2

Reproducible research in descriptive linguistics: integrating archiving and citation into the postgraduate curriculum at the University of Hawai'j at Mānoa

Andrea L. Berez

On valuing reproducibility in science and linguistics

The notion of *reproducible research* has received considerable attention in recent years from physical scientists, life scientists, social and behavioural scientists, and computational scientists. Some readers will be familiar with the criterion of *replicability* as a tenet of good execution of the scientific method, in which sound scientific experiments or studies are those that can be recreated elsewhere leading to new data, and in which sound scientific claims are those that are confirmed by the new data in a replicated study. For example, if a researcher conducts a scientific study by surveying 5000 people selected at random, that study and claims arising from it are replicable if another researcher can make the same claim based on new data that come from a random survey of 5000 different people. Likewise, claims can be disproven in a replication, if the new researcher finds different results arising from

Berez, **Andrea L. (2015). <Reproducible research in descriptive linguistics: integrating archiving and citation into the postgraduate curriculum at the University of** Hawai'I **at Mānoa.' In *Research, Records and Responsibility: Ten Years o!PARADISEC,* edited by** Amanda Harris, Nick Thieberger and Linda Barwick. Sydney: Sydney University Press.

new data. Nonetheless, because the original study was replicable, the research method itself is considered to be sound, even if the original results are later disproven.

Reproducibility is similar to replicability, but reproducible research aims to provide accountability by allowing other researchers to reach the same (or different) conclusions using the same data set as the original publication, rather than from new data arising from the same experimental conditions. The term *reproducible research* was developed mainly in computer science (e.g., Buckheit and Donoho 1995; de Leeuw 2001; Donoho 2010), with the intention that researchers should provide not only the academic paper, but also the data and computer code upon which the paper is based, thus allowing readers to reach the same conclusions about the same data sel. Summarised by Dan Gezelter of the Open Science Project:

> If a scientist makes a claim that a sceptic can only reproduce by spending three decades writing and debugging a complex computer program that exactly replicates the workings of a commercial code, the original claim is really only reproducible in principle ... Our view is that it is not healthy for scientific papers to be supported by computations that cannot be reproduced except by a few employees at a commercial software developer ... it may be *research* and it may be *important*, but unless enough details of the experimental methodology are made available so that it can be subjected to true reproducibility tests by sceptics, it isrit Science. (Gezelter 2009, emphasis original)

Reproducibility is potentially useful in other scientific enterprises beyond the physical sciences and computer science. In many fieldwork-based life and social sciences, precise replicability is impossible to achieve. The variables contributing to a particular instance of observation are too hard to control for - for instance, the mechanisms by which frog-eating bats find prey in the wild (Ryan 2011). Even in semi-controlled situations like studying primate tool-use in captivity (Tomasello and Call 2011) it is difficult to reproduce every environmental or non-environmental factor that may contribute to which tool, for example, a chimpanzee will select in a given situation.

Reproducible research in language documentation and description

Linguistics, which can also be considered a social science dealing with observations of complex behaviour, is another field that would seem to lend itself to the kind of scientific rigour that reproducibility provides, but until now there has been little discipline-wide discussion of how we might implement reproducibility, or even a widespread identification of a need to do so. The goal of reproducible research as discussed here is intended to increase accountability in the search for understanding the nature of language, rather than to reproach colleagues. The discussion has not been so benign in other fields, however: compare the recent controversy in social psychology, in which Diederik Staple was found to have fabricated data in 15 to 20 years' worth of publications (Crocker and Cooper 2012). Fang et al. (2013) surveyed more than 2000 biomedical and life sciences journals and found that while 21.3% of article retractions were due to honest investigator error, fully 67.4% of retractions were due to 'misconduct, including fraud or suspected fraud (43.4%), duplicate publication (14.2%) and plagiarism (9.8%)' (Fang et al. 2013, 1). This has led to discussions of solutions including a 'transparency index' (Marcus and Oransky 2012) and a 'retraction index' (Fang and Casadevall 2011) for journals, watchdog websites (e.g., http://retractionwatch.coml) and indices, and blogs (e.g., http://reproducibleresearch.netlblog/).1

Within linguistics, investigations into possibilities for reproducible research have mostly been in the context of language documentation and description, in which the documentary fieldwork methodology has been noted for its potential to provide substantiation of scientific claims by promoting attention to the care and structuring of language data (Himmelmann 1998, 2006; Woodbury 2003, 2011; Thieberger 2009; Thieberger and Berez 2012; among others). Digital multimedia and annotations including transcripts and translations ostensibly allow readers to confirm claims about language structure by allowing direct access to the original observational data. This would mean that not only could

