Linguistics Data Interest Group (LDIG)

Meeting - August 30-September ~~8~~ 15 2017
Pre-RDA Plenary 10

-- Document editing now closed --

How to participate

Welcome to this asynchronous meeting of the Linguistics Data Interest Group (LDIG).

There are three ways to participate:

1. Enter text or add a comment to this document. Make sure you include your name and list yourself as a meeting participant (with your email address) at the end of the document.
2. Email feedback to Lauren Gawne (l.gawne@latrobe.edu.au), which will be incorporated into the final version (anonymously if requested).
3. Add your feedback in this Google Form, which asks the same questions as in this meeting document. Your feedback will be incorporated into the final version (anonymously if requested).

Aims of this meeting

1. Get feedback on the Austin Principles of Data Citation
2. Find specific examples with which to illustrate the Austin Principles
3. Get feedback on other issues and ideas to take to the RDA's 10th Plenary
4. Keep LDIG members up to date on what we are working on

Context

This meeting has been set up to get feedback on the first draft of the Austin Principles of Data Citation in Linguistics. The LDIG chairs will be attending the RDA 10th Plenary in Montreal from the 19th-21st of September. At the meeting, the feedback will be used to produce a version of this document that will be more widely circulated. Once general principles of data citation are established we can focus on working together to create specific guidelines. By holding this meeting asynchronously in this document we hope you are able to participate, and contribute to this current work (fulfilling aim 4 of this meeting).

Read the Austin Principles of Data Citation

The Austin Principles of Data Citation are available on the Data Citation and Attribution in Linguistics project page:
https://sites.google.com/a/hawaii.edu/data-citation/austin-principles-of-data-citation-in-linguistics

These principles are based on the Force 11 Joint Declaration of Data Citation Principles, and annotated specifically for linguistics.

Feedback on the Principles (Aim 1)

What are your first impressions reading this document?

Stan Dubinsky: I am mostly quite pleased with the principles as laid out in the document, although I do think that they will be more easily applied to some sorts of research (e.g. archival data collection and cultural curation) than some others (e.g. syntactic, morphological, phonological analysis of linguistic data).
Hiram Ring: I am also quite pleased - I think the main thrust is that data underpinning any linguistic work should be available for replication, if nothing else, which I see as a positive thing. I think Stan's point is valid, but I would also say that analysis of linguistic data can be a dataset (of forms, sentences used in analysis, recordings of sounds, scripts used in R or Praat, Python code, or a host of other kinds of data). The main thing is that it is incumbent on the linguist to track and explain their particular method of reaching particular conclusions, which is just good science.
Alon Lischinsky: I concur. Building on Stan's point, I think that the principles work better when they are formulated in terms that suit the whole linguistics community than when they provide examples (e.g., the DELAMAN archives) that will only be familiar to people in specific subfields. I think it might make sense to reformulate #4 in that way, e.g., 'Data repositories, corpora and other resources for housing and providing access to linguistic data should provide the means for such identification in the form of a PID…'. [Incidentally, it seems that the TROLLing Dataverse is down; if that's permanent, it would be wise to remove it as an example]
Tyler Kendall: I also concur. Stan's (good) point, perhaps, relates to questions some readers might have about what constitutes "data" for these principles. I realize this comes from FORCE11, but the preamble states "data should be considered legitimate, citable products of research" without explaining what "data" are, and whether the consideration is from the creator of the data's perspective or the users' or… But, I realize this comment drifts away from the principles itself and gets into some of the related muddy waters we've talked about in our meetings (e.g. how much of this is really just about citation practices vs larger principles about open-science-based approaches)... In sum, I really like the document and think it does a great job of hitting the main points while staying brief and to the point.

Does the document represent linguistic data the way you work with it?

Stan Dubinsky: In many instances not. I think, for instance, that credit and attribution would need to be examined closely, in terms of whether a participant is credited as a co-author/collaborator and awarded authorship, or not. The example citation, given in point #7, is illustrative. Here, the researcher (Sherzer) did not compose, create, or (based on the description) even analyze the data object. He did collect it, render it accessible, and commented on it (although his commentary CUK001R002I700.pdf appears not to have been made accessible on the webpage hosting the object). In contrast, a theoretician or psycholinguist working with data is going to use that data to create original output, whether in the form of acceptability judgments or experimental results, and

s/he will do this by creating testable hypotheses or designing an experiment that manipulates measurable conditions. In a case like this, I (as a theoretician) might have a data source (native speaker and their judgements), help in creating testable data (a native speaker helping to create/edit/transcribe examples), or help in interpreting data (bilingual native speaker reflections on the examples and their meaning). I would not, unless the native speaker were a party to discovering the data or in analyzing it, accord coauthorship to a native speaker in the manner shown in point #7, although it would be completely appropriate to acknowledge such contributions in a note at the beginning of the paper. So, I think that participation will be quite distinct depending on whether the research involves linguistic analysis and description or whether it involves archiving, curation, and maintenance.

