Linguistics Data Interest Group[1] Charter Statement

Version 1.0 14th June 2017

Introduction

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. Publications typically include only short excerpts from data sets, and where citations are provided, the connections to the data sets are usually only vaguely identified. At the same time, the field of linguistics has generally viewed the value of data without accompanying analysis with some degree of skepticism, and thus linguists have murky benchmarks for evaluating the creation, curation, and sharing of data sets in hiring, tenure and promotion decisions.

This disconnect between linguistics publications and their supporting data results in much linguistic research being unreproducible, either in principle or in practice. Without reproducibility, linguistic claims cannot be readily validated or tested, rendering their scientific value moot. In order to facilitate the development of reproducible research in linguistics, The Linguistics Data Interest Group (LDIG) plans to develop 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. The LDIG is for data at all linguistic levels (from individual sounds or words to video recordings of conversations to experimental data) and data for all of the world's languages, and acknowledges that many of the world's languages have high cultural value and are underrepresented with regards to the amount of information that is available about them.

This interest group is aligned with the RDA mission to improve open sharing of data through forming transparent discipline-specific data citation and attribution conventions to be adopted by the international research community. This interest group will add value to the RDA community by providing breadth to the current roster of RDA interest groups: linguistics is a discipline that straddles social/behavioral sciences and the humanities, and thus we have a great deal to contribute to the general RDA discussion on a multiplicity of data types. This group ties in with other initiatives in transparent research methods in linguistics at all stages of the workflow, including Open Access data archiving and publishing, reproducible methodologies and critical consideration of data licensing. The LDIG seeks to support these initiatives while focusing on data citation specifically. The LDIG provides an ongoing space for linguists to come together to improve how we manage and cite our data, and how we train linguists in good practice.

[1] www.rd-alliance.org/groups/linguistics-data-interest-group

Who this group is for?

The LDIG is for people who work with linguistic and language data. This work includes, but is not limited to, the collection, management and analysis of linguistic data. We encourage participation from academic and speaker communities.

Objectives and outcomes

Our overarching objective is to contribute to a positive culture of linguistic data management and transparency in ways that are in keeping with what is happening in the larger digital data management community. To do this we aim to be a group that is able to provide tangible tools (e.g. guidelines, software) for improving the culture of data citation and attribution within linguistics. This will also involve understanding the breadth of data types linguists work with, and current uses of persistent identifiers. We outline three main objectives. For each objective we also suggest specific outcomes, which would be the focus of shorter term timelines (e.g. Working Groups):

- Development and adoption of *common principles and guidelines* for data citation and attribution by professional organizations, such as the Linguistic Society of America and the Societas Linguistica Europaea, academic publishers, and archives for linguistic and language data. Principles and guidelines will follow the recommendations in the Joint Declaration of Data Citation Principles.[2]
  Potential WG topics include:
  - Development of a *common stylesheet for citation of linguistic data*
  - Adoption of the style sheet by publishers, archives, organisations and individuals
  - Integrating RIS with linguistic data services like the Open Language Archives Community
- *Education and outreach efforts* to make linguists more aware of the principles of reproducible research and the value of data creation methodology, curation, management, sharing, citation and attribution. Practical training also helps.make proper data preparation less burdensome for researchers, and normalises this work as an expectation of the discipline. While much of this work will be practical training, outreach also needs to take into account the complex and varying attitudes towards creation of open access data sets across linguistics.
  Potential WG topics include:
  - Development of training modules
  - Delivery of training at conferences and workshops
  - Development of tools for the management of linguistic data
- Efforts to ensure *greater attribution of linguistic data set preparation* within the linguistics profession.
  Potential WG topics include:

[2] https://www.force11.org/group/joint-declaration-data-citation-principles-final

    - Framework for valuing the development of linguistic data sets in job appointments, tenure and promotion applications and in research degrees and postdoctoral research projects.

It will be up to the LDIG to decide if any of these specific outcomes would be best met by forming short term working groups with specific timelines for the deliverables. Other outcomes may be worked on within the LDIG on a more open timeline. Further goals include fostering greater transparency in research methodology, and data access rights. We expect that other outcomes will be developed as LDIG grows and responds to the changing research environment.

Mechanism

The co-chairs will hold a conference call every two months. The wider LDIG will convene quarterly meetings. The timezone spread of LDIG members means that these meetings will be held asynchronously in an editable document. The agenda will be posted with discussion points, and will be open for comment for a week, before actions are decided upon and delegated. We will also host face-to-face meetings at relevant linguistics conferences, such as Societas Linguistica Europaea, Linguistic Society of America, the Australian Linguistics Society, and at the RDA plenaries.

Interaction with groups in RDA

The following RDA groups have been identified as having interests that are relevant to LDIG, both in terms of technical and ethical issues in linguistic data management:

- Data policy standardisation and implementation IG
- Data Versioning IG
- Reproducibility IG
- RDA/NISO Privacy Implications of Research Data Sets IG
- Ethics and Social Aspects of Data IG
- Metadata IG
- Data Citation WG
- BoF on Data Champion Communities
- RDA/WDS Publishing Data IG

While setting up the LDIG we will ask at least four of our members to nominate themselves to participate in one of these other groups and be officially named as our cross-group co-ordinator. This will facilitate cross-group relevance.

