Scholarly Communication in Linguistics: Resource Workshop and Poster Session

100% Virtual (Weblink TBD)
Thursday, January 6
1:30 PM – 3:15 PM

Organizers: Lauren B. Collister (University of Pittsburgh)
Andrea Berez-Kroeker (University of Hawai'i at Mānoa)

Sponsor: Committee on Scholarly Communication (CoSCIL)

Participants:

Workshop:
Lauren B. Collister (University of Pittsburgh)
Angelina McMillan Major (University of Washington)
Emily M. Bender (University of Washington)
Batya Friedman (University of Washington)
Andrea L. Berez-Kroeker (University of Hawa'i at Mānoa)
Bradley McDonnell (University of Hawa'i at Mānoa)
Eve Koller (Brigham Young University Hawai'i)
Helene Andreassen (UiT The Arctic University of Norway)
Paul De Decker (Memorial University of Newfoundland)

Poster Session:
Shirley Gabber (University of Hawai'i at Mānoa)
Susan Smythe Kung (University of Texas, Austin)
Claire Bowern (Yale University)
Valerie Fridland (University of Nevada, Reno)
Tyler Kendall (University of Oregon/National Science Foundation)
Kristine A. Hildebrandt (Southern Illinois University, Edwardsville)
Alena Witzlack-Makarevich (Hebrew University)
Johanna Nichols (University of California, Berkeley)
Taras Zakharko (University of Zurich)
Balthasar Bickel (University of Zurich)
Kathleen Currie Hall (University of British Columbia)
J. Scott Mackie (University of British Columbia)
Roger Yu-Hsiang Lo (University of British Columbia)
Lina Hou (University of California, Santa Barbara)
Ryan Lepic (Gallaudet University)
Rolando Coto-Solano (Dartmouth College)
Sally Akevai Nicholas (Massey University)
Brittany Hoback (University of Wellington)
Gregorio Tiburcio Cano (Office of the Secretary of Education, Guerrero, Mexico)
Philip Duncan (University of Kansas)
Harold Torrence (University of California, Los Angeles)
Travis Major (University of California, Los Angeles)
Jason Kandybowicz (City University of New York)
Cristina Guardiano (Università di Modena e Reggio Emilia)
Hilda Koopman (University of California, Los Angeles)

"Scholarly communication" refers to the many different ways that scholarly work is created, shared, disseminated, evaluated, and preserved. Linguistics as a discipline thrives on robust and varied forms of creating and sharing research work. In recent years, the growing importance of broad awareness of the principles of scholarly communication has become apparent for linguists at all levels of our field. Trends across the social sciences toward Open Science, Open Education, Open Access, and Open Data, as well as forays into Reproducible Research, mean

that advancing scholarly communication is an essential responsibility of the LSA to its membership. In fact, the Committee on Scholarly Communication (CoSCIL) sees education and outreach in this realm to be its main charge. Along these lines, CoSCIL has recently developed the LSA's Statement on Open Scholarship, which was approved by the Executive Committee in May 2021.

In this workshop plus poster session, sponsored by CoSCIL, we will review some of the latest developments in tools and resources for scholarly communication in linguistics from members of the LSA and the broader global community. Our goal, beyond spreading awareness of these resources, is for workshop participants to consider their own linguistic research, teaching, and creation practices with an eye towards 1.) how to apply these resources in their own work, and 2.) what gaps exist in scholarly communication in linguistics.

Five presentations will focus on resources that participants can use in their own linguistics work. For the "creation" and "preservation" parts of scholarly communication, we will cover a robust data management resource, the *Open Handbook of Linguistic Data Management*. For "sharing" and "dissemination", we will highlight the Tromsø Recommendations for Citation of Research Data in Linguistics and the broader goals of the Austin Principles for Data Citation in Linguistics; we also discuss why and how to develop data statements for communicating about data in Natural Language Processing. In the realm of "evaluation", we share the document, resources, and use cases from the LSA Statement on the Merit and Evaluation of Open Scholarship in Linguistics. Finally, as a way to unify scholarly creation, preservation, sharing, dissemination and evaluation, we present resources for developing community-based documentary media as an essential component of linguistics research.

