

3

Language Documentation & Conservation Special Publication No. 15
Reflections on Language Documentation 20 Years after Himmelmann 1998
ed. by Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton, pp. 22–32
<http://nflrc.hawaii.edu/ldc/>
<http://hdl.handle.net/10125/24804>

*“It is simply a feature of a scientific enterprise
to make one’s primary data accessible to further scrutiny”*
(Himmelmann 1998: 165)

Reflections on Reproducible Research

Lauren Gawne
La Trobe University

Andrea L. Berez-Kroeker
University of Hawai'i at Mānoa

Reproducibility in language documentation and description means that the analysis given in descriptive publication is presented in a way that allows the reader to access the data on which the claims are based, to verify the analysis for themselves. Linguists, including Himmelmann, have long pointed to the centrality of documentation data to linguistic description. Over the twenty years since Himmelmann’s 1998 paper we have seen a growth in digital archiving, and the rise of the Open Access movement. Although there is good infrastructure in place to make reproducible research possible, few descriptive publications clearly link to underlying data, and very little documentation data is publicly accessible. We discuss some of the institutional roadblocks to reproducibility, including a lack of support for the development of published primary data. We also look at what work on language documentation and description can learn from the recent replication crisis in psychology.

1. Introduction¹ Himmelmann 1998 seeks to highlight the distinctiveness of language documentation from linguistic description, as well as their “bilateral mutual dependency” (p. 165). Fundamentally, however, the paper is a discussion of the role of data in linguistic analysis. Documentation is the collection and organization of data, and description is the analysis of that data. Himmelmann is adamant throughout the article that the only way documentation and description can be successful is if claims about how a language works

¹Acknowledgments: Our thanks to Peter Austin and Suzy Styles for fruitful discussion about data. Thanks also to our grammar citation and methods collaborators Barbara F. Kelly and Tyler Heston. We would also like to thank our colleagues in the Linguistics Data Interest Group of the Research Data Alliance, particularly our co-chair Helene Andreassen.

can be supported by allowing the reader access to the data on which those claims are founded.

In brief, *reproducibility* (e.g., Buckheit & Donoho 1995; de Leeuw 2001; Donoho 2010) in research means that the data on which publications are based are made available so that other scientists could ostensibly verify the results for themselves. This is distinct from the process of *replication*, in which the steps of a research project are replicated by another scientist, yielding new data which can confirm or contradict previous data. Replication is well-suited to laboratory research, but language documentation data is essentially behavioral data, and analysis of data that is grounded in a specific interactional context that is arguably impossible to replicate. So while replication is not a fruitful aspiration for most documentation-based description, we wholeheartedly agree with Himmelmann that reproducibility is.² Descriptive work needs to be based on sound documentary research methods, and those methods should be made clear by authors of descriptive publications. Relatedly, any published claims about language should be supported by evidence that the audience can, with reasonable considerations for privacy concerns, also access.

Below we examine the context in which Himmelmann (1998) was published, and the developments in linguistics in the last two decades, with regard to the development of digital archiving and open access. We then look at what we can learn as a field from the unfolding crisis of replicability in the field of psychology, and the future of language documentation, description, and data.

Himmelmann (1998) was participating in a larger discussion about the role of data in language documentation, and linguistics more broadly. Sally Thomason, writing as the editor of *Language* in 1994, also articulated concern for clarity regarding the data sources. She called upon linguists to provide “...detailed information about sources of data and methodology of data collection” (Thomason 1994: 413). Some linguists were already actively engaging with data citation in their descriptive linguistic writing. Simpson, at the beginning of her 1983 PhD dissertation on Warlpiri morphology and syntax, states “I have tried to indicate the source of each example sentence where I know it. If the example sentence is made up, I have indicated this, unless the sentence is elementary” (1983: 4). Data citation is an important feature of reproducible research, but it is only of use if the interested reader can resolve that citation to the original data. Digital archives have provided an important development in data sharing.

2. Development of archives One of the most immediately obvious developments in documentation since Himmelmann (1998) is the network of digital archives that provide a persistent and secure location for the storage of linguistic data. Himmelmann voices his concern that “In recent decades, hardly any comprehensive collections of primary data have been published” (1998:164), a concern that is objectively no longer true thanks to the rise of digital archiving. The permanent preservation of one’s materials, once a privilege reserved for only the most senior linguists, is now a common part of the documentary linguist’s workflow.

