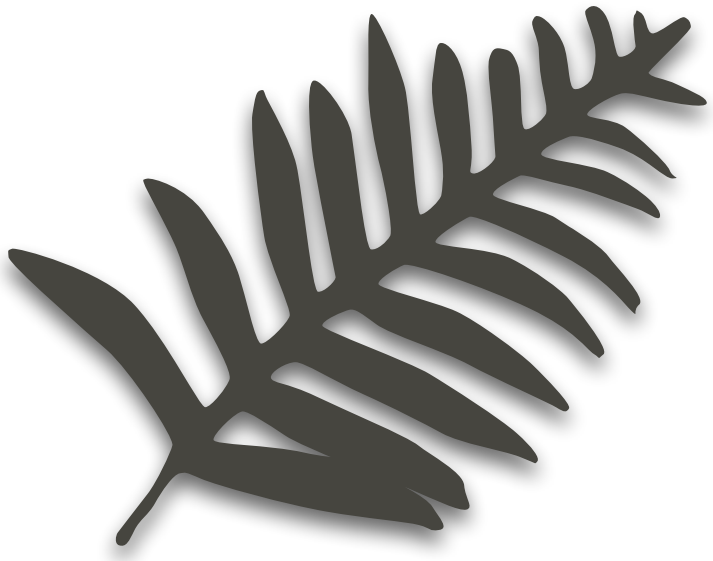
Integrating archiving into the Language Documentation curriculum at the University of Hawai‘i at Mānoa

RRR Conference
Melbourne
2 December 2013

Andrea L Berez
andrea.berez@hawaii.edu



Goal this morning

- To address the topic of training Language Documentation Masters- and Doctoral-level students in making reproducible claims via proper archiving and citation.

On valuing reproducibility

- In science, claims must be *falsifiable*, *verifiable*, and *reproducible*.
- Valued by the *Reproducible Research* movement:
  - The product of academic research is the paper *and* the full data so that claims can be reproduced (e.g., http://reproducibleresearch.net/index.php/Main_Page, http://cran.r-project.org/web/views/ReproducibleResearch.html, http://biostatistics.oxfordjournals.org/content/10/3/405.full, http://www.computer.org/csdl/mags/cs/2009/01/mcs2009010005.pdf)
- Linguistic science values reproducibility too.
- Open Science Project:

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

- Language Documentation can make linguistic claims reproducible:
  - "[Language] documentation […] will ensure that the collection and presentation of primary data receive the theoretical and practical attention they deserve." (Himmelmann 1998:164)
  - "[…] it is our professional responsibility to provide the data on which our claims are based […] It enhances the scientific basis of the linguists' work." (Theiberger 2009: 365-6)
  - "Establishing open archives for primary data is in the interest of making analyses accountable." (Himmelmann 2006:6)

