Data citation in journals
Ameka, Dailey (note-taker), Dubinsky, Legate, Rice, Thieberger (facilitator)

First report back
   We need a unified style sheet for citing data that everyone can agree on
   Issue of supplementary material and how to cite it
   What constitutes the data set?
   Next steps

Second Report Back
   Evaluation
   Co-authoring
   Citation style guide

Raw notes
   Evaluation of archival materials
   incremental improvement of materials
   peer review vs published review
   distinction between collection and corpus
   getting started
   Evaluation of Merit
   Evaluation in Journals
   If an archive announces items, are journals ready to review anything?
   evaluation and effect on community

First report back
Ideally primary data (including introspective and elicited data) that is used in journal articles should be provided with sources and be cited in the best way it can be. The reader should be able to know how the data was created and how to access it.

Documentary linguistics has particular requirements around preparation of re-usable data and the special responsibility of the linguist to provide primary material to support claims of the occurrence of phenomena in the language.

Where possible data should be cited to a source that can be accessible to the reader:
   - this then requires that the journal think about persistence of the source (best if it is housed in a trusted repository rather than a personal webpage which has no guarantee of persistence)
   - the author needs to think about making primary sources available (such as depositing data, for example, field recordings or digitised images of manuscripts in a repository)

If it cannot be cited to a source then a reason should be given for that (e.g. it is elicited and was not recorded)
We need a unified style sheet for citing data that everyone can agree on.

- first a few straightforward improvements that are easy to implement (low hanging fruit)
  - guidelines for elicited data and data from sources
  - making sure everything has proper citation
    - even if it is own native speaker intuition
    - need to differentiate between own data and someone else’s data
      - own, elicited data is new data and makes an empirical contribution which must be clearly and easily identified
  - move next to citing archived data (usually a citation form is provided by the archive)

**Issue of supplementary material and how to cite it.**

Should a journal house it or encourage the author to place it in a suitable repository for access?

Particular issues:
- how to cite web-based data - especially closed sites like Facebook
- Twitter feeds?

**What constitutes the data set?**

A data set is the material on which an analysis relies and so is contextually constituted. It could be made up of a range of data, e.g. the set of particular sentences elicited on the topic; e.g. the story from which the sentences were taken; the media on which the performance of the text is recorded.

A data set may have different definitions depending on the subfield of linguistics, so that a finegrained analysis of a particular phenomenon in English may not require a source if the language is commonly used and examples are not exotic.Child language data in CHILDES has a long tradition of being citable. Corpus data is similarly cited in publications in that field.

**Next steps**
- first draft of guidelines (who will do this?)
- send to editors of relevant journals
- elicit feedback
Second Report Back

Evaluation

Journal reviews of collections as a form of evaluation, treated much as book reviews currently are

From a journal’s perspective all citation should point to the relevant piece of information

A survey article of archives should be included in a journal (these would include DELAMAN, CHILDES ...)

- Process:
  - journals have to first say they will commission reviews of collections and archives
  - then say to repositories to announce collections like publishers announce books
  - editorial board votes and decides what’s worth review
  - go to scholar who created archive and announce that they will be reviewed
    - solicit ideas for desired reviewers

Ask all repositories to get behind this idea and agree to publish quarterly reports of collections announcing what’s new and reviewable. (Maybe OLAC could provide quarterly reports on updates which journals could subscribe to, but an archive would choose the best items to suggest for review and perhaps annotate the collection title with more information)

Clear need for more archives to deal with a wider range of linguistic data (not just LD data)

Co-authoring

not such a big issue for journals who can have as many authors as required. But there could be a role for Journal editors to suggest that additional co-authors be added (note that we are very sympathetic to the idea of including language teachers as co-authors in LD for example).

Citation style guide

- what should journals consider in citation style guide?
  - archive name, file name, time codes, file creation date
    - is deposit date the same as publication date?
    - what date would go with a collection? deposit date?
  - on the archives to make some citation decisions regarding format
  - items are arbitrary units (like a chapter)
    - item 1995 in collection 1997?
    - keep levels of organization separate, cite either item or collection?
when citing items, would need to be able to do in text citations
- time stamps would be easy, like page numbers
- text items, single sentences, how to cite?
  - follow convention of transcription system
  - existing convention of TEXT:LINE/SENTENCE #
- granularity
  - item is what you want, but how do you get from there to collection?
  - show example citation of different data types (word, sentence, media aligned sentence ...)