1 Misconduct in the life sciences has arguably greater consequences than it does in linguistics, and I am not necessarily advocating policing publications in our field.

example sentences in a grammar be confirmed as correctly transcribed, parsed and translated, but a sceptical reader could also determine whether or not she would reach the same conclusions about the phenomenon the example is meant to illustrate by providing access to the utterance in context. As with the example of frog-eating bats above, it is too cumbersome to require that descriptive linguistic claims be fully replicable, but it is not too cumbersome - in fact, it is desirous for the sake of 'good science' - to make them reproducible. A creative rewording of the Gezeltner quote above makes this clear:

> If a linguist makes a claim that a sceptic can only reproduce by spending three decades working in the same language community in the same sociolinguistic and fieldwork conditions, the original claim is really only reproducible in principle '" Our view is that it is not healthy for linguistic descriptions to be supported by examples that cannot be reproduced except by doing one's own fieldwork ... it may be *research* and it may be *important,* but unless enough details of the utterances in context are made available so that it can be subjected to true reproducibility tests by sceptics, it isn't Science. (modified from Gezeltner 2009, underlined words replaced, emphasis original)

Clearly, linguists cannot expect their colleagues to replicate fieldwork conditions (and doing so would not even necessarily lead to replicated utterances), but reproducibility may not be out of the question. Several authors have explored possibilities for providing direct access to the data upon which grammars are written, usually involving some appeal to the extensibility of structure that digital formats provide. Thieberger (2009), representing perhaps the most ardent endorsement of the benefits of reproducible grammar writing, outlines general principles for linking descriptions to corpora and lexica, but notes that generalised tools for doing so are not yet Widely available. Thieberger was able to create such a tool for his own (2006) grammar of South Efate, but software development is not often part of the ordinary working linguist's skillset. Maxwell (2012) provides an even more specific menu of data structures and software needs for producing a fully replicable grammar, including data structured as robust XML and a series of parsing engines and tokenisers. Unfortunately, the publishing industry

upon which most linguists rely has not yet caught up with these digital visionaries and we are still years away from a discipline-wide endorsement of radically linked grammars and source texts.

A simpler, albeit less robust, apparatus for linking linguistic claims to data may be available through a mechanism that already exists in academic publication: citation. If authors of descriptive linguistic materials can provide resolvable citations to original data in context - that is, a citation via a permanent handle to an archived language resource - it would be at least a step in the right direction. This of course presumes that the linguist has prepared source materials for archiving and has then deposited them in an appropriate digital archive (i,e., not a website, but a digital repository with an institutional commitment to preserving and migrating data in perpetuity) (Thieberger and Berez 2012, 100), a practice that is increasingly becoming the norm. Then the linguist need only provide an identifying handle or URL and a time code for each example in the grammar.

Practices in descriptive linguistics

In theory, providing sufficient citation sounds fairly simple. Provided the linguist archives well-structured digital files that link textual annotations to specific points in a media file (e.g., an ELAN file and an audio file), simple citation should be a straightforward process.? Many descriptive linguists have already long been providing at least some **form of citation for** examples, **for instance) the initials of the speaker** who uttered the example, or a reference to a field notebook. But linguistics has not fostered a culture of providing full citation Or even of making data locatable. Berez et al. (in prep) is a study of data citation practices in descriptive grammars, descriptive PhD theses, and linguistics journals from a ten-year span between 2003 and 2012, beginning five years after Himmelmann's (1998) position paper on language documentation as, among other things, a way to provide accountability in linguistics: '[Language] documentation ... will ensure that the collection

2 **In practice, linguists will need** to **decideon discipline-wide formats for citing many kinds of primary data, not just digital media.**

and presentation of primary data receive the theoretical and practical attention they deserve' (1998, 164). Berez et al. (in prep) have found that by and large, authors of grammars, theses and articles rarely even indicate *if* or *where* data is stored, let alone provide some indication of from where in a corpus a particular example was retrieved

In a sample of 45 published grammars, among those authors who stated explicitly where their data were located, the largest number of them (eight) had archived the data in a dedicated repository. Five authors indicated a plan to archive data in the future, and nine made some textual materials available via paper publication, either in the same volume or in a different volume. Two authors clearly stated that their data were unpublished (with no indication of a plan to make data available); one made data available on an accompanying website; one stated that data was backed up at his place of residence; one stated the data had fallen victim to a political uprising. Importantly, fewer than half the authors surveyed even considered it important to mention the location of data: 29 of the 45 surveyed did not make explicit statements about the location of their data. A sceptical reader would not even know where to look.

