The notion of reproducible research has received considerable attention in recent years from physical scientists, life scientists, social and behavioral scientists, and computational scientists. 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, where easy access to data and code allows other researchers to verify and refute putative claims.

Reproducibility in research is an evolution of replicability, a long-standing tenet of the scientific method. Replicable research methods are those that can be recreated elsewhere by other researchers, 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, as is the case in many behavioral and fieldwork-based sciences. Linguistics, as a data-driven social science in which inferences about human cognition and social structure are drawn from observations of behavior, is well positioned to benefit from principles of reproducible research. Subsequent researchers are unlikely to have access to the same speakers or speech situations from which the original data set was collected. However, given access to the original data set referenced in a publication, subsequent researchers can attempt to reproduce the original research. Crucially, reproducibility requires access to underlying data sets, not merely the examples contained in the body of a scholarly paper. This means that data need to be prepared, preserved, and cited in a manner that allows discovery of the entire set, as well as an individual datum within it.

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 methodology has been noted for its potential to provide substantiation of scientific claims by promoting attention to the care and structuring of language data. However, linguistics more broadly has not generally fostered a culture of creating and preserving data sets, citing underlying data in publications, or providing academic rewards to people who do so. In a study of data citation practices in academic linguistic works from multiple linguistic subfields spanning a ten-year period, we have found that by and large authors of journal articles, grammars, and dissertations rarely indicate if or where data is stored or shared, let alone provide a citation indicating from where in a data set a particular datum was retrieved. Where citations are provided, the connection to the data set may be only vaguely identified. For example, an excerpt might be given a citation which refers to the name of the text from which it was extracted, but in practice the reader has no way to access that text. Additionally, while the LSA’s resolutions on Cyberinfrastructure and Recognizing the Scholarly Merit of Language Documentation support attribution of academic credit for the preparation and preservation of data...
sets in hiring, tenure and promotion decisions, they stop short of providing standards for doing so. Thus, in spite of the potential generated by recent shifts in the field, a great deal of linguistic research created today is not reproducible, either in principle or in practice.

In order to facilitate the development of reproducible research in linguistics we propose the discipline-wide adoption of common standards for data citation and attribution. In our parlance citation refers to the practice of identifying the source of linguistic data, and attribution refers to mechanisms for assessing the intellectual and academic value of data citations. Citation and attribution are key to reproducible research, as the former ties linguistic examples to the data sets from which they were extracted, providing skeptical readers the opportunity to test hypotheses against those data sets; and the latter incentivizes reproducible practice by ensuring that the work of preparing, archiving, and making available linguistic data sets is properly credited.

This session grows out of our current NSF project, funded under grant number SMA-1447886, to develop standards for data citation and attribution in linguistics at the grassroots level, with the ultimate aim of creating a discipline-wide culture of reproducible research. Session participants have all been directly involved in the project, which began in 2015. The session will bring the conversation to the broader membership of the LSA for the first time, and will highlight both issues of wider concern and more detailed topics. Please note that discussion is our primary objective, and the panel discussion format as well as the poster session will encourage participation from the LSA membership more generally.

Abstracts (Papers)

**Andrea Berez-Kroeker** (University of Hawai‘i at Mānoa)
**Gary Holton** (University of Hawai‘i at Mānoa)
**Susan Smythe Kung** (University of Texas at Austin)
**Peter Pulsifer** (University of Colorado at Boulder)

*Reproducible research in linguistics: toward a data-driven science of language*

Data are fundamental to the field of linguistics. Examples drawn from natural languages provide a foundation for claims about the nature of human language, and validation of these linguistic claims relies crucially on these supporting data. Yet, while linguists have always relied on language data, they have not always facilitated access to those data. This presentation provides an introduction to the notion of reproducible research and its application to the field of linguistics. Inspired by parallel efforts in other disciplines and drawing on recent technological advances, we outline a multi-pronged strategy to develop standards for data citation and attribution in linguistics. Ultimately we envision a science of linguistics grounded in reproducible research—a science where claims about linguistic structure and practice can be readily validated. Such a data-driven linguistic science has the potential to provide substantiation of scientific claims by promoting attention to the care and structuring of language data.

