Andrea L. Berez-Kroeker*, 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 and Anthony C. Woodbury

Reproducible research in linguistics: A position statement on data citation and attribution in our field

https://doi.org/10.1515/ling-2017-0032

Abstract: This paper is a position statement on reproducible research in linguistics, including data citation and attribution, that represents the collective views of some 41 colleagues. Reproducibility can play a key role in increasing

***Corresponding author: Andrea L. Berez-Kroeker,** Department of Linguistics, University of Hawai'i at Mānoa, 1890 East West Road, Moore 569, Honolulu, HI 96822, USA, E-mail: andrea.berez@hawaii.edu
Lauren Gawne, Department of Languages and Linguistics, SOAS University of London, London WC1H 0XG, UK; La Trobe University, Melbourne, VIC 3086, Australia, E-mail: l.gawne@latrobe.edu.au
Susan Smythe Kung, Archive of the Indigenous Languages of Latin America, University of Texas at Austin, Austin, TX 78712, USA, E-mail: skung@austin.utexas.edu
Barbara F. Kelly, Department of Languages and Linguistics, The University of Melbourne, Parkville, VIC 3010, Australia, E-mail: b.kelly@unimelb.edu.au
Tyler Heston, Payap University, Chiang Mai 50000, Thailand, E-mail: tylerheston@earthlink.net
Gary Holton, Department of Linguistics, University of Hawai'i at Mānoa, 1890 East West Road, Moore 569, Honolulu, HI 96822, USA, E-mail: holton@hawaii.edu
Peter Pulsifer, National Snow and Ice Data Center, Boulder, CO 80303, USA, E-mail: pulsifer@nsidc.org
David I. Beaver, Department of Linguistics, University of Texas at Austin, Austin, TX 78712, USA, E-mail: dib@utexas.edu
Shobhana Chelliah, Department of Linguistics, University of North Texas, Denton, TX 76203, USA, E-mail: Shobhana.Chelliah@unt.edu
Stanley Dubinsky, Linguistics Program, University of South Carolina, Columbia, SC 29208, USA, E-mail: DUBINSK@mailbox.sc.edu
Richard P. Meier, Department of Linguistics, University of Texas at Austin, Austin, TX 78712, USA, E-mail: rmeier@austin.utexas.edu
Nick Thieberger, Department of Languages and Linguistics, The University of Melbourne, Parkville, VIC 3010, Australia, E-mail: thien@unimelb.edu.au
Keren Rice, Department of Linguistics, University of Toronto, Toronto, ON M5S, Canada, E-mail: rice@chass.utoronto.ca
Anthony C. Woodbury, Department of Linguistics, University of Texas at Austin, Austin, TX 78712, USA, E-mail: woodbury@austin.utexas.edu

verification and accountability in linguistic research, and is a hallmark of social science research that is currently under-represented in our field. We believe that we need to take time as a discipline to clearly articulate our expectations for how linguistic data are managed, cited, and maintained for long-term access.

Keywords: reproducibility, attribution, data citation

1 Introduction

The notion of *reproducible research* has received considerable attention in recent years from physical scientists, life scientists, social and behavioral scientists, and computational scientists. In this statement we consider reproducibility as it applies to linguistic scientists, especially with regard to facilitating a culture of proper long-term care and citation of linguistic data sets.

This paper grows out of one effort to initiate a discipline-wide dialog around the topic of data citation and attribution in linguistics, in which some 41 linguists and data scientists convened for three workshops held between September 2015 and January 2017. Participants in these workshops addressed issues related to the proper citation of linguistic data sets, and the establishment of criteria for academic credit for the collection, preservation, curation, and sharing thereof. These workshops were supported by a grant from the National Science Foundation (*Developing standards for data citation and attribution for reproducible research in linguistics* [SMA-1447886]).[1] The 41 participants represented diverse subfields of linguistics (syntax, semantics, phonetics, phonology, sociolinguistics, typology, dialectology, language documentation and conservation, historical linguistics, computational linguistics, first and second language acquisition, signed linguistics, and language archiving). Other data scientists came from library and information science, climatology, archaeology, and the polar sciences. The group included academics from every career stage from graduate students to professors to department chairs to provosts, and they represented institutions of higher learning in North America, Europe, and Australia. These participants are:

