research linguists. We hope that the ensuing standards would not be dismissed out of hand on the premise that they break with current practice.

Contact us: lg21@soas.ac.uk (Gawne), andrea.berez@hawaii.edu (Berez-Kroeker), b.kelly@unimelb.edu.au (Kelly), theston@hawaii.edu (Heston)

References

Austin, Peter K. 2013. Language documentation and meta-documentation. Keeping languages alive: Documentation, pedagogy and revitalization, ed. by Mari C. Jones and Sarah Ogilvie, 3–15. Cambridge: Cambridge University Press.
Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. Language 79.557–582.
Bowern, Claire. 2008. Linguistic fieldwork: A practical guide. Basingstoke, New York: Palgrave Macmillan.
Chelliah, Shobhana L., and Willem J. de Reuse. 2011. Handbook of d escriptive linguistic fieldwork. London: Springer.
Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker & Tyler Heston. Submitted. Putting practice into words: The state of data and methods transparency in grammatical descriptions.
Gippert, Jost; Nikolaus P. Himmelmann; and Ulrike Mosel. 2006. Essentials of language documentation. Berlin: Mouton de Gruyter.
Lüpke, Friederike. 2010. Research methods in language documentation. Language documentation and description, vol 7, ed. by Peter K. Austin, 55–104. London: SOAS.
Nakayama, Toshihide, and Keiren Rice (eds.) 2014. The art and practice of grammar writing. (Language Documentation & Conservation Special Publication No. 8). Honolulu: University of Hawai'i Press.

For the full list of references please email the authors.