After presentation of these resources, participants will be invited to separate conversation groups specifically about these resources, hosted by the presenters of the resources. Each presenter will prepare a hands-on demonstration of the utility of the resource, and invite questions and comments from participants in the group. In addition to introducing the tools and resources to participants and allowing space for exploration of their applications, these smaller group sessions will aim to answer a big question: *what gaps exist in our knowledge and resources for scholarly communication in the field of linguistics and its subdisciplines, and how can we fill them?*

Following the group discussions, the workshop will reconvene with each group reporting out on major topics and themes discussed, as well as identifying gaps in needs for scholarly communication needs across the discipline. Organizers will document these desiderata and submit them to the sponsoring committee for consideration and distribution. In this way, participant interaction with these current resources will invite future developments that actively address the needs of scholars in the field.

In addition to the presentations in this workshop, we will have a virtual poster session for all LSA attendees to explore. These ten posters include applications of the highlighted tools and resources and case studies of scholarly communication approaches in linguistics.

Workshop Abstracts:

Lauren B. Collister (University of Pittsburgh)

The LSA Statement on Open Scholarship

In May 2021, the Linguistic Society of America's Executive Committee approved the *Statement on the Scholarly Merit and Evaluation of Open Scholarship in Linguistics*. This Statement was developed by the Committee on Scholarly Communication in Linguistics (CoSCiL) and driven by a need for advocacy tools for linguists who develop, maintain, and share openly available research and educational works. In this session, I will present a definition of Open Scholarship and outline the struggles with recognition for Open Scholarship relayed to CoSCiL that led to the development of this Statement. I will then present the advocacy tools and approaches set out by the Statement, including ways to show the impact and importance of Open Scholarship works. Finally, I will share use cases for the Statement that have been collected from the community, including stories from individual linguists who have included Open Scholarship in their tenure or review documents, as well as chairs of linguistics departments who have used the Statement to advocate on behalf of linguists who work on Open Scholarship projects. This session will be a frame for the other presentations and posters in this workshop and to situate the role of Open Scholarship in scholarly communication in linguistics, and to show the value that the LSA sees in Open Scholarship for the health and future of our discipline.

Angelina McMillan-Major (University of Washington)
Emily M. Bender (University of Washington)
Batya Friedman (University of Washington)

Linguistic Data Statements: Documenting the datasets used for training and testing natural language processing systems

Research in natural language processing (NLP) is driven by datasets: collections of naturally occuring linguistic behavior with and without annotations that are used for both training of machine learning systems and testing of all kinds of NLP technology. Documentation of such datasets are critical for both scientific validity and ethical practices in NLP, helping practitioners understand the domain of generalization for results on a dataset, the potential for mismatch between training data and a deployment context, and potential biases that may be learned and amplified by systems trained on the data. Data statements (introduced by Bender & Friedman 2018) are a practice structuring that documentation, including information about how the data were selected (curation rationale), the language varieties and speakers represented in the data (including annotators), the speech situation, preprocessing steps, data capture quality, limitations, and pointers to source datasets, licensing information, annotation guidelines, and dataset quality metrics. We present version 2 of the data statements schema and a how-to guide providing best practices for creating data statements, developed on the basis of a workshop where researchers from all around the world developed data statements for diverse datasets. A data statement is a critical kind of scholarly communication that focuses on metadata, helping scholars understand products we create for each other and positioning the general public to advocate for appropriate deployment of so-called "AI" systems.