Suggest a joint session between journal editors and archivists for the next meeting, especially on citation forms for primary data

+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

Raw notes
A+B
Introductions:
- data citation is something that would benefit the field if we had better ways to do it
  - encourage people as journal editors
  - suggest to authors that there are better ways to cite material in their articles
- what is a better way and what are the principles
  - practices are so poor right now that we don't have far to go to get better
  - having citation form in a journal means we have to have a repository
  - but journals, should they send back “bland” materials with suggestions on how to resolve citations
    - suspects it won't go down well
- clarification between distinction between data and data set
- as an editor, it's easy to insist on each piece of data being cited
  - but the question is what constitutes the data set
  - it becomes more complicated regarding archival and accessibility
    - this is a different issue than the examples in a paper
- issue of data being in the form of the elicitation
- if taken from a story, different from the sentence where the story is the data set and the sentence is the data
○ when eliciting, it needs to be a prior convo before going to the field
○ difference between examples collected and how much makes it to a paper
  ■ how much of a data set needs to be represented in a publication?

○ we are inclined to talk about the “weird” stuff, but in the context of the language it might be historical rather than synchronic problems
  ■ encourages getting a lot of data of a single type to make a stronger hypothesis
    ● frequency

○ have an example sentence and you have to cite it from some source

○ what do you do with in text citations?
  ■ need a unified style sheet for citing data that everyone can agree on
  ■ if a repository gives you a citation format, that is what you should use, but there could be a different style in the bibliography
    ● balance citation style and best practices
  ■ ideally would be able to retrieve the data easily using the citations

○ what about unrecorded data points?
○ issues around citing who gave the data
○ working with journal editors would be a good, easy starting point to drive home the importance of citing data, get editors on board
  ■ but the difficulty would be getting individuals of different subfields to agree on citation guidelines
  ■ would have to work out a concrete definition of “data set”
    ● in areas like psycho, would the amount of data be overwhelming?
    ● what do you provide beyond analysis and stimuli?
  ■ where is the raw data?
    ● issues of accountability and verifiability
    ● data as a supplement?
    ● unreadable data sheets filled with mnemonics which make replicating ones own analysis difficult

○ two issues:
  ■ how is the data cited in print?
  ■ how do you manage your own data when you are creating it so it can be useful later?
    ● would need a guide that is accessible
    ● would have to use a set of conventions when developing data

○ problem on style:
  ■ we have to make up guidelines every time when developing data sets
    ● make sure to do it consistently
    ● need to educate students on set styles
Leipzig glossing rules have made a big difference with gloss lines, shorthand

- is this possible in other areas? are there standards that people more or less follow already when setting up excel sheets?

- definitions vary across subfields as well
- the principle is, if its replicable, its not just for others, it is also for yourself
  - write down own guidelines and follow them consistently

  - we want to try to encourage replicability and what flows from that?
  - consistent guidelines
  - explain steps along the way
  - having guidelines would aid in people doing this
    - first step: teaching people, billing it as “better for oneself later”
    - need to focus away from presenting needs altruistically?

- concerns:
  - people pulling data from an investigator’s data and “undercutting” on undone work
    - archived, openly accessible data; rich corpus with annotated portions (usually small portion that has to be done anyway)
      - but the raw data is hard for someone not familiar with the data to be able to do analysis
      - so you add value to the recording with some analysis
      - tendency to not use already archived data, but people want to use the archived data

- appealing to self interest:
  - easier to pull/analyze data later on for the purpose of publications

