Reproducible Research and the Americanist Tradition in Linguistics

Andrea L. Berez-Kroeker
University of Hawai'i at Mānoa
andrea.berez@hawaii.edu

19th Workshop on American Indigenous Languages
UC Santa Barbara
May 7-8 2016

Moving toward Reproducible Linguistics Research

I. Values of linguistic fieldwork in the Americas
→
Values of Language Documentation

II. Records of American Indigenous languages and Endangered Language archiving

III. Reproducible Research Movement in science

IV. Reproducible Research in Linguistics

Values of linguistic fieldwork in the Americas: Foundation in Boas

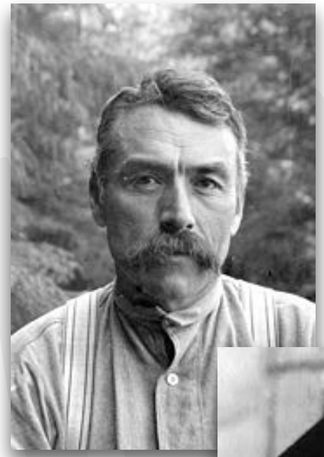
- Modern documentary linguistics has roots in American Indigenous language tradition
  (Woodbury 2011, Rosenblum & Berez 2010)

- Franz Boas (1858-1942): 4 important tenets

- 1. Charter for ethnography: linguistics at center
  - *Culture itself* contained in oratory, narrative, verbal art, ritual
  - To know how to use a language is to know a culture
  - No division between language knowledge and language use

Values of linguistic fieldwork in the Americas: Boasian era

- 2. First theorization of documentary corpus: Boasian Trilogy
  - Naturalistic texts a central component of holistic description
  - Bemoaned limits of dictation
  - Better: texts written by Native speakers
- 3. Role of speakers as documenters
  - Trained and credited
- 4. Longterm interaction with language community
  - Better for the ethnographic record than brief encounters

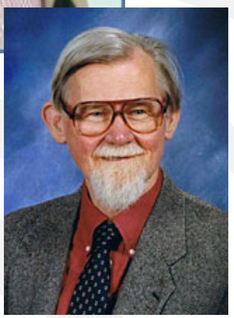


Values of linguistic fieldwork in the Americas: 20th century

- Early 20th c. - European structuralism de-theorized the documentary corpus
  - Separated generalizations about linguistic structure from examples of language use
  - Linguists' work now article-length, not corpus-sized
- Yet Americanists continued Boasian-style descriptive work
  - Sapir, Bloomfield, Haas, Chafe, Krauss, Bright, Shipley, Golla…
- Mid 20th c. - Chomskyan revolution further separated *linguistic performance* from *linguistic competence*
- But some pockets of research still put *language use* at forefront
  - 1960s - Hymes "Ethnography of speaking"
  - 1970s-80s - Rise of Functionalism







Values of linguistic fieldwork in the Americas: 20th century

- 1980s-90s -Endangered Language crisis
  - Brings heightened concern, agency
  - Calls for documenting ELs before they disappear:
- "…lest linguistics goes down in history as the only science that presided obliviously over the disappearance of 90% of the field to which it is dedicated." (Krauss 1992:10)

Values of Language Documentation worldwide

- 1990s-2000s - charter for Language Documentation:
  - Centrality of long-lasting records of language in use in many situations
  - Importance of transcripts and translations that are widely useful
    - A-theoretical, interpretable
  - Recognition of collaboration between all stakeholders:

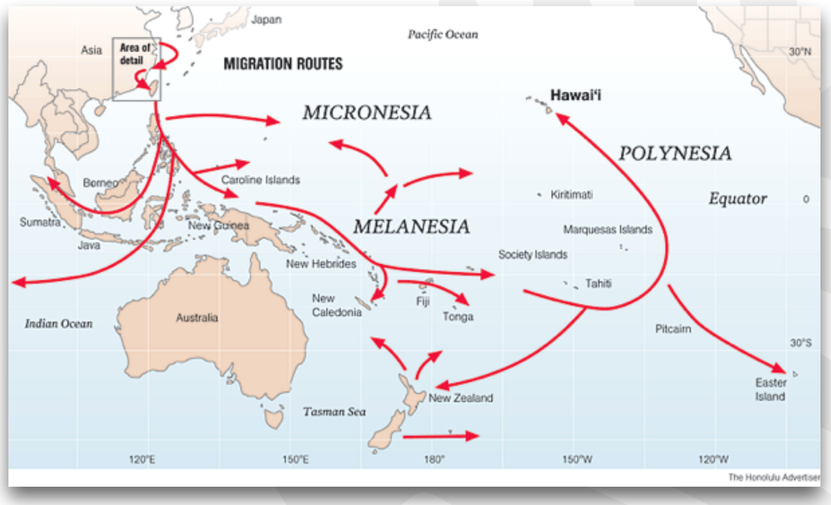
"Humans experience their own and other people's languages viscerally, and have differing stakes, goals, and aspirations for language records and language documentation" (Woodbury 2011:159).

Values of Language Documentation worldwide

- These are very much *Americanist* concerns
  - Big question has been “How do these languages work? How are they used?”
  - Description based on rich examples has been primary activity
  - Collection of grammatical structures in situ (texts) has guided research
  - Time commitment, personal relationships essential

Values of Language Documentation worldwide

- Contrast with work in Oceania
  - Big question has been “How did they get there?”
  - Historical linguistics to determine peopling of the Pacific has been primary
  - Collection of lexical records more common than collection of texts
  - Quick survey work essential



Moving toward Reproducible Linguistics Research

I. Values of Linguistic fieldwork in the Americas
→
Values of Language Documentation

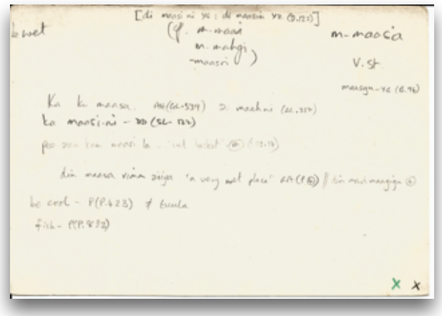
II. Records of American Indigenous languages and
Endangered Language archiving

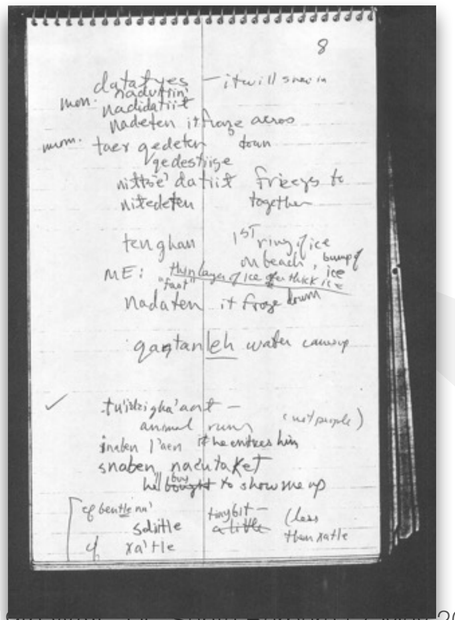
III. Reproducible Research Movement in science