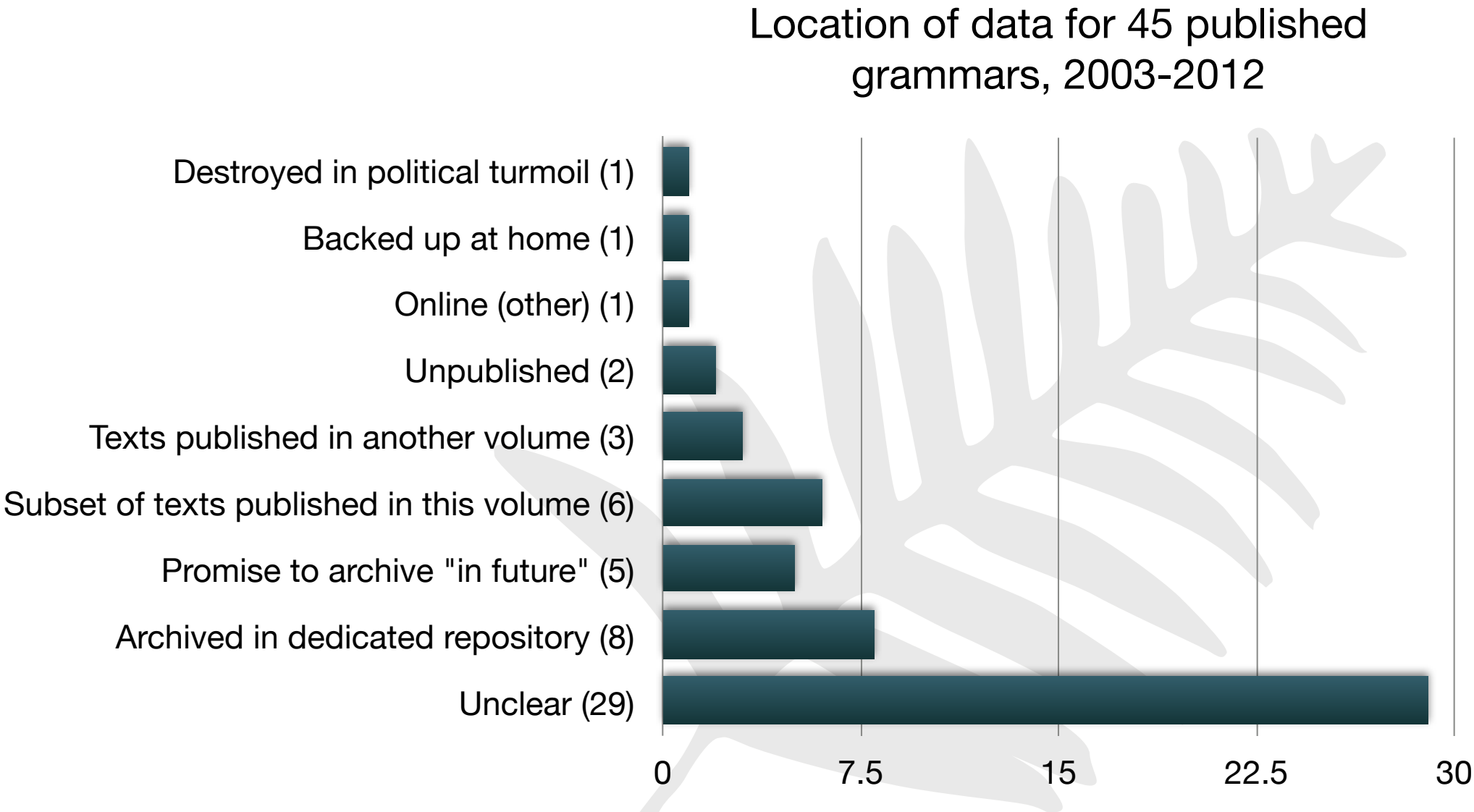
On valuing reproducibility

- Despite this, most descriptive publications make make reproducibility difficult.
- Boasian history (cf. Woodbury 2010):
  - Raw textual data separate module from the descriptive grammar that generalizes over it
  - No tradition of linking generalizing claims to data
  - Old habits are hard to break!

Berez & Heston (in prep)