**Anthony Woodbury** (University of Texas at Austin)

*Data citation: broad principles and guidelines*

Linguistic data are important resources in their own right. For a linguist to be able to work with their own records or to use records created by others, there need to be established citation standards that can be adopted easily by all relevant members of the research community. Citation implies a preservation strategy, for whatever is cited needs to be in a location to allow the curious scholar to access it. The work of the original scholar needs to be attributed properly, and citation formats have to allow for that attribution. Therefore, data should be made open as soon as possible, with consideration to ethical exceptions. They should be in usable formats, with sufficient machine and human readable documentation to allow informed re-use. These responsibilities are an integral part of linguistic research.

**Keren Rice** (University of Toronto)

*Data collections: what is the intellectual value?*

There is a tradition of valuing datasets in linguistics. At the same time, the field often has an uneasy relationship with data regarded as raw. The importance of data has grown exponentially in this digital era – consider documentation, the LSA resolutions on cyber-infrastructure and recognizing the scholarly merit of documentation, and the Digging into Data initiative addressing theoretical issues through large-scale digital data. ‘Big’ data opens questions about the scholarly merit of data collection and preparation – the research before one gets to what is often considered ‘real’ research. I focus on documentation
data, examining what characterizes a good collection and what must happen for its scholarly merit to be recognized. I review literature on qualities of a good collection, criteria by which its intellectual merit can be evaluated, and ways to help job candidates and tenure and promotion committees evaluate the scholarly value of this work.

**David Beaver** (University of Texas at Austin)
**Stanley Dubinsky** (University of South Carolina)

*The role of the journal in linguistic data citation and attribution*

Academic journals are the main medium for transmitting empirical discoveries and theoretical innovations from those who generate them to those who consume them, and to the extent that linguistic data needs to be discoverable, retrievable, citable, and properly attributed, it is they who must insure that this information is presented in a legible fashion, and that there are efficient and coherent standards for doing so. As management, curation, and citation of linguistic data becomes increasingly important to the conduct and publication of linguistic research, journal editors and publishers have an increasingly important role to play. Some (but certainly not all) of the issues needing discussion are (i) authorship, (ii) access, (iii) attribution, and (iv) the dynamicity of data. Our presentation will spell out issues needing to be addressed, suggest some means of addressing them, and hopefully (and more importantly) stimulate a productive discussion, bringing others’ excellent ideas into the conversation.

**Shobhana Chelliah** (University of North Texas)

*Outreach and education on data management to effect a culture shift in linguistics*

Consistent data attribution and citation practices will require a culture shift in the discipline. This shift can be set in motion through enhancements to existing linguistics curricula and through outreach to the disciple and beyond. This presentation will provide examples of how modules on data management, attribution, and citation can be added to existing linguistics courses in order to bring changes to citation practices rapidly and consistently. The presentation will include a proposed set of topics for this curriculum. Additionally, we explore how modules on digital language resources can be incorporated into class research projects in disciplines across campus and discuss how to utilize untapped sources of dissemination such as a university’s digital library, mobile technologies and MOOCs.

**Ruth Duerr** (University of Illinois at Urbana-Champaign/Ronin Institute)

*Data citation in the sciences*

While many domain repositories in the sciences have promoted citations for their data holdings for decades, it has only been in the last few years that the research community has begun to come to grips with the issues that arise when the sources used and created as a part of the research process are neither cited nor available. Spurred on by highly publicized cases of fraud and other errors, recognition of publication biases, and given a variety of political pressures, we are beginning to recognize the need to redefine the norms of the scientific process, a process based on transparency and reproducibility. These norms require that all sources used and created as part of the research process, be they data, code or publications, be cited and that the underlying data and code be available. Yet, defining exactly what this means is something each discipline needs to address for its community.

**Richard Meier** (University of Texas at Austin)

*Facilitated discussion*

The short presentations described above will be followed by 40 minutes of facilitated discussion. The aim of the discussion is to understand the views and concerns of the LSA membership about developing standards for data citation and attribution so that we can move forward in a more inclusive manner.
Abstracts (Posters)

Andrea Berez-Kroeker (University of Hawaiʻi at Mānoa)
Gary Holton (University of Hawaiʻi at Mānoa)
Susan Smythe Kung (University of Texas at Austin)
Peter Pulisifer (University of Colorado at Boulder)

*Developing standards for data citation and attribution for reproducible research in linguistics: project summary and next steps*

Developing Standards for Data Citation and Attribution for Reproducible Research in Linguistics is an NSF-supported project (SMA-1447886) that brings together relevant stakeholders to collaboratively develop and promote standards for linguistic data citation and attribution. Project participants include linguistics journal editors; language archivists; linguists representing various subfields and academic career stages from graduate students to provosts; and “Big Data” specialists. The first workshop was held at the University of Colorado Boulder in September 2015; the second was held at The University of Texas at Austin in April 2016. A panel presentation and the final workshop will take place in conjunction with this LSA Annual Meeting. This poster summarizes the aims and accomplishments of the first two workshops, describes the plans for the final workshop and the suggests the next steps in the development and promotion of a model for data citation and attribution in linguistics.