Hiram Ring: It does represent much of the way I work with linguistic data, and increasingly how I am thinking about working with it. Stan makes a really good point about the participation of speakers, though. I think the way that they deal with this in Psychology is that the author of a dataset is generally the person who creates it, and if there is any confusion you can use a general "time spent" test to determine this - the person who has "spent the most time" with a particular set of data is the primary author. In the case of language documentation, this is generally the person who made the recording and sat for hours with multiple speakers of the language to transcribe, translate, and annotate the recordings. At the same time, within that dataset, each recording should credit the source (speaker) of the video/audio or anonymize the source if the person requested this. I think a good rule of thumb is "when in doubt… attribute!". Ideally, when you are getting IRB (ethics) approval, questions of ownership and attribution will already have come up and have been thought through. With datasets of anonymized participants (which are required by many quantitative kinds of research anyway) you at least need an identifier to be able to differentiate individuals. The point in my mind is that datasets are not the same as papers, and so the authorship rules we traditionally assign to papers don't really apply here - yet if we're writing papers but don't allow anyone to view the source material that it's based on, it becomes very difficult for anyone to replicate/question our findings.

Alon Lischinsky: Stan and Hiram make good points. It would be a good idea to acknowledge/illustrate the diversity of linguistic data, perhaps in the Preamble, pointing out that they include not only naturalistic attestations of speech and writing, but also elicited grammaticality and felicity ratings, experimental eye-tracking scanpaths, morphological/syntactic/pragmatic annotation, etc. One particular omission that surprised me is that no mention is made of linguistic corpora as datasets (and, often, stand-alone repositories).

Tyler Kendall: Yeah, the emphasis on attribution for participants is nice and of course there are lots of places where this is very appropriate, but I wonder if how it is framed in the document is too heavy-handed. This works in cases where there are a limited number of primary consultants or informants, but not so much in cases where many people each contribute small amounts to the data, as in sociolinguist community studies, or "conventional" corpus development, or psycholinguistic lab studies. There it might be more appropriate to emphasize that fieldworkers

should get some attribution, as they often don't. But, I think the principles need to be careful to not insist there's a right way, more to remind people to think of all of the people involved in the creation of the data.

Is there anything you would like to see added to the charter?

Alon Lischinsky: regarding principle #4, I think it's worth pointing out that the usual practice in corpus linguistics is for the corpus compilers to recommend a specific way of referring to the data, often citing a published paper or report where the corpus is described rather than the corpus itself (see, e.g., this thread, especially the responses by Eric Atwell). It might be sensible to acknowledge this state of affairs, and perhaps recommend that this be included *alongside* (but not instead of) a proper reference to the data.

Tyler Kendall: Related to Alon's point, and my point above, it's not entirely clear to me whether the "users" of the Austin Principles are users of data or creators of data. Creators of data, such as corpus compilers, can do a lot by offering suggested citations for their published data. Perhaps this could be made more explicit? And perhaps some guidance can be given to users of data in cases where there is no existing recommendation or standard for how to cite a used dataset? More generally, differentiating target audiences would be helpful. I hope the outcome of this isn't just that individuals "clutter up" their CVs and papers with lots of citations to themselves. I know that sounds a little crass - sorry - I hope it makes sense though.

Examples to illustrate the Austin Principles (Aim 2)

We want to illustrate these principles with examples of good practice from across linguistics. If you know a great example of a particular principle, please share it!

Lauren Gawne: I'm hoping that we'll get examples from a range of languages, and a range of linguistic subdisciplines. Don't be modest about nominating your own good work (or if you must be modest, email me at l.gawne@latrobe.edu.au and I'll post it for you!)