IV. Reproducible Research in Linguistics

Records of American Indigenous languages

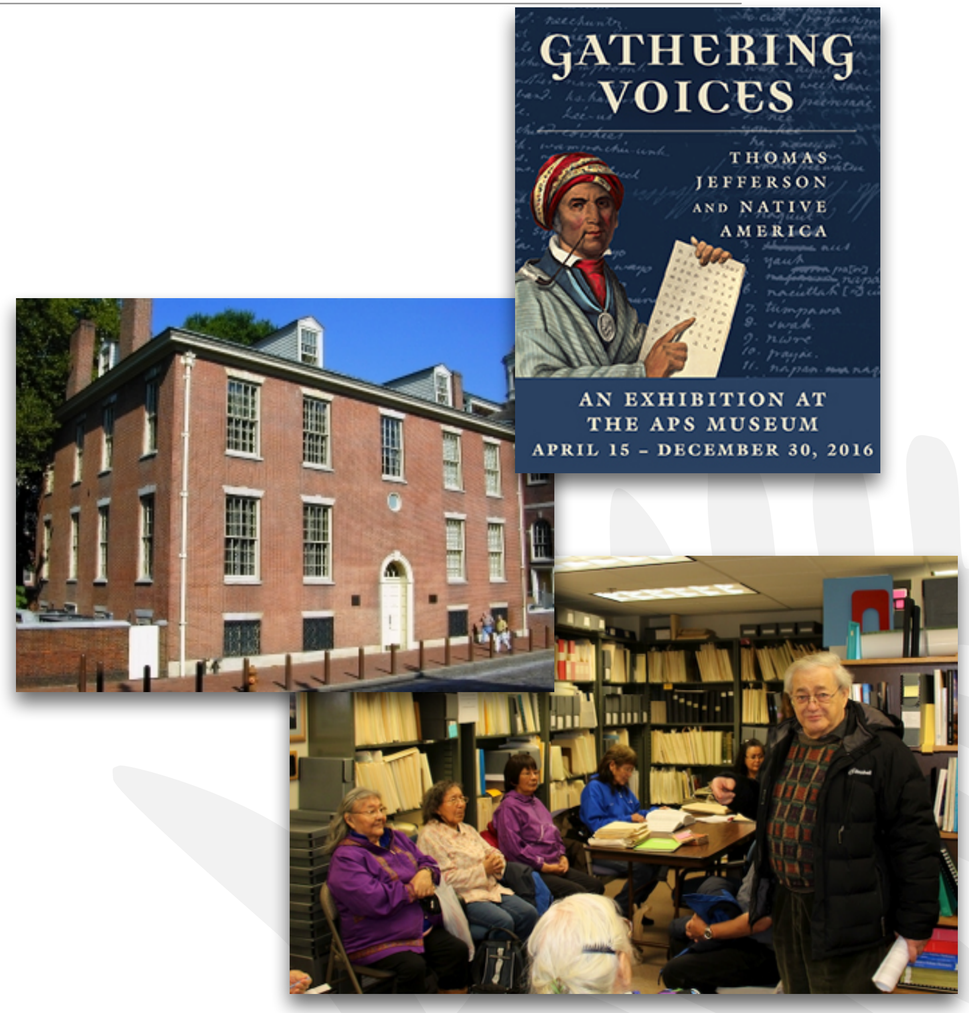
- Types of records (Golla 1995)
  - Lexical lists
  - Texts
  - File slips
  - Field notes
  - Sound recordings





Records of American Indigenous languages

- Repositories
  - American Philosophical Society
  - National Anthropological Archives
  - UC Berkeley
  - Jacobs Collection
  - Alaska Native Language Archive
- Also: IJAL as a “published repository”



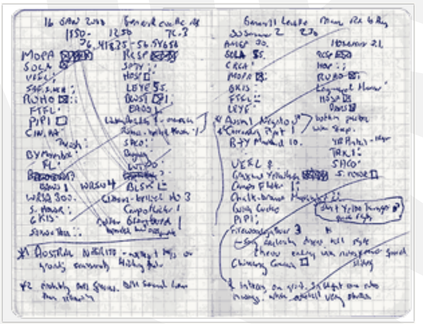
  - All physical archives for analog materials
    - Paper, tape, cards, cylinders
    - Informed by archival/library science

Development of Endangered Language archiving

- Endangered languages and endangered data
- The more advanced our media become, the more ephemeral they are
  - Hard drives: 5 years <
  - CDs/DVDs: 10 years <
  - Cassette tapes: 30 years <
  - Paper: 100-200+ years <
  - Stone tablets: ∞
- Media degrade, devices unavailable
  - (e.g. 5.25” floppies, Zip disks)

Development of Endangered Language archiving

- Digital language archiving arose in response
- But mostly by field linguists, not trained archivists
  - Legacy analog materials being digitized
  - New materials born digital
- Digital language archiving developed mostly independent of archival sciences. (Henke & Berez-Kroeker To Appear)



Endangered Language archiving

- Since 2000 we've seen
  - Development of standards for the creation and preservation of digital records of endangered languages (E-MELD)
  - Development of endangered language archiving consortia (OLAC, DELAMAN)
  - Outreach to educate practitioners
- All *inside* language documentation
  - Little discussion with other fields of linguistics
  - Little discussion with other social sciences outside of linguistics
  - Little discussion with archival/library scientists

I. Values of Linguistic fieldwork in the Americas
→
Values of Language Documentation

II. Records of American Indigenous languages and Endangered Language archiving

III. Reproducible Research Movement in science

IV. Reproducible Research in Linguistics

I. Values of Linguistic fieldwork in the Americas
→
Values of Language Documentation

II. Records of American Indigenous languages and Endangered Language archiving

III. Reproducible Research Movement in science

IV. Reproducible Research in Linguistics

Reproducible Research Movement in science

Good scientific research is *replicable*

Recreate a controlled study >

New data >

[Dis]confirm previous results

Reproducible Research Movement in science

- Some studies can't be truly replicated
  - Behavioral research, like linguistic studies
  - The factors are too hard to control for
- **Reproducible research** instead
  - Reuse of another's data > same or different conclusions

Reproducible Research Movement in science

- Comes from computer science
- "The product of academic research is **the paper and the full data** so that claims can be reproduced."
  (http://biostatistics.oxfordjournals.org/content/10/3/405.full)
- Article + Code + Software

Reproducible Research Movement in science

- Linguistics also values reproducibility!

…but we don't often make it explicit.

Open Science Project:

(Dan Gezelter. 2009. http://www.openscience.org/blog/?p=312)

If a scientist makes a claim that a skeptic can only reproduce by spending three decades writing and debugging a complex computer program that exactly replicates the workings of a commercial code, the original claim is really only reproducible in principle.

If a linguist makes a claim that a skeptic can only reproduce by spending three decades working in the same language community in the same sociolinguistic and fieldwork conditions, the original claim is really only reproducible in principle.

Our view is that it is not healthy for scientific papers to be supported by computations that cannot be reproduced except by a few employees at a commercial software developer. […] It may be research, and it may be important, but unless enough details of the experimental methodology are made available so that it can be subjected to true reproducibility tests by skeptics, it isn't Science.

–Modified from Dan Gezelter, The Open Science Project

Our view is that it is not healthy for linguistic papers to be supported by examples that cannot be reproduced except by doing one's own fieldwork. […] It may be research, and it may be important, but unless enough details of the utterances in context are made available so that it can be subjected to true reproducibility tests by skeptics, it isn't Science.