Helene Andreassen *TROLLing, UiT The Arctic University of Norway*	Ruth Duerr *Ronin Institute*	Keren Rice *University of Toronto*
Felix Ameka *Leiden University*	Colleen Fitzgerald *National Science Foundation*	Loriene Roy *University of Texas at Austin*

(*continued*)

1 https://sites.google.com/a/hawaii.edu/data-citation/

(*continued*)

Anthony Aristar *University of Texas at Austin*	Lauren Gawne *SOAS University of London and La Trobe University*	Mandana Seyfeddinipur *SOAS University of London*
Helen Aristar-Dry *University of Texas at Austin*	Jaime Perez Gonzalez *University of Texas at Austin*	Gary F. Simons *SIL International*
David Beaver *University of Texas at Austin*	Ryan Henke *University of Hawai'i at Mānoa*	Maho Takahashi *University of Hawai'i at Mānoa*
Andrea L. Berez-Kroeker *University of Hawai'i at Mānoa*	Gary Holton *University of Hawai'i at Mānoa*	Nick Thieberger *University of Melbourne*
Hans Boas *University of Texas at Austin*	Kavon Hooshiar *University of Hawai'i at Mānoa*	Sarah G. Thomason *University of Michigan*
David Carlson *World Climate Research Programme*	Tyler Kendall *University of Oregon*	Paul Trilsbeek *The Language Archive, Max Planck Institute for Psycholinguistics*
Brian Carpenter *American Philosophical Society*	Susan Smythe Kung *University of Texas at Austin*	Mark Turin, *University of British Columbia*
Shobhana Chelliah *University of North Texas*	Julie Ann Legate *University of Pennsylvania*	Laura Welcher, *Long Now Foundation*
Tanya E. Clement *University of Texas at Austin*	Bradley McDonnell *University of Hawai'i at Mānoa*	Nick Williams *University of Colorado Boulder*
Lauren Collister *University of Pittsburgh*	Richard P. Meier *University of Texas at Austin*	Margaret Winters *Wayne State University*
Meagan Dailey *University of Hawai'i at Mānoa*	Geoffrey S. Nathan *Wayne State University*	Anthony C. Woodbury *University of Texas at Austin*
Stanley Dubinsky *University of South Carolina*	Peter Pulsifer *National Sea and Ice Data Center*	

The position described here is an outcome of these meetings, and represents the collective opinion of the participants. In Section 2, we discuss reproducible research in science generally, and in linguistics in particular. In Section 3, we review some recent findings about current practices by authors of linguistics publications with regard to transparency about data sources and research methodologies. Section 4 is our summary position statement on the importance of linguistics data and the citation thereof; the need for mechanisms for evaluating "data work" in academic hiring, tenure, and promotion processes;

and the need to engender broad sociological shift in our field with regard to reproducible research through education, outreach, and policy development. Section 5 contains summary recommendations on actions that can be taken by linguistics researchers, departments, committees, and publishers, as well as some concluding remarks.

2 On valuing reproducibility in science and linguistics

Reproducible research aims to provide scientific accountability by facilitating access for other researchers to the data upon which research conclusions are based. The term, and its value as a principle of scientific rigor, has arisen primarily in computer science (e.g., Buckheit and Donoho 1995; de Leeuw 2001; Donoho 2010), where easy access to data and code allows other researchers to verify and refute putative claims. In a 2009 post on *The open science project*, a blog dedicated to open source tools and research, Dan Gezelter summarizes reproducible research thus:

> If a scientist makes a claim that a skeptic 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 skeptics, it isn't Science. (Gezelter 2009; emphasis original)

Reproducibility in research is an evolution of *replicability*, a long-standing tenet of the scientific method with which most readers are likely to already be familiar. Replicable research methods are those that can be recreated elsewhere by other scientists, leading to new data; sound scientific claims are those that can be confirmed by the new data in a replicated study.