Urek, Olga; Taurina, Agrita; Westergaard, Marit, 2017, "Adjectival gender agreement in monolingual and bilingual Latvian- and Russian-speaking pre-schoolers", doi:10.18710/8ECVSD, DataverseNO, V1

Ring, Hiram. 2017. "Replication Data for: A grammar of Pnar", doi:10.21979/N9/KVFGBZ, DR-NTU (Data), V1.

Hiram Ring: *This is a recent dataset that is now linked to my PhD thesis. It took me a couple of years to get it up, mainly because the library server couldn't host 1 GB or more of data in 2015. Copy-pasting the DOI link directs you to dataset hosted on the NTU Singapore Dataverse site, which is a very recent initiative for NTU, patterned after the Harvard Dataverse site (https://dataverse.harvard.edu/) which I encourage everyone to check out. Fortunately I had archived the data on the recommendation of one of my PhD panelists (a computational linguist), and so this version is nearly the exact same dataset that I used for my PhD. The only difference is*

that I anonymized some participants (speakers) who had given me permission to record them and use the data for analysis, but who hadn't given explicit permission to use their names.
Another good example dataset is the AUTOTYP dataset, found here: *https://github.com/autotyp/autotyp-data and cited (by their own documentation) thusly:*

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga & John B. Lowe. 2017. *The AUTOTYP typological databases*. Version 0.1.0 https://github.com/autotyp/autotyp-data/tree/0.1.0

It's worth noting that as code/databases are released publicly, it may be that they are just as likely to be located on repositories like GitHub, and use versioning similarly to software.

Tyler Kendall: This may not be "good practice" exactly, but I thought I'd take advantage of this meeting format to talk about a personal example, for what it's worth. Please ignore this if it's not helpful ;) These kinds of issues are especially relevant to me right now, as a team I lead is getting close to publicly releasing the first parts of a series of public corpora under the umbrella of the *Corpus of Regional African American Language* (CORAAL). CORAAL is composed of a few sub-corpora and our intent is to continue to publish new supplements over time. Basically, the main data come from two collections from Washington DC. One collection was recorded ca. 1968 and was donated to the project by the original researcher, who is now retired, to preserve and make available those data, and the second collection was collected by my team over the last couple of years. We aren't live yet (we're scheduled to host a special, organized session at LSA in January where we will officially release these two collections, from there we will be focusing on development of a series of smaller components that provide regional coverage, across a number of cities/communities in the US). In the meantime, some beta-testers and early adopters are using the data, which is currently hosted on a non-publicly accessible website. On that website I offer some suggested ways to cite the data. Those look like this:

Recommended/Requested Citation and **Version Number** for the main project:

- Kendall, Tyler and Charlie Farrington. 2017. *The Corpus of Regional African American Language.* Version # 0.5. BETA. Eugene, OR: The Online Resources for African American Language Project.

Recommended/Requested Citations and **Version Number** for CORAAL:DCA (1968):

- Kendall, Tyler, Ralph Fasold, Charlie Farrington, Jason McLarty, Shelby Arnson, and Brooke Josler. 2017. *The Corpus of Regional African American Language: Washington DC 1968.* Version # 0.9. BETA. Eugene, OR: The Online Resources for African American Language Project.
- Fasold, Ralph. 1972. *Tense marking in Black English: A linguistic and social analysis.* Washington, DC: Center for Applied Linguistics. [Access on ERIC]

Recommended/Requested Citations and **Version Number** for CORAAL:DCB (2016):

- Kendall, Tyler, Minnie Quartey, Charlie Farrington, Jason McLarty, Shelby Arnson, and Brooke Josler. 2017. *The Corpus of Regional African American Language: Washington DC 2016.* Version # 0.5. BETA. Eugene, OR: The Online Resources for African American Language Project.

(We don't have DOIs at this point, something I realize we'll need to do before we go live, and we don't indicate roles for the different authors.) Charlie Farrington and I are the main

“editors/curators/compilers” for the corpus and after a long set of deliberations we decided that he and I should be the authors, and the only authors, on the citation for the “main” (umbrella) corpus, but for each of the components I also want to credit the main fieldworker(s) (Minnie Quartey, for the new recordings; Ralph Fasold, who was the PI on the original 1968 recordings) and the main RAs who produced the bulk of the annotation (the same three RAs worked heavily on both of the components). The plan is then to acknowledge the other RAs and fieldworkers who helped on the project through acknowledgements, not listing them as authors. One issue with having tons of authors on these kinds of things is that it diminishes the contributions of the main contributors to have all of the less main people listed as well. Without a standard way to list roles, it’s hard to figure out how to do this. Even with roles, it seems hard to adequately attribute people when say one transcriber worked on a single file for less than a week while another worked on the project for a year and a half and had major impacts on the project. (This is hard; hence my enthusiasm for this group’s work!)