–Modified from Dan Gezelter, The Open Science Project

On valuing reproducibility

- Prominent in the language documentation literature:
  - Himmelmann 1998
  - Thieberger 2009
  - Himmelmann 2006:6

- …but relevant across all fields of linguistics:
  - Thomason 1994, about checking data in *Language:*

On valuing reproducibility

"[…] so frequently, in fact, that the assumption that the data in accepted papers is reliable began to look questionable […]" (Thomason 1994: 409)

"The advice I've offered here is simple: always consult primary sources; use sources with care; consider all relevant data; and provide detailed information about sources of data and methodology of data collection." (Thomason 413: 409)

How are we doing?

- Berez-Kroeker, Gawne, Kelly & Heston (subm.).
- Surveyed linguistics publications for how well we link back to the underlying data.
- Four questions:
  1. Where does our data come from?
  2. What kind of data are we using?
  3. Where is the data now?
  4. Are we citing our examples? If so, how?

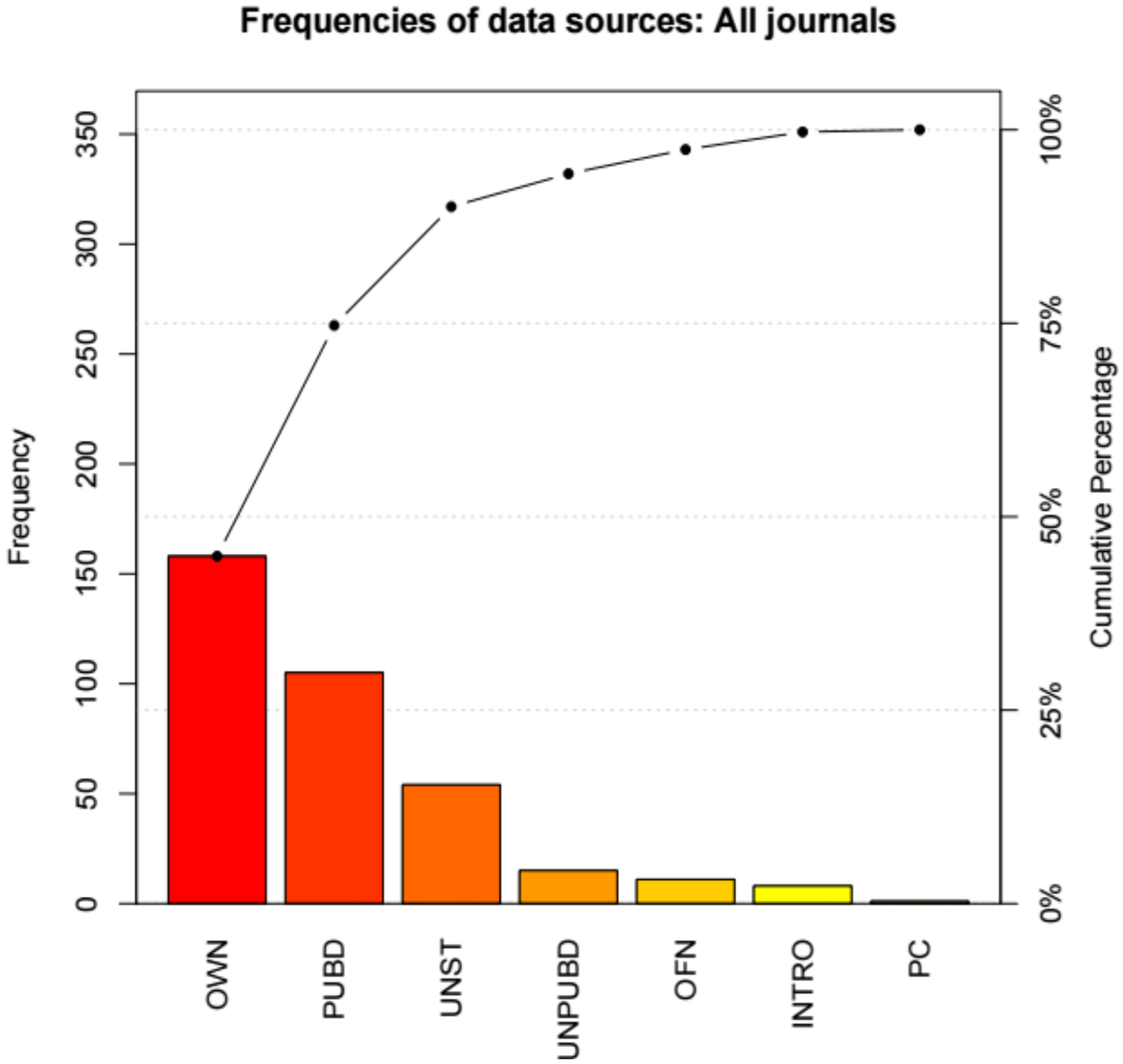
How are we doing?

- 371 publications
  - 100 grammars (50 published, 50 dissertations)
  - 271 journal articles from 9 journals
    - Areal spread: *IJAL, Oceanic Linguistics, Linguistics of the Tibeto-Burman Area, J. of African Languages & Linguistics*
    - Subfields: *J. Second Language Acquisition, J. Sociolinguistics*
    - Theoretical persuasion: *Natural Language and Linguistic Theory, Studies in Language*
    - Top J: *Language*
- 10 year span 2003-2012

1. Where does our data come from?

- OWN: data collected by author
- PUBD: published data
- UNPUBD: unpublished data collected by someone other than the author (excluding fieldnotes)
- INTRO: introspection
- OFN: other person's fieldnotes
- UNST: source of data unstated

