Data citation practices in JALL

Felix K. Ameka
f.k.ameka@hum.leidenuniv.nl
What kind of data do authors in my journal use?

- Text & audio corpora
- Experimental (limited)
- Elicited data
- Introspection
- Examples captured in Field notes
How does my journal already handle the citation of data?

- Is citation required?
  - Citation is not required currently but we are moving to at least make it desirable and later mandatory

- Is citation common, even if not required?
  - relatively

- Is there a particular format for citations?
  - From written sources e.g. fiction etc author date page format
  - For elicited data some authors provide consultant name, the stimulus for the elicitation e.g. MPI stimulus BowPed
Example from a paper in process

Footnote 1

- All recordings have been archived in The Language Archive at the MPI for Psycholinguistics, Nijmegen. In this paper, each example contains a reference to the filename of the recording from which it is taken and which can be found in the archive.

- But there is no citation provided to the corpus as such and no hyperlinks are provided

- The file naming system is also not explained in the paper
- bo-mɔɛ-ŋɔ a fua
- c2s-orange-def:foc Afua
- a-ŋa
- c1s .sbj-eat:cfh
- ‘Afua ate [the orange]FOC .’ (elic-foc_100602_SO)

(S0811171_WO)
(kadzidzia_110406_QM)
(funeral_100531-MM-EM)
(conv-rice_110411_3-2)
(elic-adv-placement_110318_SO)
(chiefs-meeting_100619_03)
(pear_100719_Pha-DQ)
Beyond this particular journal, what are the expectations for data citation in my subfield?

(Possible topics include:
- Do researchers in my subfield usually/sometimes/never cite data?  
  - Sometimes
- In print? In presentations?  
  - Sometimes in presentations and the sound embedded with examples
- Is it important in my subfield to make data available to others?  
  - It is important but there is the old culture of the fieldworker having monopoly of the data.
What are the barriers to implementing data citation in my journal?

- Author resilience
- Data is not available in a form that it can be consulted by others
- The role of data verification in the discipline
- Level of granularity of information that is helpful to the interested user
- ?? Publisher
What are the barriers to implementing data citation in my subfield?

- Lack of recognition of the production of data as a legitimate academic scientific endeavour
- Lack of financial support for processing and storage of data (including processing of “legacy” material)
- Lack of data citation culture
What are the potential benefits to implementing data citation in my journal and/or subfield?

- Better accountability
- Potential for verification of analysis
- Improved accessibility of material
- The material in corpora gain value
- The creation of corpora could then become part of the merit system in academic career development
- Binging about a sea change in the discipline
- The citations can be harvested and become part of citation indexes increasing the citation impact of the creator
Additional thoughts

- For these “advantages” to be maximally exploited there has to be different types of information in the citation:
  - The location of the material
  - The creator
  - The file (with information about “genre” of data e.g. elicitation; story; conversation; topic relevant speaker identification etc. etc.)
Data citation practices in the International Journal of American Linguistics (IJAL)

Keren Rice (former editor)
University of Toronto
(with David Beck, current co-editor)
What kind of data do authors in my journal use?

- Textual data
- Elicited data
- Data based on introspection (minor role)
- Data from literature
- Text and audio corpora
- Experimental data
How does my journal already handle the citation of data?

- Citation of sources is required.
- Some authors identify individual speakers, as appropriate.
Beyond this particular journal, what are the expectations for data citation in my subfield?

- Citation of data is of great importance.
- Availability of data depends on a range of conditions, particularly related to preferences of speaker and community.
What are the barriers to implementing data citation in my journal?

- Subject to ethics protocols (both community and university), there should be no particular barriers.
- Linguists have worked extensively on languages, but not made their data easily available, if available at all, making data citation difficult.
- Citation to accurate and available data may be the only way to get data for many languages.
What are the barriers to implementing data citation in my subfield?

- These are the same as those previously noted – ethics protocols of community and university
What are the potential benefits to implementing data citation in my journal and/or subfield?