The difference between reproducible research and replicable research is that the latter produces new data, which can then ostensibly be analyzed for either confirmation or disconfirmation of previous results; the former provides access to the original data for independent analysis. The benefit of reproducibility is evident in cases where faithfully recreating the research

conditions is impossible. For example, if a researcher conducts scientific research studying the bacteria in human navels by surveying sixty people at random, that study is considered replicable because another researcher could make the same (or different) claims based on new data coming from a survey of sixty other randomly selected human navels (Hulcr et al. 2012). But in many fieldwork-based life and social sciences, true replicability is not possible to achieve. The variables contributing to a particular instance of field observation are too hard to control in many cases – for example, 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 replicate every environmental or non-environmental factor that may contribute to which tool a chimpanzee will select in a given situation. Thus reproducibility is a potentially useful metric for rigor in scientific investigations that take place outside of a fully controllable setting.

Because linguistics can be considered a social science dealing with observations of complex behavior, it is another field that would seem to lend itself to the kind of scientific rigor that reproducibility provides; however, we are not aware of any substantial discipline-wide discussion of how we might implement reproducibility, nor of any widespread identification of a need to do so. Like the example of the frog-eating bats, 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 semi-controlled setting for linguistic observation – researchers cannot conceivably control for every possible variable, such as the previous experience of the individual, that leads to an utterance or judgment.

These natural limitations to our research methods are well accepted and noncontroversial, but they do not relieve us of the obligation of scientific accountability. The discussion of reproducibility has had serious professional consequences in other fields; consider for example the recent controversy in social psychology, in which a prominent researcher was found to have fabricated data in 15–20 years' worth of publications (Crocker and Cooper 2012). In addition, Fang and colleagues (2013) surveyed more than 2000 biomedical and life sciences journals and found that while 21.3% of 2,047 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 lead to discussions of solutions including a

"transparency index" (Marcus and Oransky 2012) and "retraction index" for journals[2] (Fang and Casadevall 2011), as well as the publication of watchdog websites,[3] indices, and blogs.[4]

Within linguistics, much of the investigation into possibilities for reproducible research has been in the context of language documentation and description, in which documentary fieldwork methods have been noted for their potential to provide substantiation of scientific claims by promoting attention to the structuring and sharing of language data. Himmelmann's 1998 position paper on language documentation is clear on this point: "[Language] documentation [...] will ensure that the collection and presentation of primary data receive the theoretical and practical attention they deserve" (1998: 164; see also Himmelmann 2006; Woodbury 2003; Woodbury 2011; Thieberger 2009; Thieberger and Berez 2012, among others).[5] Digital multimedia and annotations including transcripts and translations allow readers to confirm claims about language structure through direct access to the original observational data. This would mean that not only could example sentences in a grammar be linked to what is transcribed, parsed, and translated, but a reader could also determine whether or not she would reach the same conclusions about what those examples illustrate by having access to the utterances 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 we believe it is reasonable to make them reproducible. A creative rewording of the Gezelter quote above makes this clear:

> If a linguist makes a claim that a skeptic 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

2 A retraction index is a metric that tracks the number of articles that are retracted in a particular journal, and a transparency index aims to generally provide a more transparent account of how an individual journal operates. Marcus and Oransky (2012) suggest a number of factors that could be included in a potential transparency index, including review process, review times, manuscript acceptance rate, journal requirement for underlying data to be made available, journal costs for authors and readers, misconduct process, and retraction process.

3 e.g. http://retractionwatch.com/

4 e.g. http://reproducibleresearch.net/blog/

5 We note that our position paper is being published exactly two decades after Himmelmann (1998), and the collection and preservation of primary data *still* have not broadly received the theoretical and practical attention they deserve.

> *important*, but unless enough details of the utterances in context are made available so that it can be subjected to true reproducibility tests by skeptics, it isn't Science. (modified from Gezelter 2009; underlined words replaced; emphasis original)

Clearly, linguists cannot expect their colleagues to replicate data collection conditions – and doing so would not necessarily lead to replicated utterances – but reproducibility is a more realistic, and thus more achievable, goal. 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 is, the implementation of a structure that allows for future growth – that digital formats are well suited to provide. Thieberger (2009), which is 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 generalized 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 tokenizers. 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 (see Gawne et al. 2017).

In Bird and Simons' 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 to the present discussion among these is citation. Bird and Simons advocate 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" (2003: 572). 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 and 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 and Simons 2003: 558).