1. Where does our data come from?

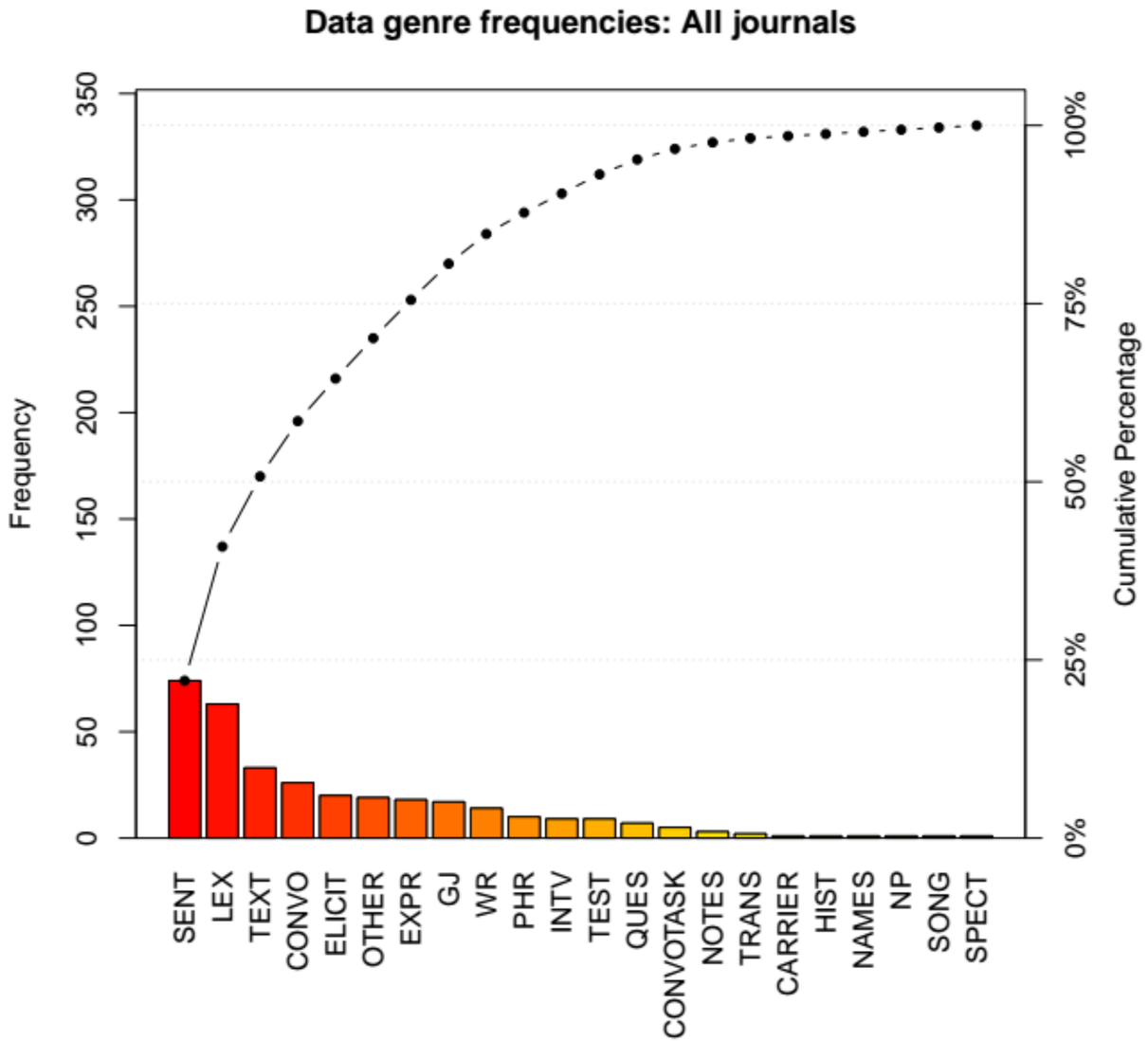


- Most data come from authors' own research ~ 50%
- Followed by published data
- Followed by...unstated

2. What kind of data are we using?

- CARRIER: data in a carrier sentence
- CONVO: conversational data (natural)
- CONVOTASK: conversational task (eg acquisition studies)
- ELICIT: elicitation
- EXPR: experimental
- GJ: grammaticality judgments
- HIST: historical data (eg correspondence sets)
- INTV: interviews
- LEX: lexical items/words
- NAMES: names
- NOTES: own fieldnotes
- NP: noun phrases
- PHR: other phrases
- QUEST: questionnaires
- SENT: sentence data (broadly defined)
- SONG: songs
- SPECT: spectrograms
- TEXT: texts (broadly defined)
- TRANS: translation tasks (eg acquisition studies)
- TEST: tests in a school environment
- WR: written data (eg newspapers)
- OTHER: other

2. What kind of data are we using?

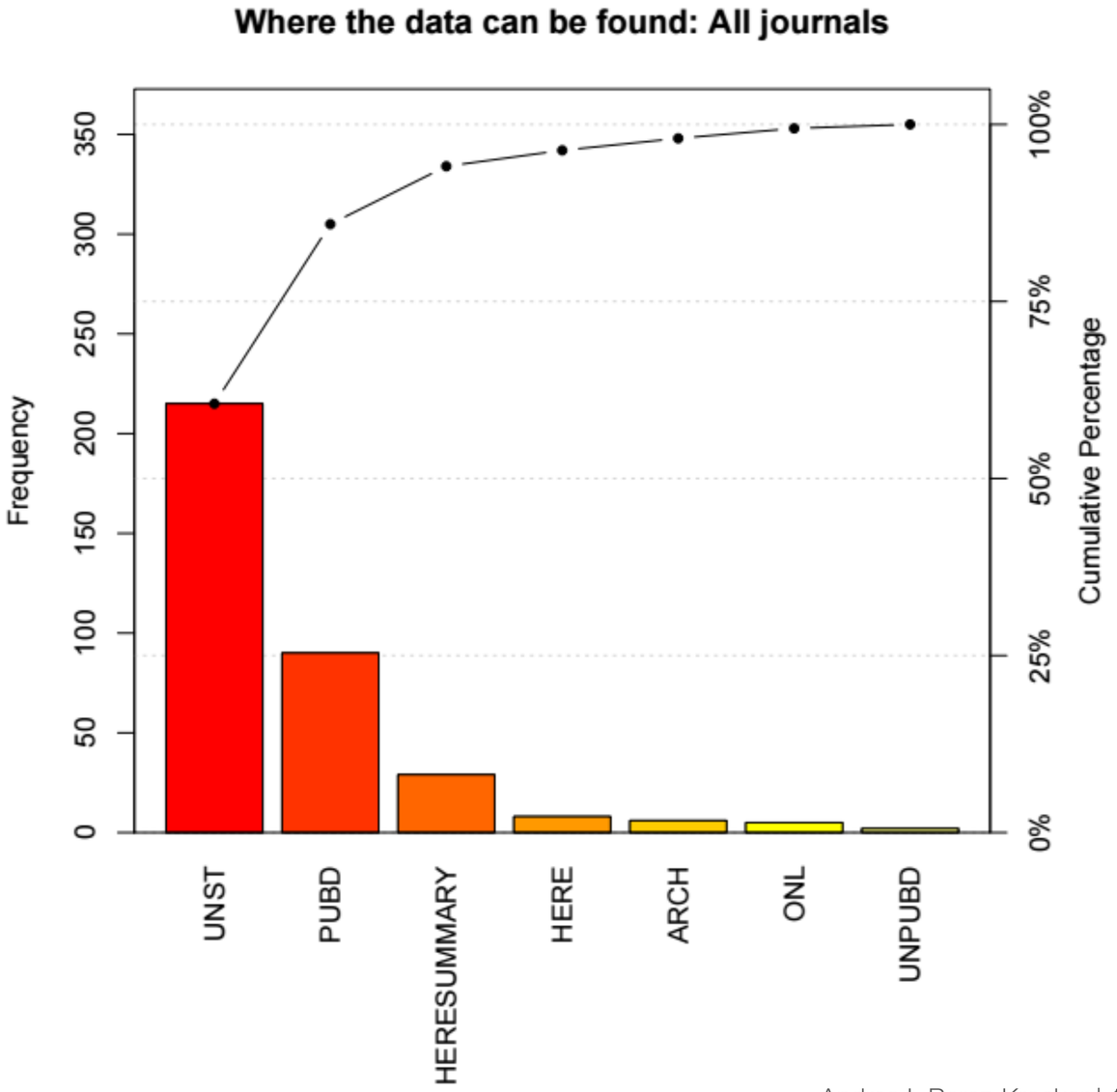


- Sentences
- Lexical items
- Texts

3. Where the data is now?

- ARCH: archived in institutional repository
- PUBD: published
- HERE: article contains the primary data
- HERESUMMARY: data summarized in the article (stats, graphs, tables)
- ONL: online (website or other non-archive)
- UNST: location of data not stated

3. Where the data is now?



- Mostly we don’t know!
- “Published” a distant 2nd

4. Are we citing our data? If so, how?

- STD: citation appears in “standard” format for publications

```
Wari’, Chapacura-Wanam
mo    ta          pa’   ta’              hwam   ca,
cond  realis.fut  kill  1sg:realis.fut   fish   3sg.m
mo    ta          pa’   ta’              carawa    ca
cond  realis.fut  kill  1sg:realis.fut   animal    3sg.m
‘Either he will kill fish or he will hunt.’
```

(Everett and Kern 1967:162)

(example from SL (Mauri 2008:23)

4. Are we citing our data? If so, how?

- AUTHPG: author + page no.

```
ma  han-ac-en    [ah-ic-0      cab-e]
neg eat-pot-b1   dawn-mcmp-b3  ground-ter
‘I have not eaten since it dawned.’
                                    (Coronel:91)
```

(example from IJAL (Yasugi 2005:27)

4. Citation conventions used in examples

- BIBLE: example from Bible, usually book + chapter + verse

```
Tuiy-ul  ganu       giLiji  ibi-l-a   a    abric-il-o;
exit-pa  insidein   people  dem-c-prox and  let-pl-pl
‘Keep away from these men and leave them alone.’
                                          (Act 5:38)
```

(example from JALL (Schadeberg & Kossman 2010:88)

4. Citation conventions used in examples

- CODEEC: citation is a code that is explained by author

So the buggies [bugíz] came out. [BN T3P12]

(endnote explains “[t]he code [BP T3P12] means speaker BN, tape 3, transcription page 12.”)