Andrea L. Berez-Kroeker (University of Hawa'i at Mānoa)
Bradley McDonnell (University of Hawa'i at Mānoa)
Eve Koller (Brigham Young University Hawai'i)
Lauren Collister (University of Pittsburgh)
The Open Handbook of Linguistic Data Management

In this talk, we present the forthcoming *Open Handbook of Linguistic Data Management*, to be published in late 2021 by MIT Press Open. The Handbook was conceived on the belief that data, in many forms and from many sources, underlie the discipline of linguistics, and proper management of data collections is essential to the future of our field. Linguistic data must be understandable, discoverable, reusable, shareable, remixable, and transformable. Although all data sets must be managed conscientiously and carefully, historically, methods for managing data in our field have been developed somewhat in isolation. Different subfields, research labs, and even individual researchers have developed their own practices and expectations regarding proper management of data. Furthermore, the discipline of linguistics still does not have a culture of broad and open discussion about data. Despite the barriers, however, the reality of linguistic practice today is that most of us use data, most of us wish to use them thoroughly and carefully, many of us share data and code with our colleagues, and most of us have some methods for managing data, whether or not those methods have been codified. Thus, the Handbook grew out of a need to provide a forum in which researchers could share their data management practices with the aim of learning more about the current state of data work across the field. In this way, we hope that the discipline can reflect deeply about the past and present, and foster an open conversation about the future of data work in linguistics. Here we present the two major section of the Handbook. Each of the full-length chapters in Part 1 delves into prominent issues surrounding data and data management. Part 2 consists of 43 shorter data management use cases, each of which demonstrates a concrete application of the abstract principles of data management in specific studies, some actual, and some hypothetical.

Helene N. Andreassen (UiT The Arctic University of Norway)
The Tromsø Recommendations for the Citation of Research Data in Linguistics

Transparency and reproducibility of research receive increasing attention in discussions on scholarly communication and good research practices (Munafò et al. 2017; Alter and Gonzalez 2018). A key element of these practices is appropriate citation of data sources, and data citation practices in linguistics are varied. While linguists have always relied on language data of many types and formats, data from publications are not always available and, when they are, the citation practices make it difficult if not impossible to understand exactly how the data were used (Berez-Kroeker et al. 2018). A great deal of published linguistic research is therefore not reproducible, either in principle or in practice.

In this presentation, I describe the Tromsø recommendations for citation of research data in linguistics (Andreassen et al. 2019), a scholar-led initiative within the frame of the Research Data Alliance which aims to support researchers and scientific publishers who wish to increase the transparency and reproducibility of linguistic research.

We first present the rationale behind the recommendation, including movements towards better data citation practices across all disciplines. Thereafter, we explain the recommendations and highlight a few issues that have caused much fruitful discussion in the development process. We end the presentation with a discussion on how linguistic researchers can proceed in order to implement the recommendations in their workflow.

Paul De Decker (Memorial University of Newfoundland)

Building community-based documentary media in linguistic research

In this presentation I introduce two open education resources designed to enhance scholarly communication and community engagement in linguistics through the creation of interactive research documentaries. The first is a multi-sector, online panel presentation, “Click Your Own Adventure” (CYOA), created in the style of the popular adventure novels from the 1980s. The current iteration features 12 interactive interviews with academics, broadcasters, producers, funders, and charities; by clicking through each interview, the user decides how far to pursue each participant’s ideas, narratives and knowledge about the use of documentary media in scholarly communication. This format allows users to pursue the interview portions and topics of greatest interest. The second, "This is not a documentary film" outlines the principles expressed in the CYOA, offering a guide to using the collaborative media process for graduate students and early career researchers in linguistics. Both CYOA and This is Not... are living texts that are actively updated and revised based on feedback from users and available through Pressbooks, an open education publishing system. The goal with releasing each under a Creative Commons licence is to make them available to linguists and the communities they work with so they might document how the process of research in linguistics works and the contexts in which knowledge is co-created.