Berez et al. (in prep) also investigates methods of citation of example sentences. In the same 45 published grammars discussed above, the tendency is to provide less information to readers, rather than more, and 25 grammars used no discernable method for citing example sentences back to the primary data from whence they came. Of those that did, however, the citation form is ad hoc, usually with minimal information like speakers' initials or name (e.g., */M),* sometimes with a date (e.g., *Tom Smith, 2009-04-07)* or, rarely, with a reference to the linguistic data type (e.g., *narrative).* In some cases, the author makes no reference to the data even being part of a larger corpus, but it some instances there is at least *some* indication that an item is a member of a group of materials conceived as unified along some parameter (e.g., *Notebook 12, p.* 16 or *KC, tape* 3 *of27).*

Only three of the 45 grammars so far examined in Berez et al. (in prep) include a citation in the format recommended by Thieberger (2009), in which an (ostensibly) resolvable permanent handle links to an item or items stored in a digital archive, with or without time offsets for the utterance. For example, this could be of the form *Peter Wee,*

*oai:scholarspace.manoa.hawaii.edu: NL1-042, 00:02:44.4-00:02:46.*0.3 In this example, the author cites the name of the speaker, the resolvable URL to a file named NL1-042 that is stored in the ScholarSpace repository at the University of Hawai'i at Mānoa (aka, Kaipuleohone, see below), and the starting and ending timecodes of the utterance in the recording (here, starting two minutes, 44.4 seconds from the beginning of the recording and ending two minutes, 46 seconds from the beginning of the recording).

Teaching postgraduate students to be more scientific

Language documentation at UHM

The Department of Linguistics at the University of Hawai'i at Mānoa (UHM) offers advanced degrees (MA, PhD) focusing on language documentation.s As such, it seems we have a responsibility to investigate novel ways to promote the methods and goals of language documentation in our curriculum. Calls for training in language documentation invariably include training in linguistic data management techniques (jukes 2011; see also the curricula of the Infields and CoLang workshops," as well as the grantee training of the Endangered Languages Documentation Programme"), and these very techniques are those that would enable linguists to prepare data in ways that facilitate archiving and citation. However, given that the publishing world does not require authors to cite back to archived primary data, how can we expect students to undertake this task? Even in the program at UHM, most of the PhD theses over the last ten years have not cited examples back to resolvable resources, with the result that most of our students' linguistic claims are not reproducible.

---

3 Many thanks to Huiyong NaIa Lee for providing this citation example.
4 http://www.ling.hawaii.edu.
5 http://www.linguistics.ucsb.edu/faculty/infield and http://linguistiCS.uoregon.edulinfield20*lO/home.*
6 http://idrh.ku.edu/colang2012, http://www.uta.edu/faculty/cmfitz/swnaI/projects/CoLang.
7 http://www.eldp.net.

Fortunately, however, we are well-positioned to instil good habits in our students because of a number of fortuitous features of the program. First, we have several required courses in which proper data management and the creation of structured digital media and annotations can be learned, practised, and then mined for evidence for claims. These include LING 710: Methods in Language Documentation; LING 630: Field Methods; LING 617: Language Revitalization; and LING 640: Methods of Language Conservation. The result is that when students begin their own fieldwork data collection, they have already had enough experience at creating structured data so that they can begin using best digital practices immediately, rather than needing to go back and retrofit a less well-structured corpus of materials later. In addition to targeted coursework, we are also home to the Kaipuleohone University of Hawai'i Digital Language Archive," a digital repository hosted in the University of Hawai'i Library's D-Space repository, Scholaropace,? Kaipuleohone is fully compliant with the Open Language Archive Community's (OLAC[10]) metadata standards, and is a member of the Digital Endangered Languages and Music Archives Network (DELAMANll) (Berez 2013).

Recent changes to *our requirements*

As mentioned above, only two of the descriptive PhD theses (i.e., grammars) from the past decade (i) indicate that field data was archived and (iil provide citations for examples back to the archived data. In response to this low level of archiving and citation, during the 2013-14 academic year the faculty in the department decided that encouraging archiving and citation was insufficient if we are to effectively communicate to our students that we value reproducibility. In the fall semester, the faculty elected to make the first of two major changes to the PhD

8 http://www.kaipuleohone.org.Kaipuleohonemeans.gourdofsweetwords·in **Hawaiian. We are grateful to Laiana Wong for suggesting this name for the archive.**

9 http://scholarspace.manoa.hawaiiedu.