-----

define the community: readers of the journals, editors and authors

-----

goals

- as editors:
  - where do you put the data?
    - no method to handling supplementary material in digital formats
      - personal webpages, unlabeled data
    - has to be put in an archive
    - who has to house it?
      - journal X in repository Y
      - journal X in allowable data house
      - journal X house on a webpage
        - worst one, unless picked up by an internet archive since webpages are impermanent
        - Internet Archive: will make a fixed HTML
should journals have a role in deciding where data is archived?
  • need for recognized repositories
  • summaries of repositories can be published
  ○ what is publishable and what should exist in a repository?
  ○ deposit through universities (but not in place of archives)
    ■ publications
    ■ could put primary material there as well?
    ■ publically available?
    ■ are libraries open to this? quite likely they will be keen on this
      • then issues of huge video and sound files
      • but page after page of text is easy
  ○ journals make submitters jump through hoops to show they have enough data to back up claim?
    ■ what is the relationship between that data and non-relevant data?
    ■ practically, do you archive data that isn’t relevant to analysis, argument, publication, etc.
  ○ necessity for both experimental and arm chair linguists to be working together on things
    ■ potentially symbiotic relationship
    ■ but while each contribute good, real data, not all of it should be or is even able to be archived
  ○ opportunity to improve elicitation methods
    ■ would be useful for the reader if the steps went through to get the cited data were known
    ■ elicitation is useful as data
    ■ if these records are available (elicitation protocol) it is useful
      • this is a type of metadata, easily implemented
      • can make sure to get explicit descriptions of elicitations
      • also helps move the field forward and allowing scholars to build upon each others work
    ■ footnotes useful for this
  ○ balancing the needs and practices of theoretical and field work
  ○ what can you find out about a language that is not in a grammar?
    ■ potentially limited or shaped by what is trendy at the time
    ■ need to know the status of an elicited example
  • its a good thing, but we can do better
  • a few straightforward ways that are easy to implement (low hanging fruit)
    ○ guidelines for elicited data and data from sources
    ○ move to pushing towards archived data (this is more difficult)
    ○ making sure everything has proper citation
      • even if its own native speaker intuition
      • need to differentiate between own data and someone else’s data
own, elicited data is new data and need to make empirical contributions clear and easily identified

- blanket statements vs every example has a citation
  - good practice?
    - depends on fields or even papers
    - refer to page that the example is on?
    - refer to whole source
      - agreed that this is bad
  - where readers should look for this info
    - always first footnote, first footnote that mentions data
      - need for some consistency that is most useful for the reader
        - first data point may be more noticeable for the reader, since some readers skip the first footnote of a paper
    - refer to whole source
      - agreed that this is bad

what can we offer the field

- guidelines
  - can distill from these notes

what is hindering us?

- inertia
- finding people who are willing to take the time out of their schedules to make a first draft

next steps

- first draft of guidelines
- send to top editors
- elicit feedback

Evaluation of archival materials

- journal point of view
- if someone writes a review article, how do you evaluate a collection or the review
- A small collection of an undocumented language would be far more valuable than a more widely spoken and documented language with a larger collection
- should quantity be that important than quality?
- how do you assess when a collection is “finished”?
can you write a review article before the collection is deemed completed?
can’t work like a book review, would have to behave more like a peer review?
divide it up by natural breaks, PhD collection; Post Doc collection, etc.
- for journals that routinely have book reviews, folding archive reviews into the same category
- editors should encourage and create possibilities for including archive reviews
  - incremental improvement of materials
    - update after update over a long term
    - like reviewing the second edition of a book
- peer review vs published review
  - deciding on what gets reviewed and how to ensure the review gets read
    - get the most significant monographs and have reviewed by people you want to hear opinions from
    - protocol for deciding which authored items are most in demand for reviews
    - protocol for publications could easily be applied to collections
      - need for mechanism to allow for repositories to announce collections
      - do you want to review later versions of collections? you don’t so much for publications
  - Process:
    - journals have to first say they will commission reviews of collections and archives
    - then say to repositories to announce collections like publishers announce books
    - editorial board votes and decides what’s worth review
    - go to scholar who created archive and announce that they will be reviewed
      - solicit ideas for desired reviewers
    - all repositories have to get behind this idea and agree to publish quarterly reports of collections announcing what’s available
    - should consider updates to archives as 2nd editions?
      - count updates with enough distinction to be considered as a newer publications
      - would be up the journals
- distinction between collection and corpus
  - collection can feed a corpus
  - maybe reviews should be of corpora?
    - could argue that collection is the corpus
    - risk: not actually reviewing collection and that might discourage some data from being deposited, when we want everything (even junk)
○ evaluation should consider everything
  ■ including less curated data
  ■ not discourage people from depositing anything

● getting started
  ○ identify how to look at collections
    ■ figure out expectations of collections
    ■ establishment of possibilities in the teaching process
    ■ identify exemplars
  ○ announcing publications/archiving of a corpus
    ■ reviews of exceptional corpora
    ■ what about reporting on the collections of an entire archive to provide a picture/survey of what is in a particular archive
      ● having a starting point of for each repository
    ■ review archives by areal focus, country, etc.