While analog language archiving had been part of anthropological practice since the late 19th century, the development of digital archiving methods for language documentation began in earnest in the early 21st century (see e.g., Woodbury 2011; Henke & Berez-Kroeker 2016). Those years saw the rise of funding schemes for language

²See Berez-Kroeker et al. 2018 for a discussion on reproducibility in linguistics in general.

documentation like DoBeS³ in 2000 and ELDP⁴ in 2003, both of which provided a repository for preserving their grantees' work. The NSF-funded Electronic Metastructure for Endangered Languages Data (EMELD)⁵ project provided much-needed education to linguists on how to digitally preserve language documentation (Boynton et al. 2010).

Alongside archiving has come a standardization of metadata. Himmelman does not actually use the term 'metadata'—instead the article refers to “information to be included” (1998: 189, see also 169–170)—while discussing what has now become known as 'metadata' at some length. In the early 2000s, the Open Language Archives Community (OLAC)⁶ was building a metadata standard based on Dublin Core specifically for describing digital language materials (e.g. Bird & Simons 2003). Similarly, the International Standards for Language Engineering Metadata Initiative (IMDI)⁷ was developed in the DoBeS context.

Digital archives not only provide persistent data storage, but they also provide access to the data thanks to improvements in the internet. Himmelman is sensitive to placing speaker attitudes at the centre of archiving models with a focus on controlled access (1998: 171–175, 189). There has been considerable discussion about ethics and access to documentation materials (Dwyer 2006; Garrett & Conathan 2009; Macri & Sarmento 2010; Shepard 2016), and some archives have implemented different levels of accessibility to materials (eg Green et al. 2011; Nathan 2010; 2014). This is an ongoing conversation, as internet access is still not globally balanced, and speakers of many of the languages represented in archives are unable to view or use deposited materials through lack of access. In terms of reproducible research at least, we have solved many of the barriers that were a concern in 1998, and can now do a lot more than meet Himmelman's minimal solution of providing an “edited version of the fieldnotes” (1998: 165).

3. Open Access The Open Access movement (OA) was beginning to coalesce in the late 1990s, and has been an important influence on the development of archiving practice in documentation. In 1997 the Association of Research Libraries developed the Scholarly Publishing and Academic Resources Coalition (SPARC),⁸ which had an early focus on encouraging open access journal publishing.⁹ OA radically altered the publishing landscape (Joseph 2013), and we see that effect today, with journals like *Language Documentation & Conservation*¹⁰ and presses like *Language Science Press*¹¹ that cost nothing to authors or readers. The OA movement is now actively involved in encouraging open access data practices (SPARC n.d.; Kitchin 2014). Language documentation archives have been leaders within the humanities and social sciences when it comes to advocating for open access, or at least mixed access for different uses.

OA publication has been aided by the creation of Creative Commons (CC) licenses.¹² Founded in 2001,¹³ CC allows copyright holders to specify how members of the public

³<http://dobes.mpi.nl/>

⁴<http://www.eldp.net/>

⁵<http://emeld.org/>

⁶<http://www.language-archives.org/>

⁷<http://tla.mpi.nl/imdi-metadata/>

⁸<http://sparcopen.org/>

⁹<http://sparcopen.org/our-work/research-data-sharing-policy-initiative/>

¹⁰<http://nflrc.hawaii.edu/ldc/>

¹¹<http://langsci-press.org/>

¹²<http://creativecommons.org>

¹³<http://creativecommons.org/about/history/>

may and may not use their work, including whether attribution is required and whether commercialization is allowable. The licenses are both machine- and human-readable for ease of use. Many archives now use CC licenses for OA data. The CC framework provides a scaffold for discussions between language documentation researchers and communities, making this issue somewhat easier to navigate than it was when Himmelmann was writing (1998: 175).

4. Data sharing in today's practice While archives have provided a robust way to share data, we still are not seeing complete uptake of archiving or other practices that lead to reproducibility. In a survey of one hundred descriptive grammars published between 2003 and 2012 that we conducted with Barbara F. Kelly and Tyler Heston, we found that data archiving before publication was only mentioned in 22 publications, and only eight publications included data citation that resolved back to a locatable corpus (Gawne et al. 2017). Many published grammars in our survey do not discuss basic methodological information like the number of speakers who contributed, or recording equipment used, which prevents the reader from understanding the nature of the data on which analysis is built. In a similar vein, Thieberger (2017) looked at 1,708 grammars published since 1967 and found that for 1,253 of the languages there were fewer than 40 items in an OLAC archive, indicating that for the vast majority of descriptive grammars the primary data on which they are based cannot be found or used.