We agree this discussion is needed. *Still* needed, in fact: even in 1994 Sally Thomason, in her capacity as editor of *Language*, called upon linguists to be vigilant in in "provid[ing] detailed information about sources of data and methodology of data collection" (Thomason 1994: 413). Her exhortation arose out of the discovery that a surprising number of papers submitted to *Language* contained erroneous data ("[w]hen I began my term as editor, I expected that there would be cases of this kind from time to time. I did not expect that these cases would occur frequently – so frequently, in fact, that the assumption that the data in accepted papers is reliable began to look questionable" [p. 409]), but the spirit of her commentary is directly relatable to problems of reproducibility (although she uses the term 'replicability'): "it is vital for all authors to ensure 'clarity and replicability of the chain of evidence' so that it will be as easy as possible for other scholars 'to evaluate the solidity of the various steps in the chain, and then to replicate and extend the work the claim is based on, if they choose to' (p.c. Mark Liberman, via email, 1993)" (Thomason 1994: 410).

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: transparency about methods of data collection and analysis, and transparency about the status of source data, including how or if data might be accessed by a reader wishing to do so.

Transparency about methods of data collection and analysis may include: a description of the conditions under which data were collected (where were data collected and for how long, what genres of speech data were collected, etc.); 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.); or information about who participated in providing the data (demographic information, language community information, etc.).

Transparency about source data may include: transparency about the nature of the data that have been used (whether published data, introspective data, corpus data, elicited data, testing data, etc.); transparency about where data can be found (in publications, in archives, in field notes in a personal collection, online, etc.), or transparency about how to locate relevant subparts of a data set (page numbers, corpus line numbers, time-codes of starting and ending points in an audio or video recording, etc.).[6]

6 Although we strongly advocate for "granular" citation to precisely recoverable source data, we also acknowledge that this must be done with sensitivity toward the privacy concerns of the people whose language is being recorded (see Chelliah and de Reuse 2011: 147–151).

3 The current state of practice

Before we can decide how to best approach building a community consensus around mechanisms for increasing transparency, let us take stock of the current state of practice. A few studies have sampled linguistics publications with regard to metrics of reproducibility. In a survey of one hundred descriptive grammars from a ten-year span between 2003 and 2012, Gawne and colleagues (2017) found that even with the benefit of years of pervasive discussion of data management methods in language documentation, very few authors in this genre make their methods or data sources explicit in their writing. In a survey of 270 articles from nine top international linguistics journals from the same time period, Berez-Kroeker and colleagues (2017a) found that scant few journal authors met any – let alone all – of the survey's metrics for basic transparency of data and methodology, including sufficient citation of numbered examples from unpublished sources, or a minimal description of methods of data collection and analysis.

These two surveys of our discipline revealed both good news and bad for the current state of reproducibility in linguistics. The bad news is that linguistics has a long way to go before we can claim to be a discipline that values reproducible research. Authors of both studies found that readers of linguistics publications are implicitly asked to make assumptions about aspects of the research process: that data are collected in an appropriate 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. It seems that 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. In short, we are in danger of being a social science asking its audience to *take our word for it*.

But these studies also revealed some very good news, which holds promise for linguistics becoming more transparent in the future. The authors found that in fact different subfields do have strengths in facets of research transparency, as represented by the publications they surveyed. Practitioners in different subfields 'do transparency' differently, and these practices could serve as models for an eventual amalgamated standard. For example, authors publishing in *Studies in Second Language Acquisition* describe research methods exceptionally well – the strong experimental focus of the journal means that a methods section is a normalized expectation. Authors in *Journal of Sociolinguists* and *The International Journal of American Linguistics* frequently provide information about research participants. Authors of phonetics papers across all journals

surveyed usually provide information about tools, hardware, and software (Berez-Kroeker et al. 2017a).[7]

Importantly, all authors in the two surveys include standard citations of published material (i.e., example sentences), 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 publications surveyed do it consistently. Our field has no such widespread expectations, recommendations, or formats for other factors those two surveys examined.