- (example from JS (Brown 2003:21, note 9)

4. Citation conventions used in examples

- CODEUNEX: citation is a code that is not explained

```
Dijokoti.
pt:take
‘(I) took (it).’ (107:936)
```

- (example from SL (Ewing 2005:100)

4. Citation conventions used in examples

- INIT: citation appears as speaker's initials only

```
Mapuche  mie-kawell-la-y-ngün.
Mapuche  have-horse-neg-ind-3pS
'The Mapuche do not own horses.'          (JA)
```

- (example from LANG (Baker et al. 2005:145)

4. Citation conventions used in examples

- LANG: citation appears as language name only

```
Words for ‘six’ in Eastern Miwok Languages
Northern Sierra Miwok      tem:ok:a
Central Sierra Miwok       tem:ok:a
Southern Sierra Miwok      tem:ok:a
```

- (example from IJAL (Blevins 2005:90)

4. Citation conventions used in examples

- LIST: article contains a list or table of sources used

- Example: Bender et al. (2003:9, note 2) OL article on Proto-Micronesian:
- footnote list of all the published dictionaries from which cognate forms are taken.
- Sources are listed by author's name and year,
- are found in full citation in the bibliography of that paper.

4. Citation conventions used in examples

- PC: citation appears as personal communication

```
kwa  lút wa?  s-náẃ-lx-s
and  neg  spc  nom-3sg.run-aut-3poss
‘But it didn’t run’      (Kinkade, p.c., 2011)
```

- (example from IJAL (Davis 2005:5)

4. Citation conventions used in examples

- SPKRAGEDIAL: citation appears as speaker's name + other demographic info

```
[T]here are times when I get stuck, and
probably all my grammar is wrong, but I can —
yeah, I can manage.
```

(Rita, f27)

- (example from JS (Chand 2011:17)

4. Citation conventions used in examples

- SPKRANON: citation appears anonymized speaker ID

- Example: Maddieson et al. 2009 (IJAL):
- M1, M2, M3, F1, F2, F3...

4. Citation conventions used in examples

- SPKRPAGE: citation appears as speaker's initials + numerical code
- Code most likely a portion of a corpus
- May or may not be explained

```
tā bǎ   nánháir   jiào-zhù
3   mom boy       call-stop
~"he call-stopped the boy" (LY:3)
```

- (example from SL (Post 2007:129)

4. Citation conventions used in examples

- SPKRTITLE: citation appears as speaker's name + title of a narrative

```
kwa?  ?íca    lut     l     cəl'án    ta?-ntitiyáx
conj  dem     neg     loc   Chelan    exist-Chinook.salmon
'And in Chelan there are no salmon.' (Friedlander: Coyote)
```

- (example from IJAL (Mattina 2006:107)

4. Citation conventions used in examples

- STATEMENT: textual statement in body of article explaining sources for numbered examples

Example: Zanuttini 2008:186 (NLLT)

- “[...] Example (2a) is from Hamblin (1987), the others from Potsdam (1998).”

4. Citation conventions used in examples

- TITLE: citation appears as the title of the story or conversation it was taken from

```
[…]
83    kyoo  desho?
      today COP
‘The day when they cook sukiyaki is tomorrow, and the day
when they bring something [to us] is today, right?
(Broccoli)
```

- (example from SL (Takara 2012: 95)

4. Citation conventions used in examples

- TITLELINE: citation appears as the title of the story + numerical code

```
moso    maezo    aut’ucu  to mo con-ci        fo’kunge.
aux.av av.also   raise    obl aux.av one-rel frog
‘They also kept a frog.’                      (Frog 1:3)
```

- example from OL (Huang & Tanangkingsing 2011:95)

4. Citation conventions used in examples

- URL: citation appears as internet URL

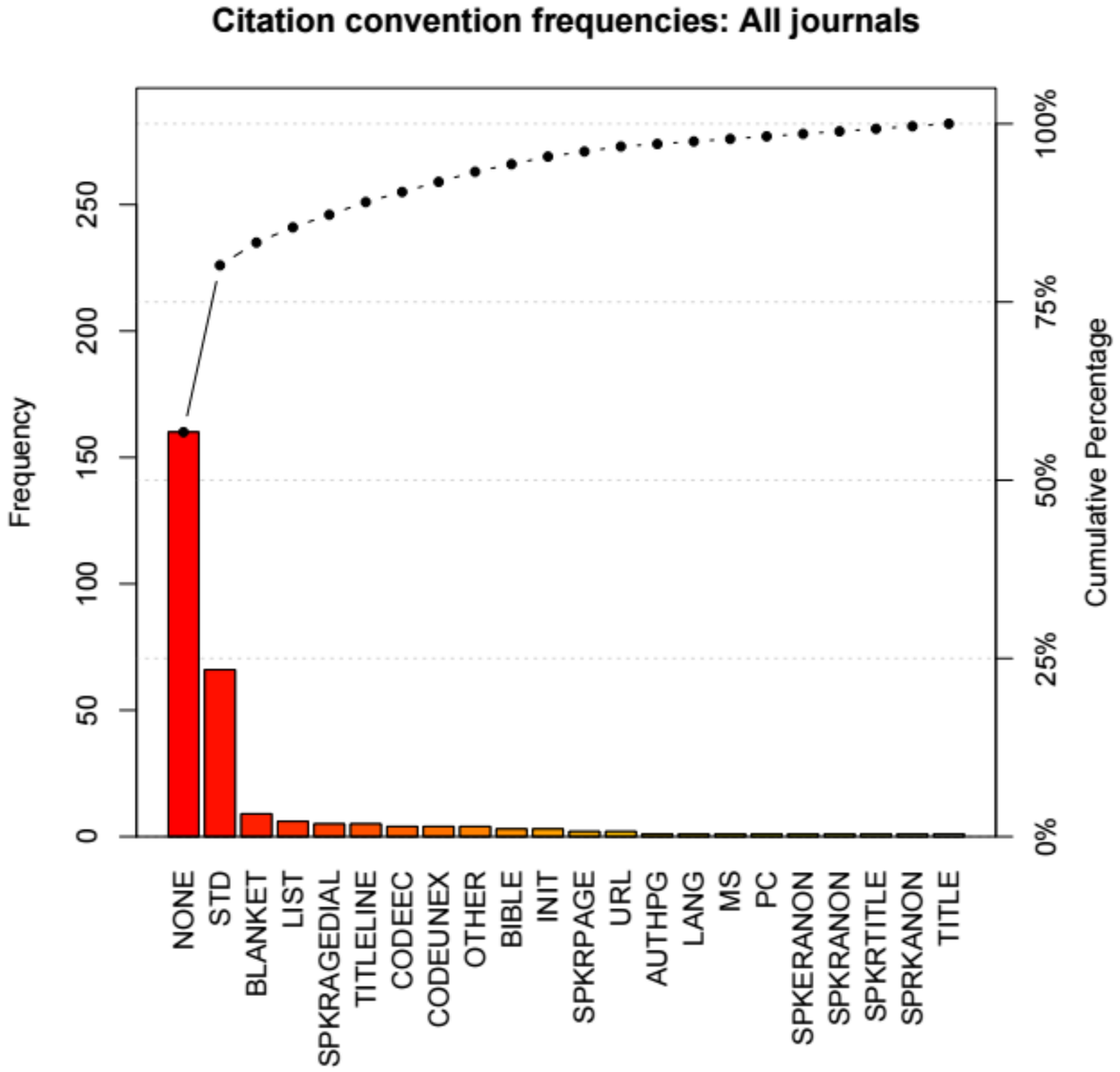
```
Eight Deadly Sins of Web 2.0 Start-Ups […]
Happinessless: Your start up has no future if you
are not happy.

(http://www.slideshare.net/imootee/eight-deadly-
sins-of-web-20-startups/)
```