Poster Abstracts:

Shirley Gabber (University of Hawai‘i at Mānoa)

The open-access companion course to The Open Handbook of Linguistic Data Management

In this poster I present the open-access online companion course to the forthcoming Open Handbook of Linguistic Data Management. The online companion course will feature units corresponding to the first thirteen chapters of the Handbook, including the need for good data management in linguistics; situating linguistics in the social science data movement; the scope of linguistic data; ethics and Indigenous peoples; the linguistic data lifecycle; copyright; linguistic data in the long view; metrics for evaluating the impact of linguistic data sets; guidance for citing linguistic data; and evaluation of data work in hiring, tenure and promotion. Each unit contains a review of key concepts as well as self-administered quizzes, related activities, and suggestions for implementing lessons into one’s own career.

Susan Smythe Kung (University of Texas Austin)

Developing a data management plan

A Data Management Plan (DMP) is a document created early in a research project that describes the types of data to be generated; how the data will be compiled, analyzed, and stored; who will have access to the data during the project; the legal and ethical status of the data; and how the data will be handled after the project is complete, including deletion or destruction of some or all of the data, long-term preservation of a subset of the data, and how preserved data will be shared. While many funders, publishers, and institutions require a DMP for all new research projects, many researchers view the DMP as a burden. However, the reality is that good data management planning from the project outset can save time, money, and frustration, while ultimately helping to increase the impact of research. This poster is intended to guide researchers through the process of developing and writing a comprehensive DMP that can be modified to satisfy any requirements.

Claire Bowern (Yale University)

Reflections on the Chirila Database

This poster provides a data management use case based on the Chirila database; a collection of lexical resources for historical linguistics from the languages of Australia. I build on earlier work which describes the data structures to document decisions about how the database was structured and forms coded. I describe some pitfalls of complex historical data and discuss pros and cons of key choices. The Chirila database has evolved over its 12 years of

development, and as the field of historical linguistics has changed. This has created both a test of flexibility in data structures and an illustration of the need to be careful about data curation decisions. I describe a dataset which is simultaneously an archive, a research tool, and a way to disseminate language information to individuals and communities.

Valerie Fridland (University of Nevada, Reno)
Tyler Kendall (University of Oregon/National Science Foundation)
Managing sociophonetic data in a study of regional variation

This poster considers the data and data practices from a multi-year, multi-pronged project studying regional variation in U.S. English, funded by the National Science Foundation, the University of Nevada, Reno, and the University of Oregon. The project collected speech production data and administered a series of speech perception tests in a number of research sites in the Northern, Southern and Western United States. The larger aim of the project was to examine the role of regionally-based social and linguistic experience in shaping speakers' production and perception of vowel quality. In this poster, we review our data collection, processing, and management practices, paying particular attention to the problems we encountered and our attempted solutions in working with a regionally diverse project team over many years. Overall, we hope the treatment in the poster and our own experiences might help illuminate potential approaches, and avoid potential pitfalls, for future projects.

Kristine A. Hildebrandt (Southern Illinois University, Edwardsville)
Alena Witzlack-Makarevich (Hebrew University)
Johanna Nichols (University of California, Berkeley)
Taras Zakharko (University of Zurich)
Balthasar Bickel (University of Zurich)
Managing AUTOTYP data: Design principles and implementation

The data management use case presented in this poster describes AUTOTYP, a large-scale research program with goals in both quantitative and qualitative typology. AUTOTYP is one of the oldest typological databases still in use and continuously developed for almost 25 years. From its first days AUTOTYP followed a radically different design philosophy than the one adopted by many traditional typological databases. This poster outlines the five major principles of AUTOTYP viz. modularity and connectivity, autotypology, the division of labor between definition files and data files, the exemplar-based method, and the principle of late aggregation. The implementation of these principles is illustrated using the example of the AUTOTYP module on grammatical relations.

