Data Citation:
The State of the Art in Linguistics

1st Workshop on Data Citation and Attribution in Linguistics
Boulder CO September 18 2015

Lauren Gawne(1), Andrea Berez-Kroeker(2),
Barbara F. Kelly(3) & Tyler Heston(2)

(1) SOAS, University of London, (2) The University of Hawaii, (3) The University of Melbourne
Developing Standards for Data Citation and Attribution

Workshop aim: to develop and promote standards for data citation and attribution for linguistic data sets.

A data-driven linguistic science has the potential to provide substantiation of scientific claims by promoting attention to the care and structuring of language data.
What is the state of the art?

How do researchers link their writing back to the underlying data?

● Where does our data come from?
● What kind of data are we using?
● Where is the data now?
● Are we citing our examples? If so, how?
Our study

We examined:

- 100 books (50 published grammars, 50 dissertations)
  - Variety of publishers, languages, institutions, supervisors
- 271 journal articles from 9 journals
  - Range of areal foci, linguistic subfields, theoretical persuasions
- Published 2003-2012
  - 5 years after Himmelmann 1998:164

“[Language] documentation […] will ensure that the collection and presentation of primary data receive the theoretical and practical attention they deserve.”
<table>
<thead>
<tr>
<th>Journal</th>
<th>No. articles included</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>International Journal of American Linguistics</td>
<td>33</td>
<td>IJAL</td>
</tr>
<tr>
<td>Journal of African Languages and Linguistics</td>
<td>29</td>
<td>JALL</td>
</tr>
<tr>
<td>Journal of Sociolinguistics</td>
<td>33</td>
<td>JS</td>
</tr>
<tr>
<td>Language</td>
<td>33</td>
<td>LANG</td>
</tr>
<tr>
<td>Linguistics of the Tibeto-Burman Area</td>
<td>18</td>
<td>LTBA</td>
</tr>
<tr>
<td>Natural Language and Linguistic Theory</td>
<td>32</td>
<td>NLLT</td>
</tr>
<tr>
<td>Oceanic Linguistics</td>
<td>33</td>
<td>OL</td>
</tr>
<tr>
<td>Studies in Language</td>
<td>30</td>
<td>SL</td>
</tr>
<tr>
<td>Studies in Second Language Acquisition</td>
<td>30</td>
<td>S2LA</td>
</tr>
</tbody>
</table>
Data distribution by year of publication

Red - published grammars
Blue - dissertations

Books

Journals
Data Coding

● Methodological variables:
  ○ participants, data collection equipment, data analysis tools/software, time spent collecting data
    (see bit.ly/GoodMethods)

● Data-related variables:
  ○ 1. Source of data
  ○ 2. Data genre analyzed (linguistic genre)
  ○ 3. Where data is now
  ○ 4. Citation conventions used to reference data, if any
Data Coding

- **Methodological variables:**
  - participants, data collection equipment, data analysis tools/software, time spent collecting data
  (see bit.ly/GoodMethods)

- **Data-related variables:**
  - 1. Source of data
  - 2. Data genre analyzed (linguistic genre)
  - 3. Where data is now
  - 4. Citation conventions used to reference data, if any
Results from Journals
1. Source of data

- 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
- NA: not applicable
1. Source of data: all journals

- Most data come from authors’ own research ~ 50%
- Followed by published data
- Followed by...unstated
1. **Source of data: individual journals**

(see handout)
2. Data genre (linguistic genre)

- CARRIER: data in a carrier sentence
- CONVO: conversational data (natural)
- CONVOTASK: conversational task (e.g. acquisition studies)
- ELICIT: elicitation
- EXPR: experimental
- GJ: grammaticality judgments
- HIST: historical data (e.g. 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 (e.g. acquisition studies)
- TEST: tests in a school environment
- WR: written data (e.g. newspapers)
- OTHER: other
2. Data genre: all journals

- Sentences
- Lexical items/words
- Texts
2. Data genre: individual journals

(see handout)
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: all journals

- Mostly we don’t know!
- “Published” a distant 2nd
3. Where the data is now: indiv js

(see handout)
4. Citation conventions used in examples

AUTHPG: author + page no.

ma han-ac-en   [ah-ic-0 cab-e]   (Coronel:91)
NEG eat-POT-B1  dawn-MCMP-B3  ground-TER

‘I have not eaten since it dawned.’

(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;*  
**Act 5:38**
*exit-PA insideinpeople DEM-C-PROX and let-PL-PL*  
‘Keep away from these men and leave them alone.’

*(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).'</n
(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.’

