Questions, curiosities, and concerns

Talking points for data citation and attribution

Ryan Henke, Meagan Dailey, & Kavon Hooshiar
University of Hawaiʻi at Mānoa

Changing the way linguistics approaches data citation and attribution means changing traditional thinking about the role of data in research and their scholarly value. Here are **FAQs and talking points** for conveying a helpful and hopeful message to colleagues.

WHY IS THIS IMPORTANT?

Because **linguistics is a data-driven social science**. The questions we ask and the conclusions we draw come from the generation and investigation of information from human cognition and social structures. As a data-driven science, we have a responsibility to make our research reproducible to allow others to find and use our data in order to test our findings and build upon our work. This means we need to find ways to **make our data available and preserve it for long-term use** by future generations of researchers. In order to achieve these aims, we must facilitate the sharing and citation of data and create pathways for researchers to receive proper credit for their work.

HOW DOES SHARING MY DATA BENEFIT ME?

Making your data available and accessible allows others to see your work and recognize your contributions. Furthermore, it helps us all **build a culture of scientific openness and progress**. That's exactly what we need to create changes in our field and make sure that people start getting recognized, hired, and promoted for generating data and making data accessible and usable for others. **This thinking has the support of the LSA**, which has issued two resolutions recognizing the scholarly value of linguistic datasets. Together, they support the recognition of datasets as "scholarly contributions to be given weight in the awarding of advanced degrees and in decisions on hiring, tenure, and promotion of faculty." By sharing your data, **you become part of the movement** to ensure researchers can credit others and receive credit for their own work.

CONTACT US

Ryan Henke	*rhenke@hawaii.edu*
Meagan Dailey	*medailey@hawaii.edu*
Kavon Hooshiar	*kavon@hawaii.edu*

HOW DOES THIS RELATE TO ETHICS?

We have an ethical responsibility to cite data and share data. **Look no further than the LSA's Code of Ethics** (*linguisticsociety.org/resource/ethics*), which says linguists should "carefully cite the original sources of ideas, descriptions, and data" and that "linguists should make the results of their research available to the general public". Our discipline and the reputations of researchers depend upon academic integrity, and we have a responsibility to share the results of science with the world outside the academy.

It's not just the LSA who thinks this. You see the same sentiments coming from organizations like the International Council of Scientific Unions. In fact, other sciences are already ahead of linguistics when it comes to data citation and attribution. **We're just asking linguists to build upon what the LSA and other scientific organizations have already confirmed as ethical imperatives**. Let's incentivize linguists to make their data available for others to use, and in turn, let's create mechanisms for other linguists to use that data and make sure the original researcher receives credit for it. That is the way we make scientific progress and advance the field.

HOW CAN I KEEP FROM GETTING SCOOPED?

The short answer is that you can never scoop-proof your shared data, but the good news is that this rarely ever happens. **It's something we all fear, but this fear almost never comes true**. The even-better news: If we put in place the right kinds of data citation and attribution pathways, researchers will get more credit and greater rewards for contributing their data to enable further work by others. We foresee a change toward incentivizing more open collaboration and sharing rather than data isolationism.

HOW CAN WE ENSURE WELL-COLLECTED DATA?

This is a question raised in other sciences as well, and it involves the training any linguist should receive at any department. This goes beyond data citation and attribution, but part of what we are doing is building mechanisms to evaluate the quality of data. Right now, **we are seeing discussions and suggestions arising in the literature**. Journals are starting to publish reviews that assess the quality of archived datasets. Organizations like DELAMAN have created awards to recognize quality datasets. Researchers and funding organizations like the NSF are also figuring out ways to make digital tools to facilitate better data collection and management. If we ensure that people receive credit and are rewarded for the data they collect, this will incentivize them to share their data. **If the data are good, people will use them, and the original researcher will benefit**. Together, this will help the larger effort to collect good data in linguistics and any other science.

WHAT IF MY DATASET ISN'T READY TO SHARE?

Nobody ever feels like their data are perfect, but we can't let the perfect be the enemy of the good. There's always more refining to do. There's always a better way to present information. That's science, and we should try to create a culture of openness and moving forward together.

WHAT'S HAPPENING NEXT?

After LSA 2017 ends, we are having a workshop to create a position paper with our final recommendations for data citation and attribution in linguistics. After that, we'll submit a proposal for an LSA resolution.