Linguists from particular subfields may find that particular interest groups are relevant to particular issues in their area, for example corpus linguists may find that the Big Data IG addresses relevant

issues. We encourage LDIG participants to also engage with other interest groups and working groups in the RDA.

Related projects and activities

There are also a number of organisations and groups outside the RDA that LDIG will engage with directly as the objectives of the group are addressed.

- Digital Endangered Languages and Musics Archives Network (DELAMAN)[3]
- Linguistic Society of America Committee for Scholarly Communication in Linguistics (CoSCIL)[4]
- Tromsø Repository of Language and Linguistics (TROLLing)[5]
- Data Citation and Attribution for Reproducible Research in Linguistics project, sponsored by the National Science Foundation (SMA 1447886)[6]
- Open Language Archives Community (OLAC)[7]
- Linguistic Data Consortium[8]
- The LINGUIST List[9]
- The Leipzig Glossing Rules[10]
- The Generic Style Rules for Linguistics[11]
- The Unified Style Sheet for Linguistics Journals[12]
- CLARIN - European Research Infrastructure for Language Resources and Technology[13]
- FORCE11 Attribution Working Group[14]

Contributors

Co-Chairs:
Andrea L. Berez-Kroeker, U Hawai'i at Mānoa
Lauren Gawne, La Trobe University
Helene N. Andreassen, UiT The Arctic University of Norway

Potential members:

[3] http://delaman.org/
[4] www.linguisticsociety.org/content/committee-scholarly-communication-linguistics-0
[5] https://dataverse.no/dataverse/trolling
[6] http://sites.google.com/a/hawaii.edu/data-citation/
[7] http://www.language-archives.org/
[8] www.ldc.upenn.edu/
[9] www.linguistlist.org
[10] www.eva.mpg.de/lingua/resources/glossing-rules.php
[11] http://www.eva.mpg.de/linguistics/past-research-resources/resources/generic-style-rules.html
[12] www.linguisticsociety.org/resource/unified-style-sheet
[13] https://www.clarin.eu/
[14] https://www.force11.org/group/attributionwg

Felix Ameka, Leiden U
Helene N. Andreassen, UiT The Arctic U of Norway
David Beaver, U Texas at Austin
Andrea Berez-Kroeker, U Hawai'i at Mānoa
Brian Carpenter, American Philosophical Society
Lauren Collister, U Pittsburgh
Meagan Dailey, U Hawai'i at Mānoa
Stanley Dubinsky, U South Carolina
Ruth Duerr, U Colorado Boulder
Colleen Fitzgerald, National Science Foundation
Lauren Gawne, SOAS, University of London
Jaime Pérez González, U Texas at Austin
Ryan Henke, U Hawai'i at Mānoa
Gary Holton, U Hawai'i at Mānoa
Kavon Hooshiar, U Hawai'i at Mānoa
Tyler Kendall, U Oregon
Susan Smythe Kung, U Texas at Austin
Richard P. Meier, U Texas at Austin
Bradley McDonnell, U Hawai'i at Mānoa
Geoffrey S. Nathan, Wayne State U
Peter Pulsifer, U Colorado Boulder
Keren Rice, U Toronto
Gary Simons, SIL International
Maho Takahashi, U Hawai'i at Mānoa
Nick Thieberger, U Melbourne
Jessica Trelogan, U Texas at Austin
Paul Trilsbeek, Max Planck Institute for Psycholinguistics
Mark Turin, U British Columbia
Laura Welcher, Long Now Foundation
Nick Williams, U Colorado Boulder
Margaret Winters, Wayne State U
Anthony Woodbury, U Texas at Austin

LDIG has also been promoted through the LINGUIST List, and we invite any interested party to participate.

Timeline

The LDIG aims to be an ongoing group, whose overall aim is to promote better practice in linguistic data management. A general timeline is given, however some of these responsibilities may be handed over to a working group specifically set up for the delivery of the data citation standards.

<u>Outreach - first 6 months (May-November 2017)</u>

- April 2017 Draft charter posted
- May 2017 Group advertised publically
- June 2017 Amended charter posted
- Sept 2017 Attend Montreal RDA plenary and connect with relevant RDA groups
- Oct 2017 Finalise LDIG structure and communication processes

<u>Groundwork - second 6 months (November 2017-May 2018)</u>

This groundwork helps us expand the reach of the LDIG and ensures that we are as relevant and inclusive as possible. Includes attendance at April 2018 RDA plenary:

- Survey of linguists on current data citation practice (individual practice and institutional level training opportunities)
- Collate possible citation practices
- Survey of linguists on current practices for academic attribution of curation of linguistic data sets in departmental tenure and promotion