For the 1968 data (CORAAL:DCA) I also suggest a citation to the original primary publication that the data are known from. (That’s a well-known study in the history of research on African American English.) We are currently collecting additional datasets and, as I’m anticipating it, those will have their own recommended citations. This is particularly tricky because there are cases (actually, like for Fasold’s data) where the data were collected completely independently, so it seems wrong for me to be first author. (I wanted Fasold to be first author on the 1968 version (CORAAL:DCA) but he insisted otherwise.) For the future datasets, my plan is to be first author if it’s a dataset that was collected specifically for the grant and under my “PI-ship” (basically, if I paid for it) but for others (e.g. a colleague is donating some recordings from her PhD work to be added to the corpus) the original researcher will be first author. My team does a lot of work to create all of the transcription and annotation so donated data still involve a lot of cost and a lot of work by the project team, but how to best attribute that and differentiate it from the core researchers who collected the data in the first place is really tricky(!)... Of course, a lot of this also hinges on what we mean by “data” (I write about this wrt sociolinguistic practices in a 2008 paper, if anyone is curious). The sociolinguistic interviews or conversations recorded in “the field” are the most important part, the groundwork for the rest, but then a lot of work goes into transcribing them and creating those additional layers of “data”. And a lot of work also goes into redacting/anonymizing the recordings, though it’s less clear how that step actually is producing “data” other than making our practices conform with our ethical guidelines and agreements with participants and our IRB(s)...

In any case, I offer this as an example, perhaps less of how the Austin Principles should be used and more as an example of why they are sorely needed! ;) I’d be very happy get input on how we could do recommended citation/attribution better for this project and to use it as an example of the Principles at work…

Thinking about other examples, I’ve created and hosted several non-traditional “scholarly contributions” over the years. The one that gets the most citations on Google Scholar (more than

any of my “traditional” publications) is the NORM website and its associated Vowels.R R package. Erik Thomas, my “co-author” and I provide recommended citations on the website (http://lingtools.uoregon.edu/norm/about_norm1.php). R also provides citations for packages from directly within that software. E.g.:

```
> citation("vowels")

To cite package ‘vowels’ in publications use:

  Tyler Kendall and Erik R. Thomas (2014). vowels: Vowel Manipulation, Normalization, and
  Plotting. R package version 1.2-1. http://blogs.uoregon.edu/vowels/

A BibTeX entry for LaTeX users is

  @Manual{,
    title = {vowels: Vowel Manipulation, Normalization, and Plotting},
    author = {Tyler Kendall and Erik R. Thomas},
    year = {2014},
    note = {R package version 1.2-1},
    url = {http://blogs.uoregon.edu/vowels/},
  }
```

While this is surely a minority of cases, I mention it because other linguists have produced R packages too and this is a place where the mechanisms involved provide a structured way for citations, and these also seem to have gotten lots of traction. I.e. it’s pretty common in sociolinguistics, psycholinguistics, corpus linguistics, etc., to see those packages cited in papers. So this could be a place to point to where examples exist for non-traditional “products” and they aren’t too sub-field specific. Finally, thanks to OLAC, many of the datasets stored in the SLAAP archive are citable, through their OLAC entries (see here).

[This ends Tyler’s, perhaps too long, contribution.]

Any other business you would like us to take to the RDA meeting? (Aim 3)

As we work on the Austin Principles we will also be looking ahead to where we can build from here. Any feedback or suggestions you have will be welcome!

Meeting Participants:

Lauren Gawne (Co-chair)
Andrea L. Berez-Kroeker (Co-chair)
Helene N. Andreassen (Co-chair)
Stan Dubinsky (via Google Form)
Hiram Ring (hiram1@e.ntu.edu.sg, via Google Document)
Alon Lischinsky (alischinsky@brookes.ac.uk, via Google Document)

Tyler Kendall (via Google Document)
Susan Kung (read, agreed, applauded, nothing to add, skung@austin.utexas.edu)
Joel Dunham (jdunham@artefactual.com, via Gotomeeting)