Linguistics Data Interest Group[1] Charter Statement

Version 0.1 draft Charter Statement 7th April 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 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.

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.

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.

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

Objectives and outcomes

Our overarching objective is to provide tangible tools (e.g. guidelines, software) for improving the culture of data citation and attribution within linguistics. We outline three main objectives, and specific outcomes for each:

- 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, and academic publishers;
  Outcomes include:
  - Development of a *common stylesheet for citation of linguistic data*
  - Adoption of the style sheet by publishers, organisations and individuals
- *Education and outreach efforts* to make linguists more aware of the principles of reproducible research and the value of data creation, curation, management, sharing, citation and attribution;
  Outcomes 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.
  Outcomes include:
  - Framework for valuing the development of linguistic data sets in job appointments
  - Framework for valuing the development of linguistic data sets in tenure and promotion applications.

We expect that other outcomes will be developed as LDIG grows.

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, and 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

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)[2]
- Linguistic Society of America Committee for Scholarly Communication in Linguistics (CoSCIL)[3]
- Tromsø Repository of Language and Linguistics (TROLLing)[4]
- Data Citation and Attribution for Reproducible Research in Linguistics project, sponsored by the National Science Foundation (SMA 1447886)[5]
- Linguistic Data Consortium[6]
- The LINGUIST List[7]
- The Leipzig Glossing Rules[8]
- The Unified Style Sheet for Linguistics Journals[9]
- CLARIN - European Research Infrastructure for Language Resources and Technology[10]

[2] http://delaman.org/
[3] www.linguisticsociety.org/content/committee-scholarly-communication-linguistics-0
[4] http://opendata.uit.no/dataverse/trolling
[5] http://sites.google.com/a/hawaii.edu/data-citation/
[6] www.ldc.upenn.edu/
[7] www.linguistlist.org
[8] www.eva.mpg.de/lingua/resources/glossing-rules.php
[9] www.linguisticsociety.org/resource/unified-style-sheet
[10] https://www.clarin.eu/

Contributors

Co-Chairs:
Andrea L. Berez-Kroeker, U Hawaiʻi at Mānoa
Lauren Gawne, La Trobe University
Susan S. Kung, U Texas at Austin
Helene N. Andreassen, UiT The Arctic University of Norway

Potential members:
Felix Ameka, Leiden U
Helene N. Andreassen, UiT The Arctic U of Norway
David Beaver, U Texas at Austin
Andrea Berez-Kroeker, U Hawaiʻi at Mānoa
Brian Carpenter, American Philosophical Society
Lauren Collister, U Pittsburgh
Meagan Dailey, U Hawaiʻi 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 Hawaiʻi at Mānoa
Gary Holton, U Hawaiʻi at Mānoa
Kavon Hooshiar, U Hawaiʻi at Mānoa
Tyler Kendall, U Oregon
Susan Smythe Kung, U Texas at Austin
Richard P. Meier, U Texas at Austin
Bradley McDonnell, U Hawaiʻi at Mānoa
Geoffrey S. Nathan, Wayne State U
Peter Pulsifer, U Colorado Boulder
Keren Rice, U Toronto
Gary Simons, SIL International
Maho Takahashi, U Hawaiʻi 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 will also be promoted through the LINGUIST List, and we invite any interested party to participate.

Timeline

Outreach - first 6 months (May-November 2017)

- April 2017 Draft charter posted
- May 2017 Group advertised publically
- July 2017 Member comment (within 6 weeks of draft going live)
- Sept 2017 Revised charter posted
- Sept 2017 Attend Montreal RDA plenary and connect with relevant RDA groups
- Oct 2017 Finalise LDIG structure and communication processes

Groundwork - second 6 months (November 2017-May 2018)

To be driven by Working Groups lead by 1-2 LDIG Chairs; 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

Building the citation standards - third 6 months (May 2018-November 2018)

- Development of the citation standards
- Development of a statement on and guidelines for tenure and promotion committees and applicants about how to weigh data set curation in linguistics
    - Includes attendance at September 2018 RDA plenary