We do, however, have a few disciplinary resolutions valuing aspects of research transparency. The Linguistic Society of America's (LSA) *Resolution on Cyberinfrastructure* encourages linguists to "make the full data sets available, subject to all relevant ethical and legal concerns," and that reviewers "expect full data sets to be published [...] and expect claims to be tested against relevant publicly available datasets;[8]" the *Ethics Statement* urges linguists to "carefully cite the original sources of ideas, descriptions, and data".[9] 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.[10] The *Unified Style Sheet for Linguistics* and *the Generic Style Rules for Linguistics* at the time of writing, do not contain advice for citing or formatting references to data sets.[11]

Another area of concern is our lack of mechanisms for valuing the work that goes into data set creation, preservation, and curation as scholarly output, a need that was voiced time and again in our workshops and in public presentations. The LSA's *Resolution on Cyberinfrastructure* and the *Resolution Recognizing*

7 According to Berez-Kroeker et al. (2017a), "[d]ifferences 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 wordlists and published dictionaries. Field linguists describe the speech community and their fieldwork conditions by tradition; phoneticians have a tradition of describing equipment."

8 http://www.linguisticsociety.org/resource/resolution-cyberinfrastructure

9 http://www.linguisticsociety.org/sites/default/files/Ethics_Statement.pdf

10 A recent review (Hammarström 2015) of *Ethnologue* (Lewis et al. 2015) takes not only that publication to task for poor transparency in both citation and methodology, but also many others: '[The *Ethnologue*] is not alone in not citing the individual justification for language listing. Nearly all modern language listings for continent-sized areas produced by linguists have the same policy of not citing sources' (Hammarström 2015: 735).

11 http://www.linguisticsociety.org/sites/default/files/style-sheet_0.pdf and http://www.eva.mpg.de/lingua.pdf/GenericStyleRules.pdf

the Scholarly Merit of Language Documentation both contain language indicating that primary data, databases, and corpora should be given "weight"' and "academic credit."[12] Unfortunately most linguists do not know how to go about advocating that "data work" be given the same kind of attribution as "analysis work" in hiring, tenure and promotion cases.

In other words, valuing reproducibility is one thing; implementing seems to be quite another. Thus, against this backdrop of increased awareness of the value of data and reproducible research across the sciences; increasingly accessible technologies for managing data; increased discussion in linguistics of the value of data to analysis; and our current lack of practices, standards, and recommendations thereof; we, the authors of this statement, articulate our position below on behalf of more than forty linguists who participated in our workshops since 2015.

4 Our position

4.1 The importance of linguistic data and data citation

Linguistic data are the very building blocks of our field. Given that linguistic theories need to be borne out through data, we believe that linguistic data are important resources in their own right and represent valuable assets for the field. Therefore, our field needs to accept responsibility for the proper documentation, preservation, attribution, and citation of these assets. The responsibility to do so is an integral part of linguistic research, and it should be collectively shared by individual scientists and researchers, data stewards, research institutions, and funding organizations.

4.2 Implementation of standards for data citation

Data citation is important as a means to verify claims made by researchers, to provide credit to data creators, and to facilitate the discovery and long-term use of data. Thus, it is crucially important that data be properly and regularly cited; however, as studies have indicated, practitioners in our field largely do not know how or when to cite data. As a first step towards moving our field toward

12 https://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation

uniform standards for citing data, we advocate the adoption of a set of general data citation guidelines like the FORCE11 Principles of Data Citation (Data Citation Synthesis Group 2014), modified for linguistics data – the *Austin principles of data citation in linguistics* (Berez-Kroeker et al. 2017b),[13] currently under development, is one such set of guidelines. As a second step, we urge journal editors and publishers to build on these guidelines and develop specific data citation formats for their publications. Appropriate citation formats would take into consideration the dynamicity of data sets, the need for suitable granularity, and the need to extend attribution for contributors in many roles (e.g., speakers, data inputters, annotators, technicians).

In order for data to be citable, it should be stored in an accessible location, preferably a data archive or some other repository with a demonstrated commitment to both preserving the data and making them accessible for many years to come. Archived data should be made "as open as possible, and as closed as necessary" (European Commission 2016), with consideration of ethical and legal exceptions; though restricted data is citable (and should be cited by those who are permitted to access it), only unrestricted data can be re-used and cited, thereby furthering the scientific principles of reproducibility and transparency. Data should be available in formats that do not require proprietary software and that will be usable or portable as technology changes. Data should have sufficient human- and machine-readable documentation (metadata) to allow for future informed re-use. Finally, archived data should have a persistent identifier or locator that can be included in the citation so that future readers who see the citation will also be able to find the cited data.