Language documentation has become a field with its own journals, conferences, network of archives and funding, but there remains a fundamental disconnect between documentation data and subsequent description. A major reason for this is the fact that the academic environment does not provide incentives for good practice in reproducibility. We add our voices to Himmelmann's in seeking better transparency in research methodology to ensure that readers can better judge the "reliability, naturalness, and representativeness of the data" (1998: 162), and we believe the best way to do this is through archiving and citation.

Preparing data for archiving is a time-consuming process that is not viewed as having academic merit on par with published analyses. Management and curation of data for archiving is a time-consuming process, even when the documentation workflow is set up to optimize the process. This means that even the best-intentioned documentation practitioner can find themselves with a large amount of work to do that is undervalued by university hiring, tenure and promotion committees. Descriptive work, in contrast, results in peer-reviewed publications, which are still the primary yardstick for measuring academic productivity.

The status quo is changing to some extent. Some initiatives have sought to use the current incentive structure to give recognition to documentation work, such as *Dictionaria*, which uses a peer-reviewed model for digital dictionary databases;¹⁴ the *Language Contexts* series in *Language Documentation and Description*, which publishes contextualising metadata for a language;¹⁵ and the publication of descriptions of archival collections in *Language Documentation & Conservation*, which act as citable proxies for datasets within current citation mechanisms (e.g. Salfner 2015).

Other efforts have been directed at raising the profile of documentation and corpus building. In 2010 the Linguistic Society of America passed the *Resolution Recognizing*

¹⁴<http://home.uni-leipzig.de/dictionaryjournal/about-the-journal/>

¹⁵www.e-publishing.org/language-contexts

the Scholarly Merit of Language Documentation, which recognized corpora and other documentation outputs as “scholarly contributions to be given weight in the awarding of advanced degrees and in decisions on hiring, tenure, and promotion of faculty.”¹⁶ In a similar spirit, the DELAMAN Franz Boas Award “recognizes and honours junior scholars who have done outstanding documentary work in creating a rich multimedia documentary collection of a particular language that is endangered or no longer spoken.”¹⁷ While it is important that there are positive motivators for archiving, a great deal of the archiving undertaken in recent years stems from a more prosaic motivation: funders increasingly require data to be archived, and open access where feasible, as part of the funding process (Austin 2014).

We are still grappling with the question of how to assess the quality of archival collections. While we acknowledge the existing peer review mechanisms for publications are not without their failings, as a discipline we have not yet come up with a commonly agreed-upon way to assess the quality of documentation collections (though note the recent draft for a *Statement on the Evaluation of Language Documentation for Hiring, Tenure, and Promotion*¹⁸ by the Linguistic Society of America and work of the committee of the Australian Linguistic Society, reported in Thieberger et al. 2016). Himmelmann also observed the need to assess documentation work (1998: 181). He focuses mainly on different types of data collection, such as elicitation, tasks and different genres of spontaneous text (1998: §3.3), however there are many factors that need to be considered including quality of recordings, number of speakers, presence of video data, and quality of metadata (Woodbury 2014; Thieberger et al. 2016).

5. Data citation in today’s practice While the move towards accessible corpora has been one challenge, another has been the lack of citation of that documentation data in publications. Editors and publishers, for the most part, have not made explicit an expectation to cite examples of linguistic phenomena (sentences, lexical items, etc.) back to the dataset whence they came. While most linguists would never dream of quoting from another author’s work without a proper citation, those same scholars will happily quote from their own extensive corpora without any citation whatsoever.

