Acknowledgments

This presentation was inspired by and draws on the work of these scholars in language documentation, citation, linguistic reproducibility, and linguistic justice:

- Dr. Andrea L. Berez-Kroeker
- Dr. Steven Bird
- Dr. Lauren B. Collister
- Dr. Anne H. Charity Hudley
- Dr. Lauren Gawne
- Dr. Gary F. Simons
- Dr. Nicholas Thieberger
Open Scholarship
What is Open Scholarship?

“Open Scholarship is a broad and somewhat loose term that has emerged in response to a need to identify and speak to a growing number of open practices within academia.

Open Access, Open Science, Open Data, and Open Education make up a sprawling landscape of interconnected approaches to and focuses on “openness”.

With roots in democracy, equality, and social justice, all areas of Open Scholarship operate under the assumption that knowledge creation and dissemination should be understood as social practices.”

From the Open UBC Statement: https://open.ubc.ca/about-us/
Values of Open Scholarship

- Community and collaboration that supports open knowledge creation
- Social justice commitment to reduce access barriers to information sources
- Openly sharing the results of research to enhance reproducibility and credibility in results
- Transparency through open sharing of grant proposals, research protocols, and data
- Speed and efficiency in sharing research findings
- Supporting open and free participation
- Engaging in technologies and networks to create collaborative communities
- Advancing community knowledge by working together for the common good
- Engaging the users of knowledge in the process of knowledge production

*Image credit:* UBC Open Scholarship [https://open.ubc.ca/about-us/](https://open.ubc.ca/about-us/) under a CC by-sa 4.0 license
Reproducible Research

“Reproducible research aims to provide scientific accountability by facilitating access for other researchers to the data upon which research conclusions are based.”

Reproducibility

Bird and Simons seminal article: Seven Dimensions of Portability for Language Documentation and Description

- present at least four domains of data management that directly support reproducible research as it is understood here: citation, discovery, access, and preservation.
- 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”
- 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
- presumes that the data themselves are also properly preserved, discoverable, and accessible.
Several recent studies have looked at linguistics publications with regard to metrics of reproducibility.

2017 survey of 50 dissertations & 50 grammars by Gawne et al

- Looked at language grammars over a 10-year span (2003–2012)

Findings

- looked at transparency of data collection methods, analysis and storage, as well as citation of primary data.
- “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.”

Reproducibility Metrics in Linguistics Journals

2017 survey of linguistics journals by Berez-Kroeker and colleagues

- Looked at 270 articles from 9 top international linguistics journals from 2003-2012

Findings:

- Over half the authors of articles do not explicitly state where the data upon which their study is based can be located
- Few authors met any of the survey’s metrics for basic transparency of data and methodology
- Articles had minimal description of methods of data collection and analysis

Citing and Linking Source Data

Language grammars typically do not reference source data or how to access it.

Eg:

● Was the data all elicited or was it recorded and transcribed?
● If it was recorded, then who was it recorded with – are the speakers old or young, male or female?
● If texts are the source of example sentences, then where in the text does the example come from?
● Where is the data itself stored?

An Example: Linked Grammars & Source Texts

“How much more useful will it be when grammars of little-known languages provide not just decontextualised and unsourced examples, but annotated corpora linked to media?”

An Example: Linked Grammars & Source Texts

Oceanic Linguistics Special Publication No. 33

A Grammar of South Efate
An Oceanic Language of Vanuatu

Nicholas Thieberger

Collection details

Collection ID
NT1

Title
South Efate (Vanuatu)

Description
Records in this collection were made during fieldwork, mainly in Erafaki Village, Efate, Vanuatu since 1996. They consist of audio recordings of South Efate speakers, originally on audio cassettes (hence there are often A and B sides of a single item). There is a total of over 40 hours of language material, and the collection includes transcribed transcripts in Efani for many recordings, and interlinear text for a small proportion of them. It also includes field notes. The material here is the basis for Thieberger’s grammar of South Efate which also cites primary material in this collection.