Some linguistic data are not easily storable or citable – i.e. some kinds of introspective data – and in these cases, authors should include in their publications an explicit statement that data come from introspection, either their own or that of others.

4.3 Academic attribution for creating, curating, preserving, and storing linguistic data

Data work is intellectually valuable, and linguists who do data work need to be recognized for their work during professional evaluation. Though linguists have long recognized the value in *using* data for linguistics analyses, the field of linguistics has been generally uneasy about valuing data in and of themselves. However, we argue that the very act of planning the collection of particular data

13 http://site.uit.no/linguisticsdatacitation/austinprinciples/

types as well as organizing data into useful and re-usable data sets requires some degree of analysis. Creating a data set is an intellectual undertaking: "just as analyzing data requires research, so does working with the data itself" (Good 2011: 233). We identify two main domains in which the academic merit of creating, curating, preserving, and storing linguistic data can be valorized: research funding, and professional evaluation (i.e., hiring, tenure, and promotion).

The first domain is funding, and particularly the expectations of the funding bodies who provide the resources for linguistics research. As researchers we can make it clear in our funding applications that data management has costs, in terms of both money and time, and that these outputs have an important and ongoing function. Funders and universities are also becoming more attentive to the management of data. Many funding bodies now require a data management plan for research projects. There is also a move towards ensuring that data from publically funded projects is made publically accessible. Major research funding bodies the United Kingdom, The Netherlands, and the United States have moved towards ensuring research articles from public funding are open access.[14] It is reasonable to assume that this practice will spread to other funding organizations.

The second domain is professional evaluation. The lack of clear guidelines and metrics for evaluating data creation, curation, sharing, and re-use in hiring, tenure, and promotion decisions has been brought up time and again in our workshop discussions. We encourage practitioners in all subfields of linguistics to develop explicit, written disciplinary standards to increase academic attribution for the important work that goes into the creation and guardianship of linguistic data.[15] Formal resolutions on the valuation of data creation from professional societies may help encourage the normalization of this practice, but it is the efforts of scholars at the level of the academic institution where hiring and promotion decisions are made that will drive the change. We support efforts to educate applicants for employment, tenure, and promotion in methods for explaining the value of linguistic data to the people who would evaluate them. We also support efforts to educate colleagues serving on hiring, tenure, and promotion committees, and those serving as department, college, and

14 For the UK, see the Research Councils UK (RCUK) Policy on Open Access: http://www.rcuk.ac.uk/research/openaccess/policy/. For The Netherlands, see open access.nl: http://www.openaccess.nl/en. For the U.S., see the National Science Foundation plan for public access to results from funded research: https://www.nsf.gov/news/special_reports/public_access/.

15 For arguments in favor of the valuation of language documentation data and collections, see Haspelmath and Michaelis (2014) and Thieberger et al. (2016).

university administrators on the value of linguistic data to the future of the discipline.

4.4 Promoting a culture shift in linguistics through education, outreach, and policy development

We support promoting a culture of data citation inside linguistics through the education of our colleagues and students. We support the development of instructional modules in the proper handling of data, at every educational level, from undergraduate to graduate to mid-career faculty. We support outreach efforts through data-oriented workshops, symposia, and other events at professional conferences and institutes. We support the broad development and sharing of training materials, handbooks, massive online open courses, and webinars in data creation, management, citation, and preservation.

A culture of citing and properly attributing data is sweeping the sciences, as can be witnessed through the formation of groups like the Research Data Alliance,[16] FORCE11,[17] the Center for Open Science,[18] and others. Linguistics should be represented in these groups so that we may participate in the broader discussion, and influence that discussion in ways that reflect the unique needs of our field.[19]

We believe that linguistics journal editors and publishers can play a key role in leading the field toward a more reproducible linguistics by developing data policies for their journals that require authors to describe their methods and cite all data. Editors can also guide authors to appropriate repositories for their data sets.