We believe in the need for data citation, and have been working alongside our colleagues in the Linguistics Data Interest Group of the Research Data Alliance,¹⁹ to bring these beliefs together in a document known as the *Austin Principles of Data Citation in Linguistics*.²⁰ At the core of these principles is the belief that “[l]inguists should cite the data upon which scholarly claims are based.” (Berez-Kroeker et al. 2017), a belief that echoes the quote from Himmelmann in the epigraph to this chapter. Data citation can help the researcher return to the original data to confirm hypotheses as analysis develops, and it can also help a reader locate the example in the corpus, to seek more contextual information to reproduce the original hypothesis, or for an analysis that the original data was not necessarily presented with a focus on (e.g. looking at the case-marking in a sentence that was originally used to exemplify a feature of tense). Although we are accustomed to seeing example sentences presented as written artefacts, we agree

¹⁶www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation

¹⁷<http://www.delaman.org/delaman-franz-boas-award/>

¹⁸www.linguisticsociety.org/content/draft-lsa-statement-evaluation-language-documentation-hiring-tenure-and-promotion

¹⁹<http://rd-alliance.org/groups/linguistics-data-ig>

²⁰<http://site.uit.no/linguisticsdatacitation/>

with Himmelmann that most contextualization of the utterance is lost in print, including prosody and gesture, as well as the possibility to “gloss over” complexities (1998: 191 fn5).

Citing your own data encourages others to cite your data in their work as well. Himmelmann notes that any language documentation corpus includes information well beyond the scope of what a single researcher or team can undertake to analyze (1998: 163). We have seen very little uptake of documentary data in descriptive work published by researchers other than the data collector(s); in our study of articles published in *Linguistic Typology* between 2012–2017 we found that the overwhelming majority of authors draw on published descriptions or their own documentation data (Gawne et al. 2017), but almost never the datasets of other data collectors. We believe that ultimately the citation of data will become standard practice, through editorial policies that make it a norm like other forms of citation.

6. The replication crisis in psychology Linguistics is not the only field to have considered the role of data and analysis in research. The ‘replication crisis’ that started in medical science (see Goldacre 2010 for a summary) and is now being played out in social psychology (Chambers 2017) has much to teach us about the importance of transparency in research methods and data presentation, as well as how we can best approach these themes as a community of researchers. As we discussed above, we do not believe that language documentation and description should strive for replication, which is more relevant to psychology, but there are lessons in this crisis for the future of reproducibility as well. Psychology, like linguistics, is interested in thresholds, not absolutes, in the often-difficult to establish nature of human behaviour.

In 2011 Daryl Bem, a social psychologist, published a paper that demonstrated, across a series of experiments, statistically significant effects of ‘precognition’, with participants appearing to contradict the flow of time and show priming effects on early parts of the experiment based on later parts of the experiment (Bem 2011). The research methods were all meticulously reported, leaving the reviewers to either decide that ‘precognition’ did exist, or the methods of social psychology were not reliable. Bem’s work appears to have been shaped by a ‘forking paths’ analysis (also known as ‘experimenter degrees of freedom’), where each decision in the analysis process appears to be sensible, in keeping with the norms and best practice of the field, but helps the researcher converge on the outcome they want. Bias in each step of data collection can lead to bias in the analysis, which can lead to bias in the meta-analysis that shapes the trajectory of the field. In linguistics, we’ve seen that some topics have been neglected as specific targets for documentation work, because they’ve been considered marginal to a particular conceptualization of language. These phenomena may eventually be shown to be less marginal than had been originally thought (e.g. ideophones, see Dingemanse et al. 2018).

Bem’s case is egregious because believing his findings contradicts the basis of causality on which our understanding of the universe is built, but there are a number of other questionable research practices that the field of psychology is critically analysing. One of these is *hypothesising after results are known*, or HARKing—where the narrative for the data is often changed to fit a more compelling hypothesis after collection is complete and analysis has begun (e.g. deciding that the variable of gender is the significant difference, even though that wasn’t the original aim of the experiment). The other is *p-hacking*, running numerous statistical processes to ‘find’ results in the data, which then leads to HARKing to create a publishable narrative (see, for example, the ‘pizzagate’ controversy surrounding work by Brian Wansink and colleagues (problems with this

research summarized in van der Zee et al. 2017), in which data about pizza consumption was sliced (like an unethical pizza) into statistically-significant subsets to fit the research narrative). Although most descriptive work does not require the explicit formation of hypotheses, researchers do have a set of expectations about what linguistic features a language might demonstrate, based on the typological profiles of related languages. Similarly, a researcher must always select the example sentences to illustrate a descriptive grammar, which is by no means an objective process. Providing the reader with additional examples through presentation of the original data can help mitigate these limitations of descriptive work, in the way that pre-registering hypotheses and presenting data sets is helping in the field of psychology.