10 http://www.language-archives.org.

11 http://www.delarnan.org.

Handbook. This change added language stating that (i) students whose theses were based on data collected during their own fieldwork, regardless of linguistic subfield, were required to properly archive their data; (ii) students were to develop an archiving plan as a component of the required thesis proposal; and (iii) proof of deposit must be given to the thesis committee before a thesis could be approved This change ensured that students would plan for archiving early in their graduate careers, and would hopefully train them to continue the practice into their professional lives.

Later, in the spring semester of 2014, the faculty elected to make the second change to the PhD Handbook, this time requiring proper citations to archived materials. Data in theses coming from a student's own archived materials must now be cited via a persistent identifier to the source file. After some discussion, it was determined that because different subfields have different practices for citation, the exact format and level of granularity - for instance, to a timecode in a specific audio file for examples from discourse, or to a collection of scanned field notebooks for historical linguistics, or to a dataset from an experiment - would be developed in consultation with the thesis advisor. The final wording of the new additions to the PhD Handbook are below:

> The Department of Linguistics values proper data management and citation. Students whose dissertations are based on data collected during the course of their own fieldwork are required to properly archive their data in an appropriate language archive in order to ensure the longevity of the data. Students will develop an archivi.ng plan early and will include a description of this plan in the Dissertation Proposal. Data can be archived with Kaipuleohone, the University of Hawai'i Digital Language Archive or with another accepted archive (for example, a member archive of the DELAMAN network). For students archiving their data in Kaipuleohone, the archivi.ng plan should be developed in consultation with the current archive director. All students will be required to submit proof of deposit to the committee before the dissertation can be approved.
>
> In addition, each student is required to cite data in the dissertation coming from his or her own archived materials via a persistent identifier URL to the source file in the archive. The exact format of

the citation and the level of granularity (e.g., timecode in an audio file; collection of files; dataset; etc.) can be developed in consultation with the dissertation advisor, and should reflect the best practices in the student's linguistic subfield.

Frequently asked questions

The new requirements for archiving and citation apply to students entering the program from fall 2014 onwards. Before the changes came into effect, three PhD students voluntarily followed the proposed requirements and provided feedback On workflow integration. We continue to work with these students, but so far their response has been positive, even enthusiastic.'! Nonetheless, a few (rhetorical and actual) questions arise. Preliminary answers to some of these are below, and no doubt we will continue to refine how we put our new policies into practice.

Q: After the thesis is submitted, the author wants to revise a transcription, gloss, parse, analysis, etc. Is the citation now incorrect?

A: No. The citation is to the primary data itself (i.e., the media file), not the transcription or analysis per se.

Q: The student wants or needs to keep data temporarily private or inaccessible, either because of privacy concerns with the data provider, or to discourage 'scooping: Is this going to put the student's degree at risk?

A: No. The student will still cite the archived materials. Requests for access to files that are not freely available will be handled like any other such request, by contacting the depositor. We are still able to balance the need for privacy.

12 One student writes, 'I have to say that archiving is one of the best things I ever decided to do for this dissertation. After four chapters, I have almost 400 example sentences, all attributed to specific timings in specific files in my archive. Metadata and a mix of software makes everything so easy to find. If it wasn't for the archive, I can't imagine even getting to this stage relatively unscathed. So thanks!' (Email from Nala Lee, 1 April 2014).

Q: The student later archives multiple 'versions' of the same primary data. Is the citation now incorrect?

A: No. Students, like other depositors, are encouraged to reference the original, unedited version of the primary data. Later versions can be associated to the original file in the archive metadata.

Q: Isn't it overwhelming to archive and cite while also writing a thesis?

A: Students are taught early to integrate preparing for archiving and citation early through required coursework. The intention is for this to become part of expected departmental culture, and for students to accept archiving and citation as part of the rigorous steps for thesis research and beyond.