5 Summary recommendations and conclusion

In this paper, we have introduced the idea of reproducibility in linguistic research, and we have argued that scientific reproducibility is not possible without proper data citation and attribution. Furthermore, we have posited that more value should be placed on the linguistic data that underlie all

16 http://www.rd-alliance.org/

17 http://www.force11.org/

18 http://cos.io/

19 For example, the incipient Linguistics Data Interest Group of the Research Data Alliance, which may be joined free of charge by anyone who agrees to uphold the values of the Research Data Alliance: http://www.rd-alliance.org/groups/linguistics-data-interest-group.

linguistic theory and form the foundations of our discipline. Data collectors should receive appropriate attribution for their work, especially when their data are accessible, re-usable, and citable. We offer the following summary recommendations that can be adopted by practitioners at various levels.

First, it is imperative for our field that individual linguistic researchers educate themselves, their colleagues, and their students in broad principles of digital data management, so that their data are more easily shared, preserved, cited, and reused. We recommend that individual linguists proactively seek to develop a relationship with a data archive.

Second, we realize that some linguists may be reluctant to share data for personal (as opposed to ethical) reasons, and such an attitude is hardly surprising given that data sharing may not previously been standard practice in the subfields many of us work in. We can only encourage such researchers to carefully evaluate the reasons for their reticence in the light of the discussion in this paper, to potentially reconsider whether their concerns are valid, and to bring any concerns into public light so that future policies and public debate on data sharing issues can take them into account.

Third, our recommendation for departments and committees is that they develop and share concrete, written guidelines for evaluating "data work" including data management, curation, annotation, citation, and sharing. Such written guidelines can play a crucial role in hiring, tenure, and promotion cases, both for internal use among colleagues in linguistics departments, programs, and research centers, and for sharing with university-level personnel committees.

As a final recommendation, we encourage editors and publishers of linguistics journals and book series to develop concrete policies for both data sharing and data citation, and to develop formats for the citation of linguistic data sets. These are non-trivial tasks, but sharing of ideas in the editorial community can lessen the burden on individual editorial teams and publishers, and is essential for the long-term development of publishing standards for data.

The discussion about the role of data in linguistic analysis is ongoing, and there are still many positive changes that linguistics as a discipline can make to ensure that research is reproducible. In 2014 the National Science Foundation called for proposals to develop standards for data citation and attribution: "NSF seeks to explore new norms and practices in the research community for [...] data citation and attribution, so that data producers, [...] and data curators are credited for their contributions" (National Science Foundation 2014). This NSF-wide effort is evidence that transparency in service of reproducible science is valued widely in the sciences, yet like many fields, linguistics is in need of standards that put those values into practice. We encourage readers to see this

moment as an opportunity to draw together increasingly common expectations regarding linguistic data and make those expectations explicit.

Acknowledgements: We are grateful to participants in the three workshops on developing standards for data citation and attribution for reproducible research in linguistics for their discussion on the ideas presented herein. Particular thanks also to Maho Takahashi, Meagan Dailey, Ryan Henke, Kavon Hooshiar, and Jaime Perez Gonzalez for their assistance. Portions of Section 2 appeared in Berez (2015). Errors of omission and commission belong to the authors alone.

Funding: This research was funded in part by the National Science Foundation SMA-1447886.

References

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 Amanda Harris, Nicholas Thieberger & Linda Barwick (eds.), *Research, records and responsibility*, 39–51. Sydney: University of Sydney Press.

Berez-Kroeker, Andrea L., Lauren Gawne, Barbara F. Kelly & Tyler Heston. 2017a. A survey of current reproducibility practices in linguistics journals, 2003–2012. https://sites.google.com/a/hawaii.edu/data-citation/survey (accessed 11 August 2017).

Berez-Kroeker, Andrea L., Helene N. Andreassen, Lauren Gawne, Gary Holton, Susan Smythe Kung, Peter Pulsifer, Lauren B. Collister, The Data Citation and Attribution in Linguistics Group, & The Linguistics Data Interest Group. 2017b. The Austin principles of data citation in linguistics (Version 0.1). http://site.uit.no/linguisticsdatacitation/austinprinciples/ (accessed 27 November 2017).

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79. 557–582.

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.

Chelliah, Shobhana L. & Willem J. de Reuse. 2011. *Handbook of descriptive linguistic fieldwork*. London: Springer.

Crocker, Jennifer & M. Lynne Cooper. 2012. Addressing scientific fraud. *Science* 334. 1182.

