An increasingly digital science is reducing the reproducibility of research. "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 true reproducibility tests by skeptics, it isn't Science." (Gezelter 2014; emphasis original). This project aims to create guidelines for reproducible research in linguistics.

**Project Summary**

This project consists of a series of workshops and one panel presentation bringing together relevant stakeholders to develop and promote standards for data citation and attribution for linguistic data.

Linguistics is a data-driven social science, in which inferences about human cognition and social structure are drawn from observations of linguistic practices. These observations, in the form of recordings and associated annotations, represent the primary data sets that underly the field.

While linguists have always relied on language data, we have not always facilitated access to those data.

Linguistic publications typically include short excerpts from data sets, ordinarily consisting of fewer than five words and often without citation. Where citations are provided, the connection to the data set is usually only vaguely identified. An excerpt might be given with a citation referring to the name of the text from which it was extracted, but in practice the reader has no way to access that text.

That is, 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. The workshops and panel presentation will facilitate development of standards for the citation and attribution of linguistic data that are responsive to these changing conditions and shift the field of linguistics toward a more scientific, data-driven model which results in reproducible research.

A primary factor hindering the development of reproducible research in linguistics is the lack of standards for data citation and attribution. Although language data are increasingly recognized as important, there are no widely established guidelines for the citation of these data.

Equally important, there are no standards for attribution. Lacking such standards, journals, academic tenure and promotion committees, and peer review processes continue to emphasize linguistic analysis over linguistic data, and as a result linguists have little incentive to make data accessible.

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.

**What do we mean by reproducible research?**

Reproducible research aims to provide scientific accountability by facilitating access for other researchers to the data and code authors used to verify and validate published claims and results.

The notion of reproducible research has received considerable attention in recent across the sciences (e.g., Bushell & Dowris 1995, de Lusay 2011, Donoho 2014). Reproducibility is in research is an evolution of reliability, a longstanding tenet of the scientific method. Reproducibility in research means that those that can be recreated elsewhere by other scientists, leading to new data. Reproducible research provides access to the original data for independent verification and, additionally, a way to assess the reliability and credibility of the published results.

Reproducible research requires access to the raw data, and tools used to analyze those data, such as code and the software environment in which the code is executed. The quality of the data and the processes that led to the data are equally important, as data are often not easily accessible, or not available at all.

The term, and its value as a principle of scientific rigor, has arisen primarily in computer science, where easy access to raw data and code allows other researchers to verify and validate published results.

"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 reproducibility tests by skeptics, it isn't Science." (Gezelter 2014; emphasis original).

**Our Activities**

**Workshop 1: September 19-20, 2015, Boulder Colorado**

All our inaugural workshops, we heard from presenters in four Working Groups:

- The Editor Community
- The Archivist Community
- The 1/Big Data Community
- The Interested Working Linguist Community

And we discussed the following topics:

- The state-of-the-art in infrasstructural data management
- Data publication opportunities and challenges
- Current practices for data citation in linguistics journals and conferences
- Evaluation of data sets for promotion and tenure
- Data packaging and archives
- Digital humanities
- Developing the reproducible community

This workshop gave us an overview of some of the questions and problems that many of our developing communities in the linguistics data citation discussion.

**Workshop 2: April 8-10, 2016, Austin, Texas**

Participants in our second workshop joined one of four Task Forces. The task forces developed overarching principles and guidelines for data citation and attribution in linguistics:

- Principles and Guidelines Task Force
- Principles and Guidelines Task Force
- Principles and Guidelines Task Force
- Principles and Guidelines Task Force

**Project Funding**

This 2-year initial project is funded by the NSF Science of Science and Information Policy Office, under the "Supporting Scientific Discovery Through Names and Practices for Software and Data Citation and Attribution" program (SMA-1447886, 2015-2017). This program also supported similar projects in other fields.

We will be seeking additional funding at the end of the current project.

Get Involved!

We want to hear from you to tell us:

- Developing standards for citation formats
- Development of software tools for reproducible research
- Developing guidelines for academic attribution of data set creation, curation, archiving and sharing
- Education and outreach

Please contact us at lingdata@hawaii.edu

Visit our website at:

http://bit.ly/LinguisticsDataCitation

*This material is based upon work supported by the National Science Foundation under grant SMA-1447886. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.*