- (example from LANG (Plag & Baayen 2009:115)

4. Citation conventions used in examples

- MS: citation appears as standard reference to unpublished manuscript.
- NONE: author did not include any form of citation
- NA: article did not contain numbered examples
- OTHER: other practice not easily classifiable here

4.Citation conventions used in examples: all



- Again, mostly nothing.
- “Standard” is a distant 2nd

Overall results

- Inconsistent citation of data sources in publications
  - Most authors do not cite data
  - Except from published paper sources
- Authors' own data is the most common source
  - Not cited!
  - Not archived!

Moving toward Reproducible Linguistics Research

I. Linguistic fieldwork in the Americas
→
Language Documentation

II. Value of records of American languages
and
Endangered Language archiving

III. Reproducible Research Movement
in science

IV. Reproducible Research
in Linguistics

Reproducible Research in Linguistics

- NSF Science of Science & Information Policy (SciSIP):
- Supporting Scientific Discovery Through Norms and Practices for Software and Data Citation and Attribution
- Science now “more open” but data creation not rewarded
- Seeks to:
  - Develop novel citation methods
  - Promote standards of academic credit
  - Citation patterns that include *roles*
    - (eg., “data provider,” “data analyzer,” “computational modeler”)

Reproducible Research in Linguistics

- “Developing Standards for Data Citation and Attribution for Reproducible Research in Linguistics”
- 2015-2017
- Collaboration between four institutions
- Three workshops
  - Working Groups → Task Forces
- Deliverables
  - LSA Panel
  - Position paper
  - LSA Resolution

Reproducible Research in Linguistics

- Grass-roots changes across entire discipline
- Culture of publishing analyses *and linking to preserved data*
- Linguistics can become more data-driven

Workshop 1: September 2015, Boulder

- Working Groups of stakeholder communities
  - Journal editors
  - Archivists
  - Information Technology / Big Data
  - Interested OWLs (=Ordinary Working Linguists)
- Identified specific goals and homework assignments
  - Research current data citation practices
  - Research potential repositories
  - Survey of data citation and the job market
  - Award for archival data set (DELAMAN)
  - Update Open Language Archives Community (OLAC)

Workshop 2: April 2016, Austin

- Task Forces
- Evolved from Working Groups to address key issues:
  - Task Force: Principles and Guidelines
  - Task Force: Attribution for Academic Credit
  - Task Force: Education and Outreach
  - Task Force: Citation Formats and Stylesheet
- Now working on a Position Paper

Task Force: Principles & Guidelines

"Linguistic data are important resources in their own right and represent valuable assets for the field. They need to be *documented, preserved, attributed,* and *cited*.

Responsibility is shared by researchers, data stewards, institutions and funding bodies."

Task Force: Attribution for Credit

- How can we incorporate into hires and promotions?
- Three approaches:
  - Precedent via LSA Resolutions:
    - Cyberinfrastructure
    - Scholarly Merit of Language Documentation
  - Assessing value
    - Metrics for quality
  - Education
    - Empowering applicants, T&P committees

Task Force: Education and Outreach

- Spread the word about Reproducible Research
- Educate ourselves about ethical data management
  - Training through CoLang, LSA institute
- Encourage a culture of responsible data sharing
  - Social media, brown bags, salons, workshops, etc.

Task Force: Citation Formats and Stylesheets

- Make use of persistent identifiers in articles, books
- Show granular datum in its larger context, people and their roles.
- Suggestions:

People. Date. *Title*. Repository, granularity. DOI.

Dilu, Muguwa (speaker, transcriber) & Andrea L. Berez-Kroeker (author). 2013. Kuman Language Documentary Corpus. Kaipuleohone Digital Language Archive, items 001-006. https://scholarspace.manoa.hawaii.edu/handle/10125/29514.

People. DOI, Location/Timestamps.

((Muguwa Dilu (speaker). https://scholarspace.manoa.hawaii.edu/handle/10125/29554, 00:01:35.67-00:01:48.55.))

Reproducible Research is inherently *ethical*

- RR allows everyone who contributes to get proper credit
  - Speakers
  - Translators
  - Assistants
  - Teachers
  - Statisticians
  - Programmers

Reproducible Research is inherently *ethical*

- What about confidential or sensitive language records?
  - RR still allows some records to remain confidential at the level of the archive
    - Avoids “locking everything up” with no exceptions
    - Researchers need to think about the archiving plan *now*
    - (Most legacy material gets locked up because nobody came up with a plan)