- Credit/honour to those who contributed the data
- Accurate reproduction of original data
- It is not so clear that reproducibility of data is an benefit given the variation within a language represented in IJAL and the rapidly changing state of many of the languages. The data might not be reproducible, but that does not necessarily mean that the data was wrong.
Additional thoughts
Data citation practices in the Language Documentation & Conservation Journal

Nick Thieberger
University of Melbourne
What kind of data do authors in my journal use?

- Archival records
- Text corpora
- Media via textual index
How does my journal already handle the citation of data?

- normal bibliographic citation
- citation is required
- embedded media
- storing media in DSpace repository
- websites cited as archive.org versions for persistence
Beyond this particular journal, what are the expectations for data citation in my subfield?

- Do researchers in my subfield usually/sometimes/never cite data?
  - pathetically poor take-up of the promise to cite primary data offered by new methods and tools
  - part of our effort is to ensure data is cited in all analysis
  - we need to build corpora that can be cited, and these need to be archived with persistent location and identification, and this will then allow all analytical work to cite its sources

- Is it important in my subfield to make data available to others?
  - LangDoc emphasizes the importance of citation as part of an ecosystem of data management, curation and re-use
What are the barriers to implementing data citation in my journal?

- We do a pretty good job at encouraging authors to cite primary data
- Some authors have not adopted modern methods and so are unable to cite their own data appropriately
- There are not enough archives to provide citability to primary data
What are the barriers to implementing data citation in my subfield?

- Many authors have not adopted modern methods and so are unable to cite their own data appropriately.
- Tools do not always provide an easy means to cite at the level of granularity required.
- Not enough archives to provide citability to primary data.

As of Elan 4.9.0 there is a citation function.
What are the potential benefits to implementing data citation in my journal and/or subfield?

- Many and various
- Replicability of claims
- Encouraging a workflow that
  - results in citability at a finegrained level (e.g. timed chunk of media)
  - relies on citable research data – archived with persistent identification
Additional thoughts

- Building digital object identifiers into archives
- Building reward structures for creation of citable primary data
Data citation practices in Language

Stanley Dubinsky
University of South Carolina
dubinsky@sc.edu
What kind of data do authors in my journal use?

- Untagged or lightly tagged raw data:
  - Textual corpora (Google Books, Twitter, researcher generated corpora, etc.)
  - Sound corpora (telephonic conversational corpora, researcher recorded spontaneous or staged data, etc.)

- Tagged raw data
  - Tagged textual corpora (COCA, COHA, Childes, author collected and transcribed data, etc.)
  - Tagged sound corpora (recorded stories or spontaneous conversations, etc.)
What kind of data do authors in my journal use?

- Elicited, interpreted, or described data:
  - Written or spoken responses to stimuli and other experimental prompts.
  - Responses to targeted questions.
  - Native speaker judgements of constructed items (stimuli).
  - Observations of subject responses to stimuli (fMRI data, eye-tracking data).
  - Sentences (or words) paired with acceptability judgements.
  - Corpus data paired with token counts or frequencies.
  - Items paired with semantic interpretations.
What kind of data do authors in my journal use?

- Analyzed data:
  - Bracketed phrasal or sentential data.
  - Data paired with semantic representations.
  - Morphologically analyzed word forms.
  - Forms paired with phonological structures or tableau.
  - Conversationally analyzed discourse.
How does my journal present data?

- Print/textual presentation of data:
  - Transcriptions of forms utilize standardized Romanization.
  - IPA representation used as appropriate.
  - Interlinear glosses and translation for non-English items.
  - Photographic and glossed presentation of signed data.
  - Non-Romanized orthographies not used.
  - Tables, maps, etc.
How does my journal present data?

- On-line presentation of data supplements:
  - Sound files.
  - Photographic and cinematic presentation of examples
  - Tabular records of elicited responses to stimuli
  - Description of corpora
  - Description of raw data sources
  - Researcher generated corpora
  - Anything that can be served through an internet connection.
  - Data is often presented in a format that omits any metadata or citation information.
How does my journal handle citation of data?