Data Citation Synthesis Group. 2014. *Joint declaration of data citation principles*, edited by M. Martone. San Diego: FORCE11. https://www.force11.org/group/joint-declaration-data-citation-principles-final (accessed 9 August 2017).

de Leeuw, Jan. 2001. Reproducible research: The bottom line. *UCLA Department of Statistics papers*. http://escholarship.org/uc/item/9050x4r4 (accessed 15 March 2014).

Donoho, David L. 2010. An invitation to reproducible computational research. *Biostatistics* 11. 385–388.

European Commission. 2016. Guidelines on FAIR data management in Horizon 2020. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf. (accessed 09 August 2017).

Fang, Ferric C. & Arturo Casadevall. 2011. Retracted science and the retraction index. *Infection and Immunity* 79. 3855–3859.

Fang, Ferric C., R. Grant Steen & Arturo Casadevall. 2013. Misconduct accounts for the majority of retracted scientific publications. *PNAS Early Edition* 334. 1–6.

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.

Gezelter, Dan. 2009. Being scientific: Falsifiability, verifiability, empirical tests, and reproducibility. *The OpenScience project*. http://www.openscience.org/blog/?p=312 (accessed 5 July 2015).

Good, Jeff. 2011. Data and language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages*, 212–234. Cambridge: Cambridge University Press.

Hammarström, Harald. 2015. Ethnologue 16/17/18th editions: A comprehensive review. *Language* 91. 723–737.

Haspelmath, Martin & Susanne Maria Michaelis. 2014. Annotated corpora of small languages as refereed publications: A vision. *Diversity linguistics comment*. http://dlc.hypotheses.org/691 (accessed 10 January 2017).

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin & New York: Mouton de Gruyter.

Hulcr, Jiri, Andrew M. Latimer, Jessica B. Henley, Nina R. Rountree, Noah Fierer, Andrea Lucky, Margaret D. Lowman & Robert R. Dunn. 2012. A jungle in there: Bacteria in belly buttons are highly diverse, but predictable. *PlosOne* 7. e47712. http://www.ncbi.nlm.nih.gov/pubmed/23144827 (accessed 9 August 2017).

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2015. *Ethnologue: Languages of the world*. Dallas, TX: SIL International.

Marcus, Adam & Ivan Oransky. 2012. Bring on the transparency index. *The Scientist Magazine*. http://tiny.cc/2012-transp-marcus.

Maxwell, Mike. 2012. Electronic grammars and reproducible research. In Sebastian Nordhoff (ed.), *Electronic grammaticography* (Language Documentation & Conservation Special Publication No. 4), 207–234. Honolulu: University of Hawai'i Press.

National Science Foundation. 2014. Supporting scientific discovery through norms and practices for software and data citation and attribution (Dear Colleague letter). http://www.nsf.gov/pubs/2014/nsf14059/nsf14059.jsp?org=NSF. (accessed 11 November 2014).

Ryan, Michael J. 2011. Replication in field biology: The case of the frog-eating bat. *Science* 334. 1229–1230.

Thieberger, Nicholas. 2006. *A grammar of South Efate: An Oceanic language of Vanuatu*. Honolulu: University of Hawaii Press.

Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data. In Patricia Epps & Alexandre Arkhipov (eds.), *New challenges in typology: Transcending the borders and refining the distinctions*, 389–408. Berlin & New York: Mouton de Gruyter.

Thieberger, Nicholas & Andrea L. Berez. 2012. Linguistic data management. In Nicholas Thieberger (ed.), *The Oxford handbook of linguistic fieldwork*, 90–118. Oxford: Oxford University Press.

Thieberger, Nick, Anna Margetts, Stephen Morey & Simon Musgrave. 2016. Assessing annotated corpora as research output. *Australian Journal of Linguistics* 36. 1–21.
Thomason, Sarah. 1994. The editor's department. *Language* 70. 409–423.
Tomasello, Michael & Josep Call. 2011. Methodological challenges in the study of primate cognition. *Science* 334. 1227–1228.
Woodbury, Anthony C. 2003. Defining documentary linguistics. *Language Documentation & Description* 1. 35–51.
Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *Cambridge handbook of endangered languages*, 159–186. Cambridge: Cambridge University Press.