Conclusion

Linguistics departments routinely make values declarations by requiring milestones to a degree. For instance, by requiring students to write qualifying papers, we are stating that we value the ability to write article-length research papers. By requiring comprehensive exams, we are stating that we value being able to talk and write eloquently about linguistics. By requiring PhD theses, we are stating that we value being able to plan and execute independent research. Given that we are ultimately training linguistic scientists, why would we value reproducibility any less than the aforementioned skills? Students, like anyone, are less likely to adopt practices that are seen as unessential, and will not often spend their time On activities they do not get credit for doing, but by teaching students how to archive and cite data properly; and then not only recommending but requiring it, we are making a statement that we value reproducibility.

Works cited

Berez, Aodrea L (2013). 'The digital archiving of endangered language oral traditions: *Kaipuleohone* at the University of Hawai'I and *C'ek'aedi Hwnax* in Alaska: *Oral Tradition 28(2)*.

Berez, Andrea 1., Lauren Gawne, Tyler Heston and Barbara Kelly *(2015).* 'Citation **and transparency in descriptive linguistics: (unpublished manuscript)**

Buckheit, Jonathan B. and David 1. Donoho (1995). 'WaveLab and reproducible **research: In *Wavelets and Statistics*, edited by Anestis Antoniadis and Georges** Oppenheim, 55-81. New York: Springer.

Crocker, Jennifer and M. Lynne Cooper *(2012).* 'Addressing scientific fraud: *Science* 334: 1182. doi: 10.1126Iscience.1216775.

Donoho, David L (2010). 'An invitation to reproducible computational research: *Biostatistics* 11(3): 385-88.

Fang, Ferric C. and Arturo Casadevall *(2011).* 'Retracted science and the retraction index' *Infection and Immunity* 79(10): 3855-59.

Fang, Ferric C; R. Grant Steen, and Arturo Casadevall (2013). 'Misconduct **accounts for the majority of retracted scientific publications: *PNAS Early*** *Edition* 334: 1-6.

GezeItner, Dan *(2009).* 'Being scientific: Falsifiability, verifiability, empirical tests, and reproducibility.' *The OpenScience Project.* http://openscience.orgi blog/?p=312.

Himmelmann, Nikolaus P (1998). 'Documentary and descriptive linguistics: *Linguistics* 36: 161-95.

Himmelmann, Nikolaus P (2006). 'Language documentation: What *is* it good for?' In *Essentials of Language Documentation,* edited by [ost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, 1-30. Berlin: Mouton de Gruyter.

Jukes, Anthony *(2011).* 'Researcher training and capacity development in language documentation: In *The Cambridge Handbook of Endangered Languages,* edited by Peter K. Austin and Julia Sallabank, 423-45. Cambridge: **Cambridge University Press.**

de Leeuw, Jan *(2001).* 'Reproducible research: The bottom line: *UCLA Department of Statistics Papers.* http://escholarship.orgiuc/item/9050x4r4.

Marcus, Adam and Ivan Oransky (2012). 'Bring on the transparency index: *The Scientist.* http://tiny.cc/2012-transp-marcus.

Maxwell, Mike *(2012).* 'Electronic grammars and reproducible research: In ***Electronic Grammaticography (Language Documentation and Conservation*** *Special Publication No.4),* edited by Sebastian Nordhoff, *207-34.*

Ryan, Michael J. *(2011).* 'Replication in field biology: The case of the frog-eating bat.' *Science* 334: *1229-30.*

Thieberger, Nicholas *(2006). A Grammar of South Efate: An Oceanic Language of* ***Vanuatu.*** **Honolulu: University** of Hawai'i **Press.**

Thleberger, Nicholas *(2009).* 'Steps toward a grammar embedded in data: In *New Challenges* in *Typology: Transcending the Borders and Refining the*

Distinctions. **edited by Patricia Epps and Alexandre Arkhipov, 389-408. Berlin; New York, NY: Moutonde Gruyter Mouton.**

'Ihieberger, Nicholas and Andrea 1. Berez (2012). 'Linguistic data management: In *The Oxford Handbook of* Linguistic *Fieldwork,* edited by Nicholas Thieberger, 90-118. Oxford: Oxford University Press.

Tomasello, Michael and Iosep Call (2011). 'Methodological challenges in the study **of primate cognition: *Science* 334: 1227-28.**

Woodbury, Anthony C. (2003). 'Defining documentary linguistics: In *Language Documentation and Description, Volume* 1, edited by Peter K. Austin, 35-5l. London: The Hans Rausing Endangered Languages Project.

Woodbury, Anthony C. (2011). 'Language documentation: In *The Cambridge Handbook of Endangered Languages,* edited by Peter K. Austin and Julia Sallabank, 159-86. Cambridge: Cambridge University Press.