Ryan Henke (University of Hawaiʻi at Mānoa)
Meagan Dailey (University of Hawaiʻi at Mānoa)
Kavon Hooshiar (University of Hawaiʻi at Mānoa)

*Questions, curiosities, and concerns: talking points for data citation and attribution*

Changing the way linguists approach data citation and attribution means changing the way we traditionally think about data, their role in research, and their scholarly value. We are accustomed to valuing only particular academic products, even though we invest just as much time, effort, and analytical skill into collecting and managing the data behind published products. Efforts to redress this status quo proceed on a variety of fronts, but none of these efforts will happen overnight. Furthermore, not all of these conversations will be smooth conversions of viewpoints and philosophies. Changing minds takes time and patience, especially when navigating decades of thought and practice. This poster presents some questions, curiosities, and concerns that have been raised during conversations about changing standards and practices for data citation and attribution. Following the tradition in public relations and politics, we offer talking points for conveying a helpful and hopeful message to colleagues.

Lauren Gawne (School of Oriental and African Studies)
Barbara Kelly (University of Melbourne)
Andrea L. Berez-Kroeker (University of Hawaiʻi at Mānoa)
Tyler Heston (University of Hawaiʻi at Mānoa)

*A survey of current reproducibility practices in linguistics publications*

In order to move forward toward reproducible research in linguistics, we first need to know where we are now with regard to our practices for methodological clarity and data citation in publications. In this poster we share the results of a study of over 370 journal articles, dissertations, and grammars, which is taken as a sample of current practices in the field. The publications all come from a ten-year span. The journals were selected for broad coverage. Grammars included published grammars and dissertations written as grammars, with broad geographic coverage, both in terms of subject language and publisher or university. These publications are critiqued on the basis of transparency of data source, data collection methods, analysis, and storage. While we find examples of transparent reporting, most of the surveyed research does not include key metadata, methodological information, or citations that are resolvable to the data on which the analyses are based.

Susan Smythe Kung (University of Texas at Austin)
Jessica Trelogan (University of Texas at Austin)

*The data management life cycle for linguists*

While the concept of managing data is not new to the field of linguistics, the reality is that there are still significant barriers to creating citable records that allow persistent access to clearly structured primary data and that enable reproducible results. With
the current emphasis from funding agencies and publishers on the importance of transparency and data sharing, it is increasingly important that good data management skills and methods be prioritized as a formal part of the academic linguist’s workflow. Although it is difficult to provide specific recommendations for all subdisciplines of such a heterogeneous field, this poster will highlight core principles and provide specific guidelines for academic linguists throughout the research life cycle, from the earliest pre-planning stages, through deposit in a trusted repository, and into the future as data are re-used for further inquiry.

Susan Smythe Kung (University of Texas at Austin)
Jaime Perez Gonzalez (University of Texas at Austin)

Citation and attribution of archived data: guidelines of the Archive of the Indigenous Languages of Latin America

Today great quantities of research data are publically available for re-use, and most academic fields are becoming aware of the need to establish recommendations for how these data should be cited so that the data creators get proper attribution for their work. To this end, AILLA has developed Citation Guidelines that provide detailed citation examples of the different hierarchical levels of AILLA's holdings, including collections (organized materials based on individual collectors), resources (materials organized around a speech event), and individual files. These Guidelines differentiate in-text and bibliographic citations. Furthermore, each collection and resource page on AILLA provides instructions for how it should be cited. In this poster, we explain AILLA's Citation Guidelines, we show how--when followed--these guidelines give appropriate credit to the various contributors of the data and allow for easy access to the data in the archive, and we demonstrate the proper implementation of these guidelines in linguistic literature.