(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

Northerm 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

\( k^\text{wa} \quad \text{lút} \quad \text{wa}\? \quad \text{s-náw'-lx-s} \)

and \( \text{NEG} \quad \text{SPC} \quad \text{NOM-3SG.run-AUT-3POSS} \)

‘But it didn’t run’ \( \quad \text{(Kinkade, p.c., 2011)} \)

(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

SPKTRANON: 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 nánhái̚r jiào-zhù
3 mōm 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

kʷaʔ ʔíca lut ɬ cəlʔán taʔ-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

STD: citation appears in “standard” format for publications

Wari’, Chapacura-Wanam (Everett and Kern 1967:162)

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.’

(example from SL (Mauri 2008:23)
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.’

(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
4. Citation conventions used in examples: indiv

(see handout)
Results from Books
1. Source of data - Books

Methodologically consistent genre compared to Journals

<table>
<thead>
<tr>
<th>Source of Data</th>
<th>Dissertations</th>
<th>Published Grammars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Own fieldwork</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>Other published sources</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Other unpublished sources</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>
3. Where is the data now - Books

We don’t know where most descriptive data ends up

<table>
<thead>
<tr>
<th></th>
<th>Dissertations</th>
<th>Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown</td>
<td>35</td>
<td>33</td>
</tr>
<tr>
<td>Archived</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>“Will archive”</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>With community</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Online</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Sizable text corpus with grammar</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
4. Data citation - book

Low rates of good practice in referencing back to original data

- Minimal reference e.g. to speaker or story ‘Rajesh’ or ‘Hunting’
- Minimal with reference to corpus: Rajesh ‘JC story’
- Resolvable to underlying corpus, not archived: RL LG1-101027-01
- Fully resolvable to underlying corpus, with time codes & archived: RL LG1-101027-01 01:09

Graph:
- Red - published grammars
- Blue - dissertations

Scores (1-5):
1 = None
2 = Minimal
3 = Partially resolvable
4 = Resolvable
5 = Fully resolvable

Number of publications:

- Minimal reference: e.g. to speaker or story ‘Rajesh’ or ‘Hunting’
- Minimal with reference to corpus: Rajesh ‘JC story’
- Resolvable to underlying corpus, not archived: RL LG1-101027-01
- Fully resolvable to underlying corpus, with time codes & archived: RL LG1-101027-01 01:09

Graph shows low rates of good practice in referencing back to original data.
Citation conventions - books

The reference number (here ref 07032.031) indicates the name of the file the sentence is extracted from (07032) and the line inside the text where the sentence appears (031).

ref 07032.031

Na-tov "Vira rar Myriam.
1SG-call Vira 3PL.DL Myriam
'I called Vira and Myriam.'

(example from Guerin 2008: 8)
Found 1 bundle in this deposit (page 1 of 1)

**07032**

07032.wav  
Access protocol: **URCS**

Download

07032.trs  
Access protocol: **URCS**

Download

ID: 07032  
Title: 07032  
Alternative title: Voices of Aese  
Date created: 13-Jun-07  
Location: Saorol, Mavea Island, Vanuatu  
Devices: Sony HI-MD MZ-RH10 - Minidisks  
Description: SLL narrating a fearful experience she had on Ais Island.

Keywords: English - Mavea - Personal story - Sera Lima Lowet

Gawne, Berez, Kelly, Heston | Delaman | Sept. 18 2015
We have good models for fieldwork


But few of these explicitly discuss management, citation and attribution of linguistic data.

[A notable exception is Bird & Simons 2003]
Attitudes are changing slowly

“...one participant wondered why she felt the need to point out the importance of citing the source of each example when presenting it in an academic publication. This seemed obvious to those present...”

Ruth Singer, Melbourne LIP
“Grammar writing: where are we now?”

http://www.paradisec.org.au/blog/2015/02/grammar-writing-where-are-we-now/
References

Berez, Andrea. 2014. “Reproducible research in descriptive linguistics: Integrating archiving and citation into the postgraduate curriculum at the University of Hawai‘i at Mānoa.” In Amanda Harris, Nick Thieberger & Linda Barwick (eds.), Research, records, and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures. Sydney: University of Sydney Press.

Berez, Andrea, Lauren Gawne, Barbara Kelly and Tyler Heston. In prep. Citation and transparency in descriptive linguistics.


Works from study included here


Thank you.

These slides are available at bit.ly/DataCitationSOTA

lg21@soas.ac.uk
andrea.berez@hawaii.edu

Special thanks to the National Science Foundation, The University of Melbourne library staff, The University of Melbourne and NTU, Singapore where Lauren worked on earlier stages of this project.