Reproducible Research in linguistics: The past and future of “data work” in our field

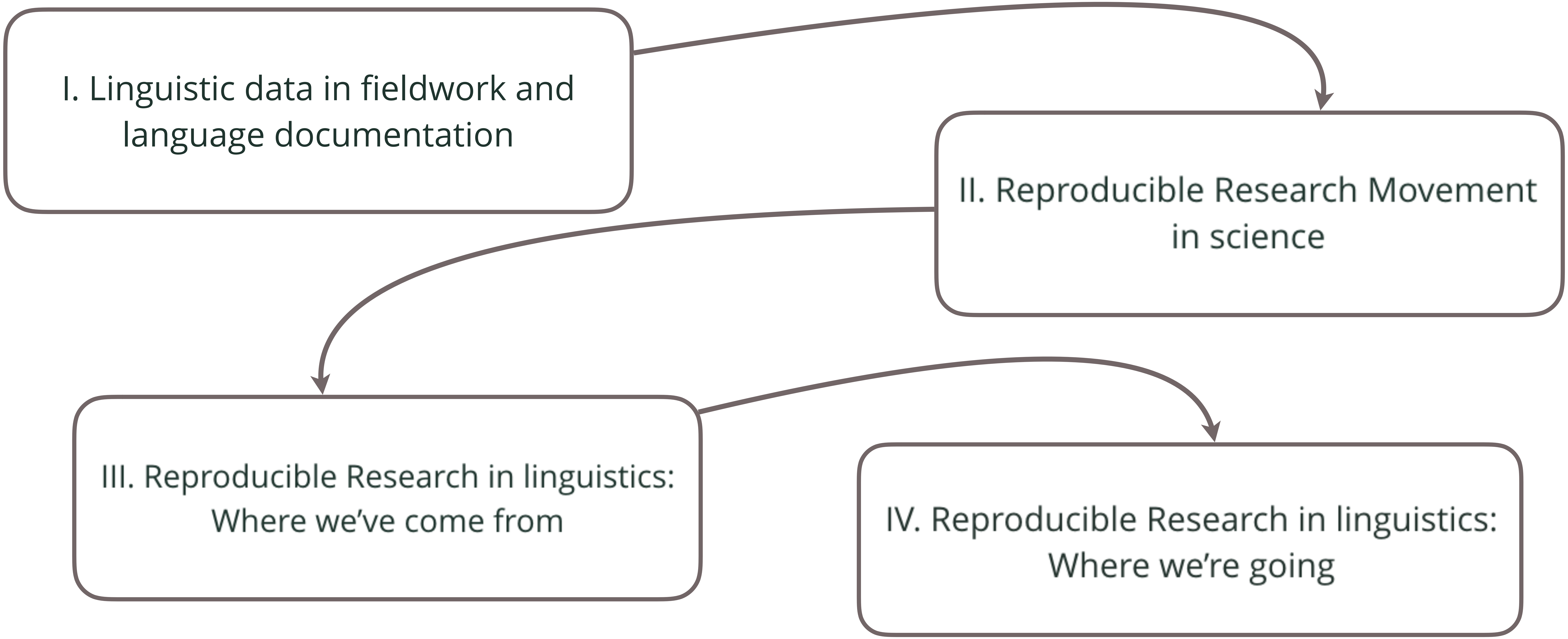
Andrea L. Berez-Kroeker
University of Hawai‘i at Mānoa
andrea.berez@hawaii.edu

University of Oregon
Department of Linguistics
November 10, 2017

Quick poll. By show of hands:

- Do you collect your own data? (alone or w/ collaborators)
  - Primarily in the field?
  - Primarily in the lab?
- Have you ever reused someone else's data?
- Do you spend at least 30% of your research time doing "data work?"
  - Collecting, managing, processing, storing, curating, sharing…

Today's presentation



I. Linguistic data in fieldwork and language documentation

II. Reproducible Research Movement in science

III. Reproducible Research in linguistics: Where we've come from

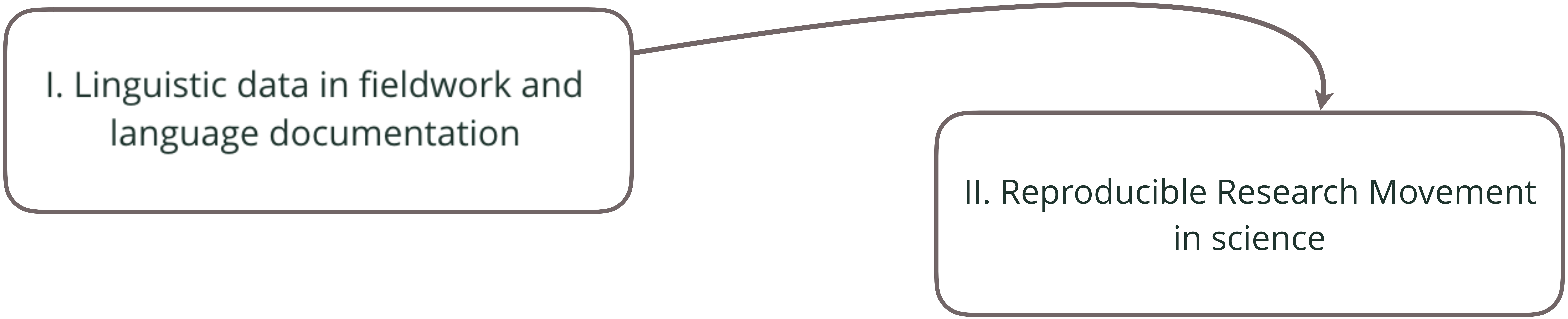
IV. Reproducible Research in linguistics: Where we're going

What I was going to say was...