Lauren B. Collister (University of Pittsburgh)

Tell the story of data with metrics

With an increasing focus at many universities on engagement and impact of scholarly work both in academia and outside, knowledge of data re-use is an important consideration when sharing a dataset. Traditional citation measures for data are difficult to track because scholars rarely cite datasets formally (Kratz and Strasser, 2015). Fortunately, there are other tools available for gathering metrics about reuse of data sets that can be used by scholars. These tools are called ‘altmetrics’, and they can capture both traditional scholarly measures of use (such as citations) as well as downloads, news and popular stories that reference a dataset, blog posts, and even social media buzz about data. These metrics can be a valuable tool for scholars to show all of the impact of their work. On this poster, these metrics will be explained along with best practices for using tools to gather altmetrics.

Helene N. Andreassen (UiT The Arctic University of Norway)
Philipp Conzett (UiT The Arctic University of Norway)
Stein Høydalsvik (UiT The Arctic University of Norway)
Leif Longva (UiT The Arctic University of Norway)
Odu Obiajulu (UiT The Arctic University of Norway)

TROLLing: Scope and operation of an open repository for linguistic datasets

TROLLing (opendata.uit.no) is an international archive for open linguistic data and statistical code (e.g. R scripts), launched in 2014 at UiT The Arctic University of Norway. With the increasing demand for archiving and sharing research data, as well as the problem of improper attribution, TROLLing aims to meet researchers’ needs by proposing safe storage of data files, and metadata templates based on international standards. Retrieval, sharing, and reuse of data is further facilitated by TROLLing being part of a global open data network. As regards attribution, the system automatically provides a dataset citation, comprising among other things the author name(s) and a persistent identifier (doi). A version control allows researchers to update their datasets at any time, previously published versions still being available open access.

TROLLing is available to all subfields of linguistics, but is limited to structural data. The metadata template, however, allows linking to primary data, stored elsewhere.

Meagan Dailey (University of Hawai‘i at Mānoa)
Ryan Henke (University of Hawai‘i at Mānoa)

Data citation, attribution, and employability
Demand from academic departments for linguists possessing data skills has remained low in the last decade despite an influx of new data-driven tools, research, and ability to manage data in ways not possible before the internet. We assessed two barometers of employability: academic job postings and course descriptions. A survey of the academic linguistic job market over 10 years reveals that despite the field becoming more reliant on digital data, employers are not asking that candidates be fluent in data management. We also surveyed course descriptions and syllabi from 25 of the top-ranked linguistics programs in the United States and abroad, finding that most universities do not offer training in basic data management, despite offering courses in data-driven subdisciplines. This poster presents the data supporting these points including data management hiring and training trends.

**Kavon Hooshiar (University of Hawai‘i at Mānoa)**

*Data management across academic disciplines*

Developments in digital technologies have increased the quantity of data being created as well as provided a means to make that data available to the public digitally. Researchers are now faced with managing such grey publications, for which there is no guarantee of persistence or accessibility, nor standards for citation and attribution.

Linguists are not alone in changing the way we think about data. Initiatives such as the e-Infrastructure Reflection Group and FORCE11 have membership across the sciences and identify citation, attribution, unique identification, access, persistence, specificity, and interoperability of data as fundamental. The linguistics community could benefit from developing our understanding of data management consistently with the larger academic community, and the overlap between our guiding principles should facilitate this outcome.

This poster outlines this overlap between our efforts and those of other disciplines, and explores ways we can proceed to facilitate our interaction with the larger academic community.

**Bradley McDonnell (University of Hawai‘i at Mānoa)**

**Patrick Hall (University of California, Santa Barbara)**

*Developing methods for reproducible research in linguistics: a first step*

Reproducible research in other fields has developed various software tools that facilitate the publishing of code and results in a single document that are linked directly to the data. In mainstream linguistics, however, such software does not exist. The workflows for including linguistic examples in published work typically involve manual methods of copying and pasting text from a database into a word processing document. These manual methods are error-prone and time-consuming--often involving tedious tasks of aligning glosses in tables or with tabs. Furthermore, the examples in these documents are in no way linked to the corpus. This poster presents a first-attempt at developing a family of scripts called glossbox that link data, code, and analysis. At present, glossbox works with the typesetting software LaTeX, allowing users to semi-automatically import examples directly from the corpus. These examples require little to no manual manipulation and automatically produce citations to the corpus.