The crisis of replicable methodology in psychology was the motivation for Brian Nosek and hundreds of colleagues to attempt to replicate 100 experimental psychology studies published in 2008 (Open Science Collaboration 2015). Fewer than fifty percent of the studies were successfully replicated. In 2013, while working on the replication study, Nosek and colleagues started the Center for Open Science,²¹ a researcher-driven organization that builds easy-to-use tools and protocols, as well as leading discussions about the nature of research practice.

7. Looking ahead The problems in psychology arose in part because research practices were not transparent enough. Researchers generally were not required to present their methodology in a way that ensured replication, nor to commit to a course of research, maintain it through to publication and share the underlying data. In recent years, the move towards greater transparency in psychology has included ‘pre-registration’ of methods, either as a peer-reviewed process that becomes the first half of the final peer-reviewed paper,²² or as a non-reviewed methodology that is time-stamped and limits the ‘researcher degrees of freedom’ that can influence the final outcomes.²³

Language documentation does not have the same experimental focus, so we would not want pre-registration as a solution to our research problems, nonetheless we still generally don’t make it easy for our readers to access the datasets on which our analyses are based and are therefore equally susceptible to research pitfalls. Even Thomason noted during her tenure as editor of *Language* that erroneous data “occur[ed] frequently—so frequently, in fact, that the assumption that the data in accepted papers is reliable began to look questionable” (1994: 409). We need to continue to develop a social and technological infrastructure in linguistics that allows us to reap tangible rewards for the creation, management, and citation of linguistic data as much as we do for linguistics publications. In short, we still don’t value language *documentation* as much as we value linguistic *description*. We can do better.

Language documentation has changed over the last 20 years thanks to the development of digital data collection methods, and online archives that allow for both the storage and dissemination of recorded materials. While we have made some moves towards a more open approach to data that would support research reproducibility, there is still more work to be done to ensure that the link is made clear between linguistic description and the documentation that it is based upon. At a minimum, we believe

²¹<http://cos.io/>

²²Very recently Timo Roettger of Northwestern University has put together an initiative to encourage more linguistics journals to adopt Registered Reports (RRs). For information on the initiative see <http://linguistlist.org/issues/29/29-3168.html>

²³<http://cos.io/prereg/>

that all data from documentation should be archived with a digital repository that has a mandate for long-term storage. Where the data are not sensitive or controversial, they should be made accessible to both the language speakers and to researchers who wish to confirm existing analyses, test new analyses or explore previously under-described phenomena in the language. Descriptive work should clearly state the research methods used in collecting the data that forms the basis of the research, make clear where the data are located and should explicitly link each piece of data to its place in the documentation data. Digital archives and the Open Access movement have given us the tools to make this happen. When all of this is common practice, and not just the practice of a subset of researchers, we will have made a clear move in the direction of reproducible research.

References


- Austin, Peter K. 2014. Language documentation in the 21st century. *JournaLIPP* 3. 57–71.
- Bem, Daryl J. 2011. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology* 100(3). 407–425. (doi:10.1037/a0021524)
- Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1). 1–18. (doi:10.1515/ling-2017-0032)
- Bird, Steven, & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 57–582.
- Boynton, Jessica, Steven Moran, Helen Aristar-Dry & Anthony Aristar. 2010. Using the EMELD School of Best Practices to create lasting digital documentation. In Lenore A. Grenoble & Louanna Furbee-Losee (eds.), *Language documentation: Practice and values*, 133–146. Amsterdam, Philadelphia: Benjamins.
- Buckheit, Jonathan B. & David L. Donoho. 1995. WaveLab and reproducible research. In Anestis Antoniadis & Georges Oppenheim (eds.), *Wavelets and statistics*, 55–81. New York: Springer.
- Chambers, Chris. 2017. *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton: Princeton University Press.
- Dingemanse, Mark. 2018. Redrawing the margins of language: Lessons from research on ideophones. *Glossa* 3(1). 1–30. (doi:10.5334/gjgl.444)
- Donoho, David L. 2010. An invitation to reproducible computational research. *Biostatistics* 11. 385–388.
- Dwyer, Arienne M. 2006. Ethics and practicalities of cooperative fieldwork and analysis. In Jost Gippert, Nikolaus P. Himmelmann, & Ulrike Mosel (eds.), *Essentials of language documentation*, 31–66. Berlin: Mouton de Gruyter.
- Garrett, Andrew & Lisa Conathan. 2009. Archives, communities, and linguists: Negotiating access to language documentation. Presentation at the *Linguistic Society of America Annual Meeting*. (http://www.ailla.utexas.org/site/lisa_olac09/conathangarrett_lsa_olac09.pdf)
- Gawne, Lauren, Andrea L. Berez-Kroeker & Helene N. Andreassen. 2017. Data citation in linguistic typology: Working towards a data citation standard in linguistics. Presentation at *Association for Linguistic Typology 12*. Canberra: December 11–15.
- Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11. 157–189.
- Goldacre, Ben. 2010. *Bad science: Quacks, hacks, and big pharma flacks*. London: McClelland, Stewart.
- Green, Jennifer, Gail Woods & Ben Foley. 2011. Looking at language: Appropriate design for sign resources in remote Australian Indigenous communities. In Nick Thieberger, Linda Barwick, Rosey Billington & Jill Vaughan (eds.), *Sustainable data from digital research: Humanities perspective on digital research*, 66–89. Melbourne: Custom Book Centre, The University of Melbourne.