Moving toward Reproducible Linguistics Research

I. Linguistic fieldwork in the Americas
→
Language Documentation

II. Value of records of American languages
and
Endangered Language archiving

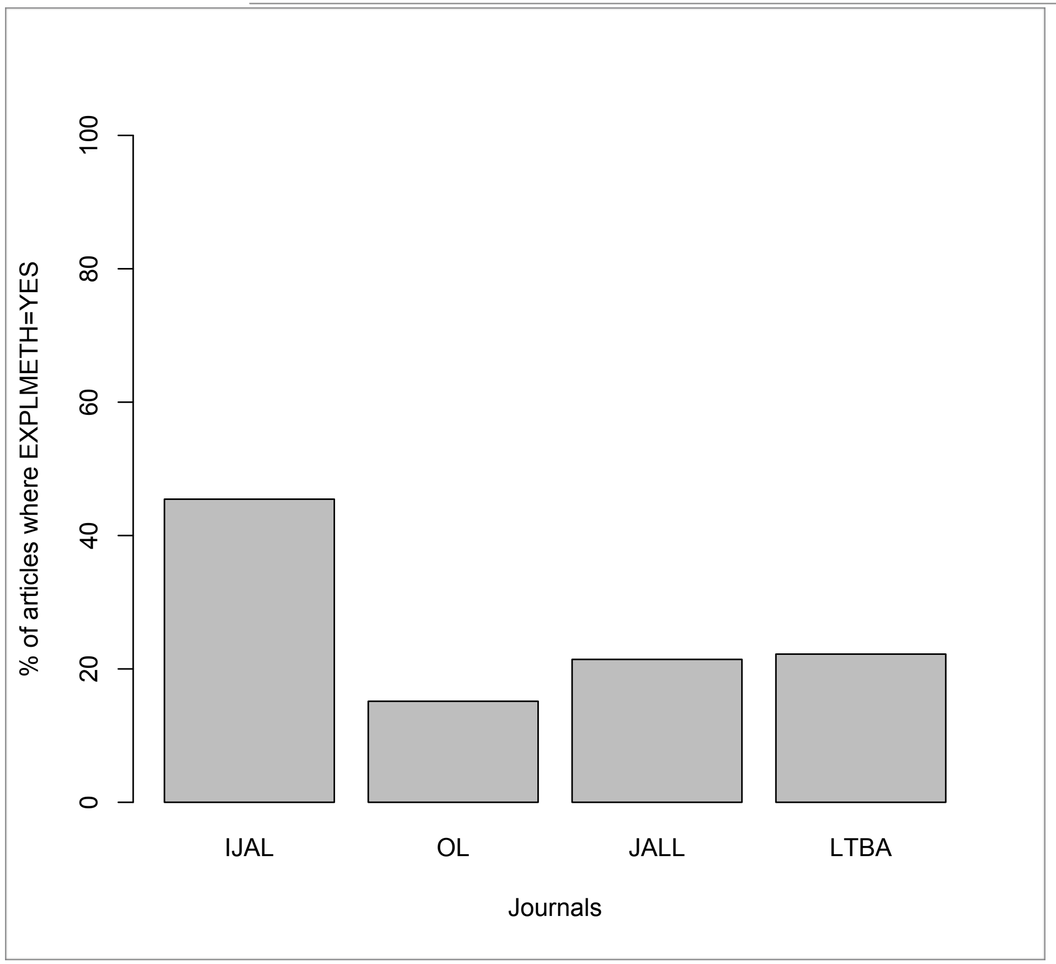
III. Reproducible Research Movement
in science

IV. Reproducible Research
in Linguistics

What is the Americanist connection?

- Reproducible Research echoes the values of Linguistic Fieldwork in American Indigenous Languages.
- Task Force Principles & Guidelines values:
  - Data and methods must be documented
  - Records must be preserved
  - Data must be attributed properly
  - Data must be cited well

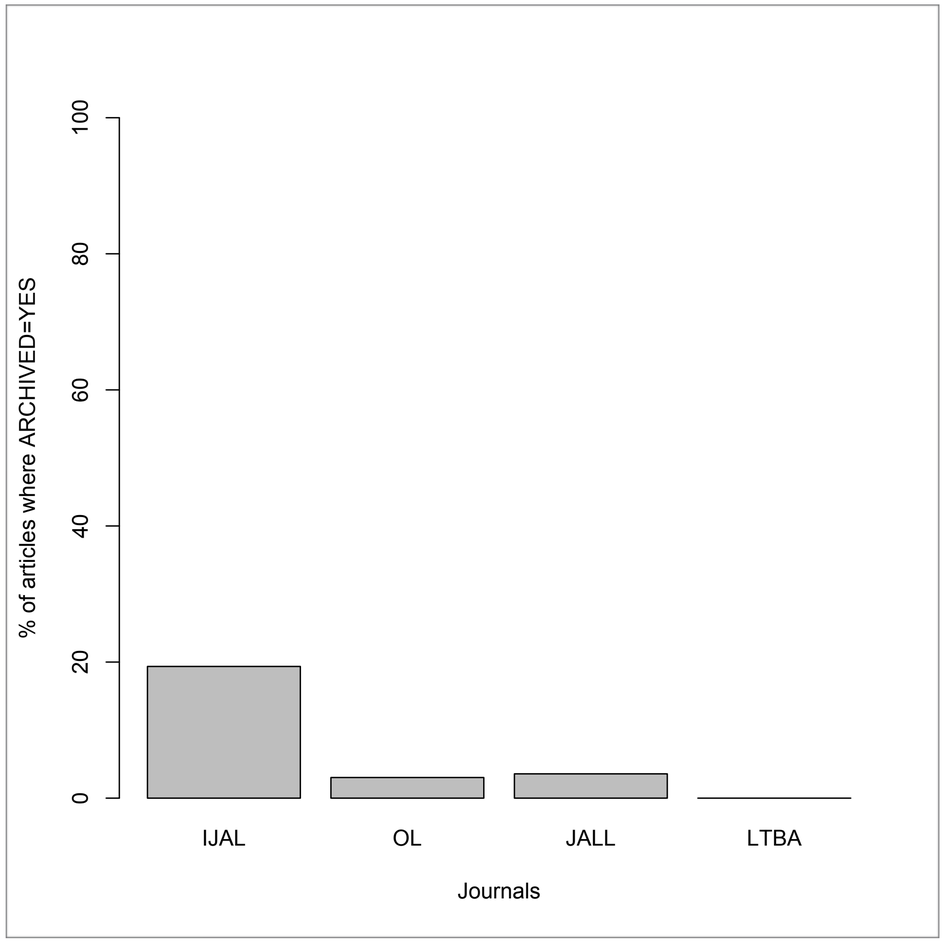
Value: Data, methods should be well documented



- IJAL authors…
  - Describe data collection methods,
  - Describe software, hardware, other tools,
  - Describe their fieldwork situations

…more often than in other areal journals.