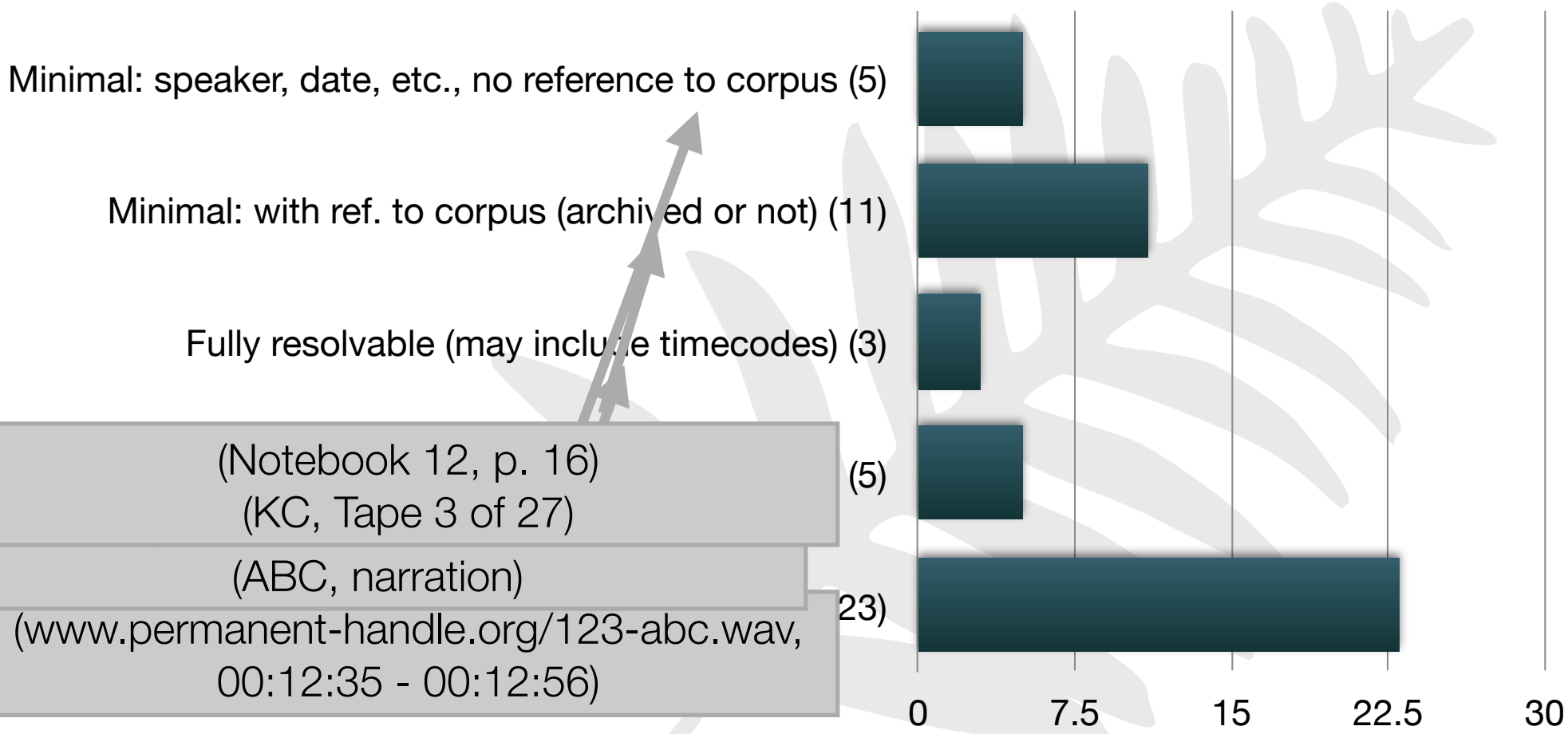
- Surveying descriptions from 2003-2012
  - Grammars, articles, PhD theses
- Is source of data made explicit?
- Is current location of data made explicit?
- Are examples cited back to raw data?
  - If so, can these be resolved?

Berez & Heston (in prep)



Berez & Heston (in prep)

Citation of examples of descriptive claims for 45 published grammars, 2003-2012



(Notebook 12, p. 16)
(KC, Tape 3 of 27)

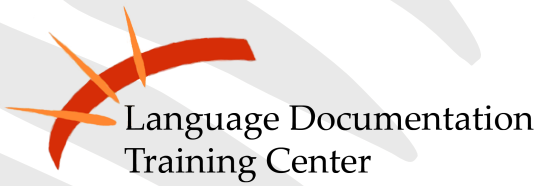
(ABC, narration)
(www.permanent-handle.org/123-abc.wav, 00:12:35 - 00:12:56)

On teaching students to be scientific

- Are we training our students to be scientific enough?
- If archiving primary data and citing back to that data is still not happening in the publishing world, how can we expect students to do it?
- Calls for training in language documentation invariably include data management. (Jukes 2011)
- Archiving often one small component of that.
  - Discussion-based, not hands-on (with exceptions).
  - Result: Students (mostly) not urged do it now. Examples in theses (mostly) not cited to resolvable resources. Claims (mostly) not reproducible.

Postgraduate program at University of Hawai‘i

- Postgrad program in Language Documentation & Conservation
- Activities include:
  - MA & PhD degree programs
  - International Conference on Language Documentation & Conservation series
  - Journal *LD&C*
  - Language Documentation Training Center



Postgraduate program at University of Hawai‘i

- Extensive opportunities for Language Documentation-type classes:
  - LING 680: Intro to Lang Documentation
  - LING 710: Methods in Lang Documentation
  - LING 630: Field Methods (2 semester sequence).
  - LING 617: Language Revitalization & Language Acquisition
  - LING 750: Seminars in (e.g.) Methods of Language Conservation, Archiving, Biocultural Diversity

Postgraduate program at University of Hawai‘i

- Are our students archiving and citing?
    - Only 2 from last 8 years of filed PhD Theses have archived field data and cited examples back to that data.
- Even at UH, archiving and citation as a core value of language documentation has not been communicated sufficiently to our students.
- So what can we do?