Kathleen Currie Hall (University of British Columbia)
J. Scott Mackie (University of British Columbia)
Roger Yu-Hsiang Lo (University of British Columbia)
Managing and analyzing data with phonological corpustools

The data management use case presented in this poster describes Phonological CorpusTools, a free, open-source, cross-platform software tool that is designed to facilitate the phonological analysis of transcribed corpora. We first explain the overall rationale for and structure of the software and then discuss how it can be used in conjunction with two different kinds of data: pre-existing corpora and original or fieldwork data. Throughout the poster, we present various aspects of the software and its use that we believe reflect good data management practices.

Lina Hou (University of California, Santa Barbara)
Ryan Lepic (Gallaudet University)
ASL internet Corpus

The development of affordable and convenient video recording technology has led to an expansion of sign language videos on the Internet, particularly for American Sign Language (ASL). The availability of such videos opens up new opportunities for researchers to work with naturalistic data, from signers who have voluntarily shared content online. We discuss the advantages and benefits of using Internet-based data to advance sign language research towards open

methods, using a few Internet-based ASL studies as examples. We also discuss some of the practical and ethical considerations for managing Internet-based sign language data, to initiate a conversation about how to promote open methods in sign language research.

Rolando Coto-Solano (Dartmouth College)
Sally Akevai Nicholas (Massey University)
Brittany Hoback (University of Wellington)
Gregorio Tiburcio Cano (Office of the Secretary of Education, Guerrero, Mexico)

Data Management in Untrained Forced Alignment for Phonetic Research: Examples from Costa Rica, Mexico, the Cook Islands and Vanuatu

Forced alignment is a technique to match a recording of spoken language with its transcription, down to the level of words and phones. This can be used in an "untrained" manner to align data from Indigenous and other under-resourced languages. Here we show case studies from six languages: Bribri, Cabécar, and Malecu from Costa Rica, Me̲'pha̲a Váthᵾá from Mexico, Cook Islands Māori from the Cook Islands and Denggan from Vanuatu. We focus on the workflow to prepare and process this data for phonetic research and we discuss the linguistic, ethical and data management challenges involved in performing this research.

Philip Duncan (University of Kansas)
Harold Torrence (University of California, Los Angeles)
Travis Major (University of California, Los Angeles)
Jason Kandybowicz (City University of New York)

Managing data for a theoretical syntactic study of under-documented languages

This poster highlights key aspects of the design and workflow of a collaborative project to document the interrogative systems of two Ghana-Togo Mountain languages, Ikpana and Avatime. The version of data management we outline here is but one example of what data management for theoretical syntax could look like. Some of what we emphasize is particular to working with un- or under-documented and Indigenous languages, though we also attend to principles that are useful across different contexts. In terms of implementing data management best practices, we aimed to maximize usability (e.g., across linguistic disciplines and for purposes of dissemination/publication), replicability (both in terms of our results as well as our management framework), and accessibility (for project-internal purposes and for sharing). For syntax in particular, we emphasize two interacting priorities that inform all domains of planning and management: first, the importance of doing good descriptive and documentary linguistics throughout to provide a solid foundation for theoretical work; and, second, the need for theoretically-informed description and data collection.

Cristina Guardiano (Università di Modena e Reggio Emilia)
Hilda Koopman (University of California, Los Angeles)

Managing Data in TerraLing

TerraLing is a database-backed web application set up to collect, store and explore data for comparative research in the linguistic sciences. TerraLing is publicly accessible and open-ended: new languages, contributors (preferably native speaker linguists), properties and databases can be added so as to allow the database to grow over time. TerraLing aims to: (a) make linguistic data widely available on a group of sister databases, whether the data come from well-studied or understudied languages (including dialects), from spoken or signed languages, or from endangered, extinct, or emerging languages, (b) to provide a common set of powerful queries and analytical tools on the web application to explore the data in each database, and (c) to enable language researchers to easily set up additional sister databases. The long-term goal is to turn TerraLing into a ready-made community tool that linguistic projects can use to gather and store their data for comparative research purposes.