- Henke, Ryan E. & Andrea L. Berez-Kroeker. 2016. A brief history of archiving in language documentation, with an annotated bibliography. *Language Documentation & Conservation* 10. 411–457.
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 6. 161–195.
- Joseph, Heather. 2013. The Open Access Movement Grows Up: Taking Stock of a Revolution. *PLoS Biol* 11(10). (e1001686.doi:10.1371/journal.pbio.1001686)
- Kitchin, Rob. 2014. *The data revolution*. London: Sage.
- Leeuw, Jan de. 2001. Reproducible research: The bottom line. *UCLA Department of Statistics papers*. (<http://escholarship.org/uc/item/9050x4r4>) (Accessed 28 March 2018)
- Macri, Martha & James Sarmiento. 2010. Respecting privacy: Ethical and pragmatic considerations. *Language & Communication* 30(3). 192–197.
- Nathan, David. 2010. Archives 2.0 for endangered languages: From disk space to MySpace. *International Journal of Humanities and Arts Computing* 4(1–2). 111–124.
- Nathan, David. 2014. Access and accessibility at ELAR, an archive for endangered languages documentation. *Language Documentation and Description* 12. 187–208.
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 349(6251): aac4716. (doi:10.1126/science.aac4716)
- Salfner, Sophie. 2015. A guide to the Ikaan language and culture documentation. *Language Documentation & Conservation* 9. 237–267.
- Shepard, Michael Alvarez. 2016. The value-added language archive: Increasing cultural compatibility for Native American communities. *Language Documentation & Conservation* 10. 458–479.
- Simpson, Jane. H. 1983. Aspects of Warlpiri morphology and syntax. Massachusetts Institute of Technology PhD dissertation. (<http://hdl.handle.net/1721.1/15468>) (Accessed 2018-03-18)
- SPARC. n.d. Open Data Factsheet (11.10-2). (<http://sparcopen.org/open-data/>) (Accessed 2018-04-2012)
- Thieberger, Nick, Anna Margetts, Stephen Morey & Simon Musgrave. 2016. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36. 1–21. (doi:10.1080/07268602.2016.1109428)
- Thieberger, Nick. 2017. LD&C possibilities for the next decade. *Language Documentation & Conservation* 11. 1–4.
- Thomason, Sarah. 1994. The editor's department. *Language* 70. 409–423.
- van der Zee, Tim, Jordan Anaya & Nicholas J. L. Brown. 2017. Statistical heartburn: An attempt to digest four pizza publications from the Cornell Food and Brand Lab. *PeerJ Preprints* 5:e2748v1. (doi:10.7287/peerj.preprints.2748v1)
- Woodbury, Anthony. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 159–211. Cambridge: Cambridge University Press.

Woodbury Anthony C. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. *Language Documentation and Description* 12. 19–36.


Lauren Gawne

l.gawne@latrobe.edu.au

 orcid.org/0000-0003-4930-4673

Andrea L. Berez-Kroeker

andrea.berez@hawaii.edu

 orcid.org/0000-0001-8782-515X