Kaipuleohone: UH Digital Linguistic Archive

- OLAC & DELAMAN member archive
- Founded 2008 to digitize and house ethnographic materials associated with UH
- 650+ items
- First 4 years: mostly served emeriti and current faculty; a few students were early-adopters

Kaipuleohone: UH Digital Linguistic Archive

- Collections include:
  - Derek Bickerton's Hawaiian Creole English collection
  - Robert Blust's Austronesian items
  - Byron Bender's Marshallese collection
  - Charlene Sato Center for Pigin & Creole Studies
- Wasn't actively integrated postgraduate student experience

Changes to our PhD Requirements

- As of Fall 2013, 1st of 2 steps: major change to PhD Handbook of Requirements:

- Students whose dissertations are based on data collected during the course of their own fieldwork are required to properly archive their data in an appropriate language archive to ensure longevity of the data. Students will develop an archiving plan early and will include a description of this plan in the Dissertation Proposal. Data can be archived with Kaipuleohone, the University of Hawai'i Digital Ethnographic Archive, or with another accepted language archive (see http://ems03.mpi.nl/delaman/ for a list of accepted archives). For students archiving their data in Kaipuleohone, the archiving plan should be developed in consultation with the current archive director. All students will be required to submit proof of deposit to the committee before the dissertation can be approved.

Changes to our PhD Requirements

- Summary:
  - Students whose theses are based on fieldwork are required to properly archive their data
  - Archiving plans will be part of the Dissertation Proposal.
  - Only accepted DELAMAN archives may be used.
  - Students required to submit proof of deposit to the committee before the thesis can be approved.
- This year 7 PhD students have developed an archiving plan under these regulations.

Changes to our PhD Requirements

- Soon step 2: Descriptive theses must cite (theoretically) resolvable resources.
- For example:

16 MP; Ten k'e ngge' nadzitez'aan.
ice on upland 3S.SUB.animal.runs.ITER.PERF
'It runs back upland on the ice.'
((Markle Pete, oai:paradisec.org.au:ALB01-059, 00:06:51.740-00:07:14.870))

- Two students currently testing, will provide feedback

Potential sticking points

- Q: Student wants to revise transcription, glossing, or analysis. Is the citation now incorrect?
    - A: No. The citation is to the primary data itself, not the transcription or analysis per se.
- Q: Student wants to keep data private/inaccessible. Is this going to put the student's PhD at risk?
    - A: No. Student's restricted data is handled like any other depositor. Still able to balance need for privacy.

Potential sticking points

- Q: Student later archives multiple “versions” of the same primary data. Is the citation now incorrect?
  - A: No. Students, like others, are encouraged to reference the original, unedited version of the primary data. Later versions can be associated to original file in archive metadata.
- Q: Isn’t it overwhelming to archive and cite while also writing a thesis?
  - A: Students are taught to integrate prep for archiving and citation early through coursework. Part of departmental culture.

In summary

Summary

- Departments declare their *values* by requiring milestones
  - We value writing article-length papers: QPs
  - We value being able to talk eloquently about linguistics: Comps exams
  - We value being able to undertake large-scale research projects: PhD Theses
- Why would archiving and reproducibility of claims be any less valued?

Summary

- Students are less likely to adopt practices that are seen as nonessential.
    - Little time to spend on what they don't get credit for.
- By teaching students how to archive, and then expecting them to do it, we show that we value *reproducibility*.

Summary

- Archiving and citation seen as another degree milestone toward training professional linguists, on par with learning how to do fieldwork, write articles, analyze data, etc.
- Students now view the proper care for their data and responsible use of it in their writings to be part of expected professional practice.
- When skeptics can interact with examples embedded in the speech context, claims based on them become reproducible, verifiable, and falsifiable.
- That's something we value.

Postscript: Linguistics setting example for college

- UH on verge of campus-wide data storage and sharing initiative
- Dean of College of Languages, Linguistics & Literature:
  - Using language documentation model of providing citable primary language data
  - Kaipuleohone is becoming the repository for the entire college

References

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.

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.

Woodbury, Anthony C. 2011. Language documentation. In Austin & Sallabank 2011, 159-186.

Thank you.

Special thanks to the UHM University Research Council and the Dean's Office of the College of Languages, Linguistics & Literature.