Where there is a transcript the file has the same name as the recording with a .txt extension. These transcriptions are not as accurate as the versions that are found in the interlinear annotation, and also have relics of earlier transcription conventions (p and m for coarticulated b and f).

Many of these transcriptions have then been interlinearised in the version that can be found in NT8. More recordings and other material in South Efate can be found in collections NT1 to NT7 in PARADISEC. A description of the collections is given here: https://go.coedl.net/NatlamMaterial

Archive Link
http://catalog.paradise.org.au/repository/NT1

Collector
Nick Thieberger

Operator

Originating university
University of Melbourne

Items in Collection (35)

1 2 3 4 Next > Last »

Item Title Actions
001 Elicitation with Sisas Alben View
002 Words read by Sisas Alben View
003 Words and Sentences read by Sisas Alben View
004 Sisas Alben telling stories View
005 Recordings in South Efate View
006 Recordings in South Efate View
1996 Fieldnotes from 1996 View
20001 Recordings in South Efate View
20002 Audio recordings in Erafaki village View
20003 Recordings in South Efate View

Show 10 | Show 50 | Show 100 | Show all 35

https://catalog.paradise.org.au/collections/NT1
Another Example: Dissertations & Linked Data

The SpiCE Corpus

SpiCE is an open-access corpus of conversational bilingual Speech in Cantonese and English. SpiCE includes high-quality audio recordings of 30-minute interviews with 34 early bilinguals in each language with accompanying transcriptions and language background information. The corpus was first released in May 2021. Detailed information about the corpus is provided on the design and transcription pages.

Contributors

- Khia A. Johnson — get in touch at khia.johnson@ubc.ca or @khia_johnson
- Molly Babel
- Ivan Fong
- Nancy Yiu
- ...and many other members of the Speech-in-Context Lab

Citing the corpus

When referring to the corpus in the body of your text or presentation, you can call it the “SpiCE Corpus” (please keep the letters in their specified cases), but should spell out “Speech in Cantonese and English” at least once!

You can cite the corpus directly (preferred):

Another Example: Dissertations & Linked Data

Facilitating the Move to Open Scholarship
Research Incentives

- Align open practices with review, tenure & promotion (RPT) guidelines and research missions & values
- RPT guidelines frequently cited as barriers to openness
Funding Agencies

March 2016: Canada’s research funding agencies launch research data management policy

Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council (NSERC) and the Social Sciences and Humanities Research Council (SSHRC)—launched the Tri-Agency Research Data Management Policy (RDM policy).

The agencies’ RDM policy establishes requirements for institutional RDM strategies, data management plans and data deposit.
Scholarly Society Statements

Linguistic Society of America

Resolution on Cyberinfrastructure (2010) 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

https://www.linguisticsociety.org/resource/resolution-cyberinfrastructure

Resolution Recognizing the Scholarly Merit of Language Documentation (2010)

● linguistic fieldwork shifting to a more collaborative endeavor based on ethical responsibilities to speech communities
● products of language documentation and work supporting linguistic vitality are of significant importance to the preservation of linguistic diversity
●

https://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation
Publishers Roles and Policies

Language and linguistics editors and publishers can play a key role in guiding the field towards openness by:

- developing data policies for their journals that require authors to describe their methods and cite all data

- guiding authors to appropriate repositories for their data sets e.g. Tromsø Repository of Language and Linguistics [https://trolling.uit.no](https://trolling.uit.no)

- developing and enforcing policies for both data sharing and data citation

- advocating formats for the citation of linguistic data sets

See: Tromsø Recommendations for Citation of Research Data in Linguistics [doi.org/10.15497/rda00040](doi.org/10.15497/rda00040)
New Book

Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences

MIT Press, 2020

This volume examines the challenges inherent in making diverse data in linguistics and the language sciences open, distributed, integrated, and accessible, thus fostering wide data sharing and collaboration.
Thank you! 
Questions?