● Evaluation of Merit
  ○ tailoring collections to fit the job market
  ○ would have to increase training in various areas outside of specialized workshops
    ■ how to make databases
  ○ balance between theory and fieldwork

● Evaluation in Journals
  ○ Open access
    ■ if you’re evaluating stuff, it has to be available
    ■ embargo periods (open access after one year)
      ● sometimes you have to pay to make items open access
    ■ degrees of open access
      ● Platinum Open Access = No-author-fee Gold Open Access
      ● Gold Open Access = Journal-led Open Access
      ● Green Open Access = Author-led Open Access (self-archiving)
      ● There’s also Author-pays Gold Open Access
  ○ If an archive announces items, are journals ready to review anything?
    ■ What if published collections are not open access
      ● cannot have a general review if the public cannot see it?
      ● but the knowing of the existence, and how to contact the author for access, may be valuable
      ● but cannot have a review of a collection that is completely closed by a community
      ● could still publish a list of collections received and mark the level of access
        ○ announcements similar to linguist.list
        ○ quarterly blog
What counts as published?
- if password protected, is it published
- what about community that does not want outside access to collections?
- with respect to tenure, if it can be evaluated by the committee, can it count as published?
  - need to have the right people writing tenure letters
    - responsibility of writers to situate the value of an item in a larger context, in cases where the interpretation of “published” is an issue
  - risk of a university level tenure committee not allowing some products
- evaluation and effect on community
  - barriers
    - restricted access to collections
    - what does published mean
    - embargoed collections
  - another way of raising visibility of archives
    - unify community
    - important to include some journals that aren’t LD&C, IJL, etc., if they were to have these sorts of reviews to reach a broader audience
    - like Language
  - Role of archiving in linguistic research and then a review of exemplar collections
    - tasked to a review editor?
    - need to create a category for these
  - short reports vs full reviews
  - language has to decide if it wants to curate announcements of collections and corpora
    - would still be good to have an article on the value and role of archives as a start to open the door for collection reviews and repository announcements

++++++++++++++++++++++
D
- data packaging
  - if a bibliography does not contain references to the sources to a word, it probably should
    - journal editors play a role
  - should there be an appendix with all the references, or what is the appropriate level of granularity
    - related to the issue of data/data set
- journals need to be able to be sure that the data provided allows for verifiability
  - Link to a particular sentence, or to an entire discourse? the sentences aren’t nearly enough
  - link to a single sound file is fine when verifying phonemes
  - level of granularity dependent on the issue being addressed
- what do editors want to enforce regarding archive citation
  - many archives are different, but it has to be a sufficient citation to resolve
  - guidelines/item on the style sheet for how to cite archives despite differences between individual archives
- how do you cite granularity?
  - what sort of information should be required?
  - like a page number? have a timestamp or filename parallel page numbers
- what should journals consider in citation style guide?
  - archive name, file name, time codes, file creation date
    - is deposit date the same as publication date?
    - what date would go with a collection? deposit date?
  - on the archives to make some citation decisions regarding format
  - items are arbitrary units (like a chapter)
    - item 1995 in collection 1997?
    - keep levels of organization separate, cite either item or collection?
    - when citing items, would need to be able to do in text citations
      - time stamps would be easy, like page numbers
      - text items, single sentences, how to cite?
        - follow convention of transcription system
        - existing convention of TEXT:LINE/SENTENCE #
  - granularity
    - item is what you want, but how do you get from there to collection?
- what info do we put in text under an example to cite an item?
  - resolve from the citation to the source
    - archived referencing formats should be able to do this
  - need a formalized linguistic stylesheet
    - elements need to be known
    - author’s name
    - date
    - DOI
      - treat like a book?
      - each paper gets a unique DOI?
    - what else?
○ variety of ways things can be arranged, but we need pointers that are standardized
  ■ need to distinguish between “page” and “example”
  ■ stable distinction between abbreviations
● suggest a joint session between journal editors and archivists for the next meeting
  ○ show example citation of different data types (word, sentence, media aligned sentence …)
● citing unavailable data
  ○ this is the same as citing an unpublished manuscript
  ○ can’t build a whole argument on data that is unavailable
    ■ up to author and reviewer to figure out
  ○ If citing an item in an inaccessible collection, cite in a way that acknowledges the restrictions on the collection
    ■ do archive citation guidelines commonly include info on access conditions?

+++++++++++++++++++++++