- Within the published article:
  - Textually or photographically presented examples (with gloss and interpretation) are cited if drawn from other publications.
  - Acceptability ratings paired with examples attributed to author, unless explicitly stated otherwise.
  - Tables and figures presenting data descriptions (e.g. distributions, frequencies, etc.) attributed to author if not cited from another source.
  - Formal representations paired with analyzed examples assumed to come from author, unless cited otherwise.
  - Sources of raw data often noted in print.
What are the expectations for data citation in my subfield (syntax)?

- **Minimal**
  - In-text examples are normally given with a source, if they were taken from other published work.
  - Researchers own constructed or elicited examples are rarely cited or sourced.
  - In more recent experimental syntax work, subjects’ responses are quantified and results reported. Metadata (subjects’ age, origins, education, etc.) are rarely taken into account, except in the most general terms (native/non-native).
What are the barriers to implementing data citation in my journal?

- Unclear what constitutes “data” in many cases.
- Unclear what needs to be cited, and how to do so.
- Reluctance on the part of some researchers to giving others access, and “doing other people’s work for them”.
- No guidelines from the journal on what is expected.
- No consistent means of archiving and serving data.
What are the barriers to implementing data citation in my subfield?

- Unclear what constitutes “data” in many cases.
- Unclear what needs to be cited, and how to do so.
- Reluctance on the part of some researchers to giving others access, and “doing other people’s work for them”.
- Data are often collected informally and not recorded in a manner that would lend itself to citation.
- The form and source of “data” can vary widely and wildly.
What are the potential benefits to implementing data citation in my journal and/or subfield?

- Would lead to more consistent and reliable results.
- Would allow for replication of studies and coherent extension of research results.
- Would result in fewer re-inventions of the same wheel, and would help to insure consistent forward progress.
Additional thoughts

• Archiving formats need to be reliable and stable.
• Metadata fields, values, format, and expectations need to be established.
• Journals should have a uniform means of hosting supplemental data and set out uniform requirements for data citation.
Additional thanks

- I am grateful to the editorial team of language (its Editor, Associate Editors, and editorial assistants) for the input that allowed me to contribute to this workshop.

- Special thanks go out to: Gene Buckley, Greg Carlson, Megan Crowhurst, Grant Goodall, and Line Mikkelsen.

- Very extra-special thanks go out to Hope Dawson, who graciously provided a broad range of actual examples of data presentation and citation on which these observations are based.
Data citation practices in Natural Language & Linguistic Theory

Julie Anne Legate
jlegate@ling.upenn.edu
What kind of data do authors in my journal use?

- Elicited data
- Naturally-occurring recorded data
- Data from introspection
- Data from secondary sources
- Text & audio corpora
- Experimental results
- Data found online
How does my journal already handle the citation of data?

- Citation of data source is standard

- Citation may take several forms, including:
  - Author’s native speaker intuitions
  - Author’s field notes
  - Secondary source
  - NAME personal communication
  - Archive
  - Corpus
  - URL

- Exceptions:
  - Data considered uncontroversial or established
  - Data from author’s native speaker intuitions
Beyond this particular journal, what are the expectations for data citation in my subfield?

- Citation is standard and expected
- Researchers may fail to provide citations for author’s native speaker intuitions
- Researchers may fail to provide citations for data deemed established or uncontroversial
- Citations are not always of maximum utility to subsequent researchers
What are the barriers to implementing data citation in my journal?

- **Space**
  - for multiple tokens of example types
  - for embedding context of example
  - for full list of experimental stimuli
  - for complete data leading to a claim/generalization

- **Partial solution: Electronic Supplementary Material**

- **Transience of URLs**
What are the barriers to implementing data citation in my subfield?

- Need for standardized data repository
- Need for standardized format for data annotation
- Time and effort required to annotate and archive data
What are the potential benefits to implementing data citation in my journal and/or subfield?

- Replicability

- Required explicit citations lead authors to consider data collection more carefully

- Subsequent researchers may access fuller data set, to reanalyse or to analyse with distinct questions in mind
Additional thoughts

- When is citation pedantic versus useful?
- How to address citations that are of limited use to later researchers?
- How to value efforts in data annotation and archival for tenure and promotion?