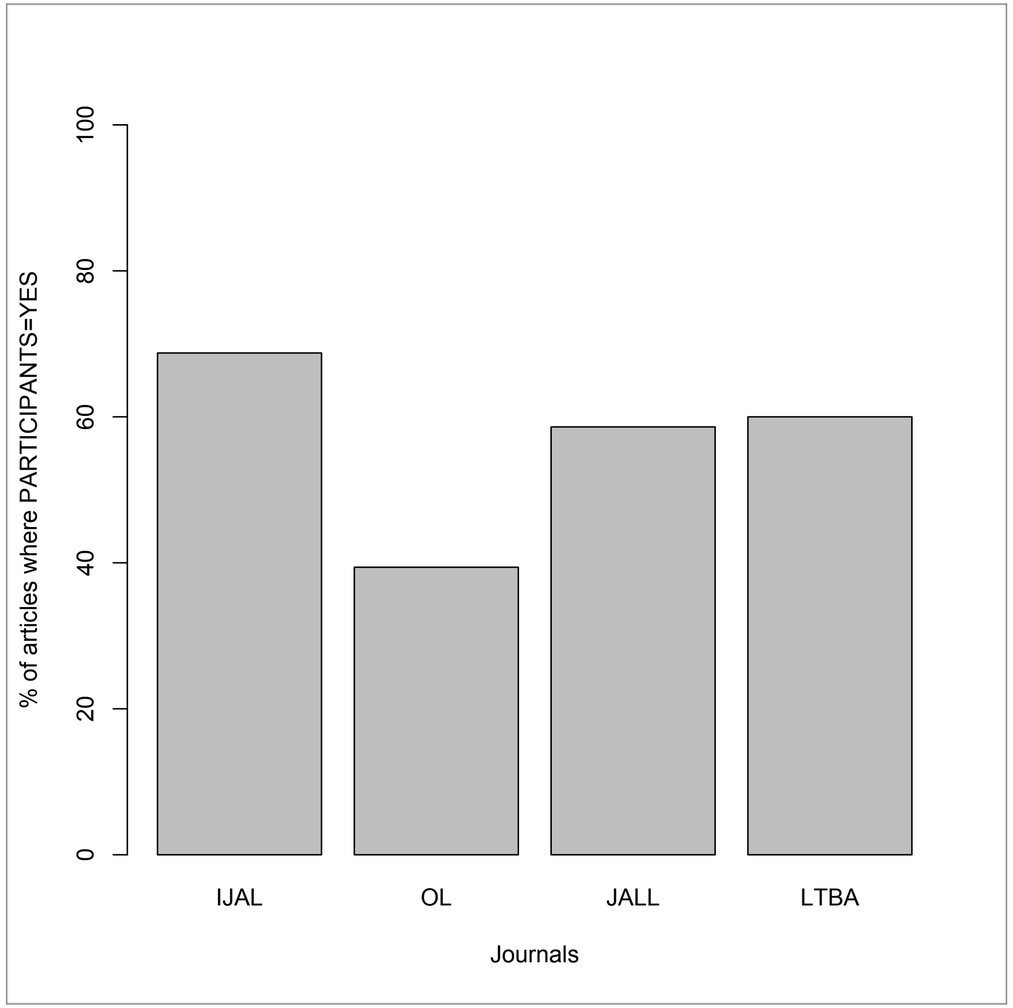
Value: Data should be preserved



- IJAL authors…
  - preserve their records in an archive

…more often than in other areal journals.

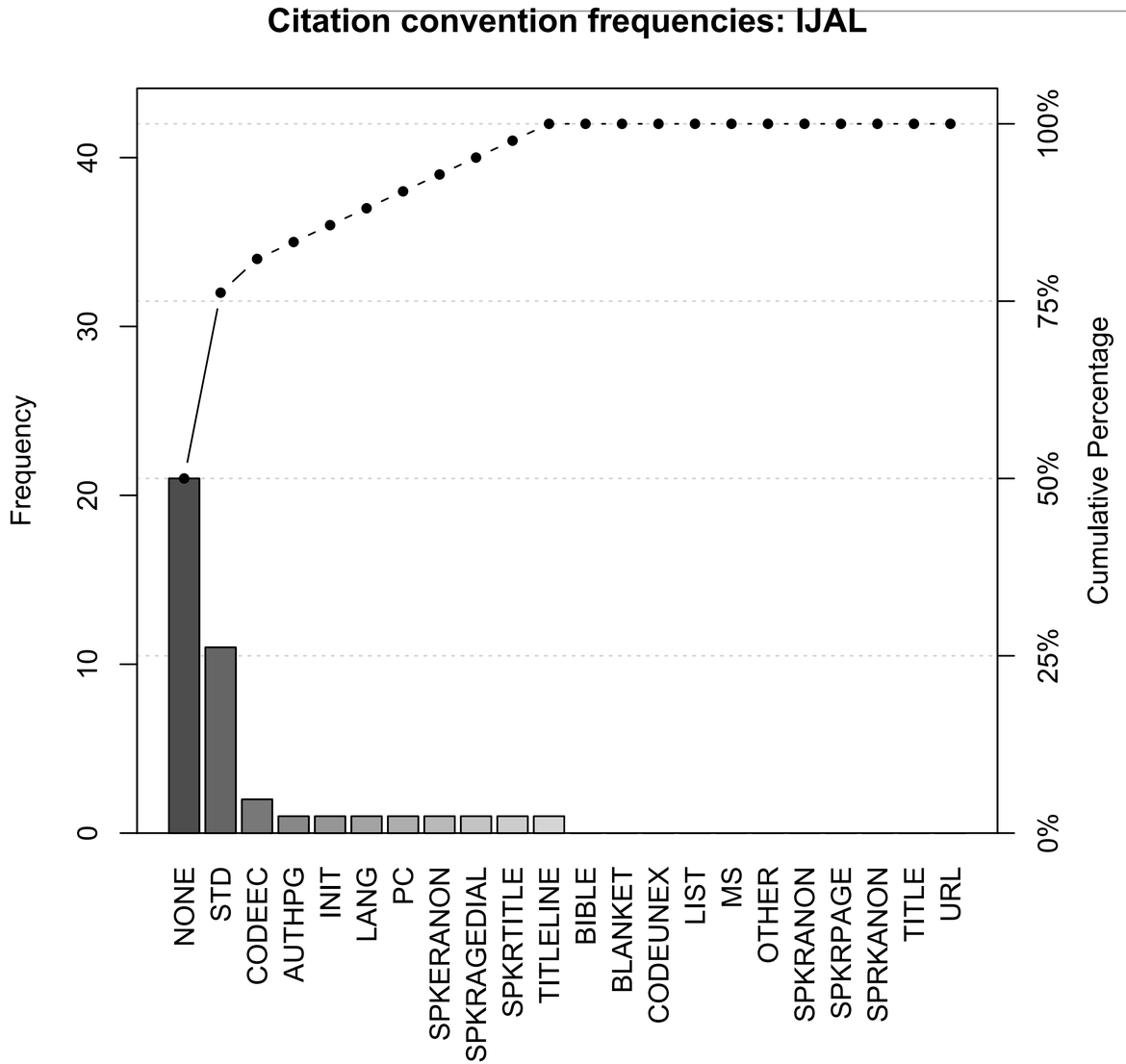
Value: Data should be properly attributed



- IJAL authors…
  - Describe, name, thank and acknowledge the speakers they work with

…more often than in other areal journals.

Value: Data should be cited



- IJAL authors…
  - cite data sources when they have guidelines

…but we need better standards for other kinds of data.

Reproducible Research and the Americanist tradition in linguistic fieldwork

- As linguistics moves toward becoming a reproducible social science
- students of American Indigenous languages will be able to serve as examples for how to do research that is
  - ethical,
  - attributable, and
  - reproducible.

References

- Berez-Kroeker, Andrea L., Lauren Gawne, Barbara F. Kelly & Tyler Heston. Submitted. Transparency of data and methods in linguistics: The state of the discipline.
- Gezelter, Dan. 2009. Being scientific: Falsifiability, verifiablility, empirical tests, and reproducibility. The Open Science Project blog, http://www.openscience.org/blog/?p=312. Retrieved 29 November, 2013.
- Golla, Victor. 1995. The records of American Indian linguistics. In Sydel Silverman & Nancy J. Parezo (eds.), Preserving the anthropological record, 143-157. New York: Wenner-Gren Foundation for Anthropological Research.
- Henke, Ryan & Andrea L. Berez-Kroeker. To appear. A brief history of archiving in language documentation, with an annotated bibliography. Language Documentation & Conservation.
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. Linguistics 36: 161-195.
- Himmelmann, Nikolaus P. 2006. Language documentation: What is it good for? In Jost Gippert, Nikolaus P. Himmelmann, & Ulrike Mosel (eds.). Essentials of language documentation, 1-30. Berlin: Mouton de Gruyter.
- Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data.
- Thomason, Sally. 1994. The editor's department. Language 70: 409-413.
- Woodbury, Anthony C. 2011. Language documentation. In In Peter K. Austin & Julia Sallabank (eds.), Cambridge Handbook of Endangered Languages, 159-186. Cambridge: Cambridge University Press.

Thank you!

bit.ly/LinguisticsDataCitation

Special thanks to **Meagan Dailey, Ryan Henke, Gary Holton, Kavon Hooshiar, Susan Kung, Peter Pulsifer,** and the participants in the Workshops on Data Citation & Attribution in Linguistics. This material is based upon work supported by the National Science Foundation under grant SMA-1447886. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Submit your nominations for the Franz Boas Award! delaman.org