- For last 2 decades, language documentation has been one of the major forces in the preservation of data in linguistics
- LangDoc has deep roots in records-focussed traditions of linguistic fieldwork (cf Woodbury 2010, Rosenblum & Berez 2010)
  - E. g., historical linguistics in Oceania, descriptive linguistics in Americas and Australia
- Endangered language crisis brought awareness of endangered data crisis
  - Linguistic data are at risk from degradation of many kinds

- Spawned rise of initiatives to preserve documentary data
  - Funding: ELDP, NSF DEL, DoBeS, ELF
  - Tech/Archiving: E-MELD, DELAMAN, OLAC
  - Education: CoLang, Workshops, degree programs
- 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

Today's presentation



I. Linguistic data in fieldwork and language documentation

II. Reproducible Research Movement in science

Reproducible Research Movement in science

Good scientific research is *replicable*

Replicate a controlled study >

New data >

[Dis]confirm previous results

Reproducible Research Movement in science

- Some studies can’t be truly replicated
  - E.g. 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

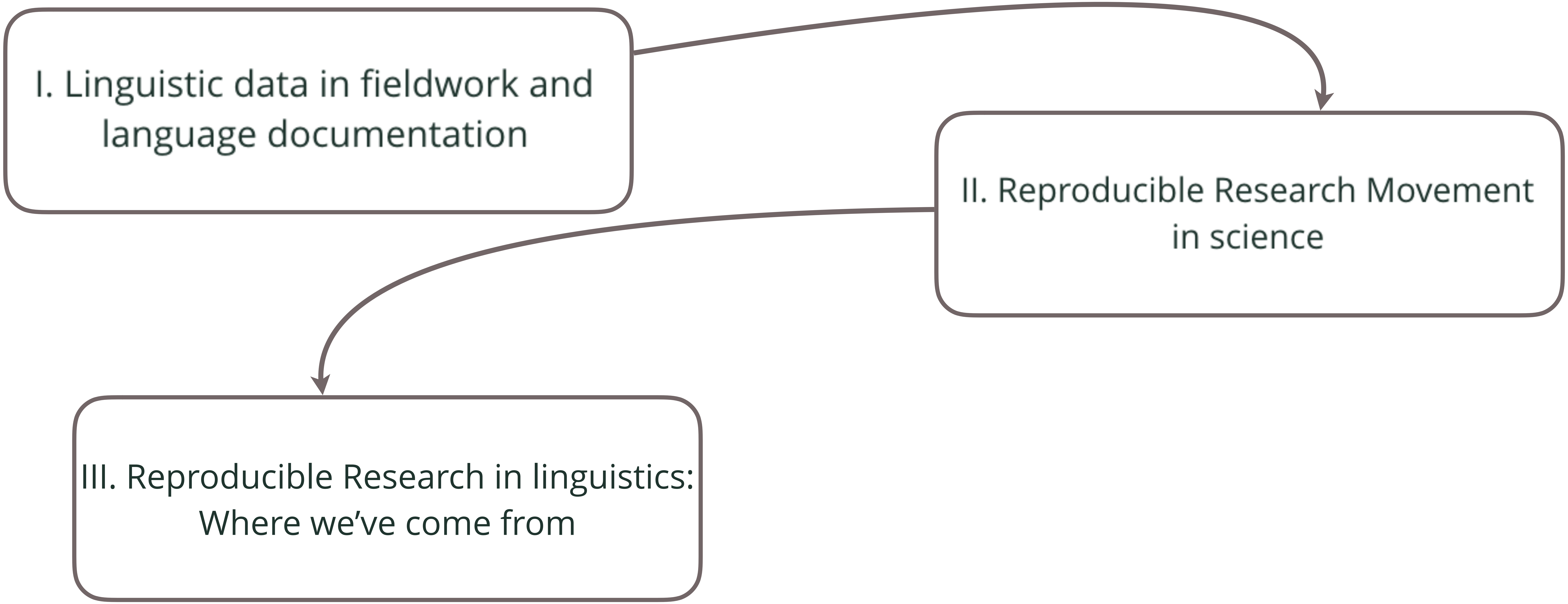
"[…] so frequently, in fact, that the assumption that the data in accepted papers is reliable began to look questionable […]" (Thomason 1994: 409)

"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)

On valuing reproducibility

- That was 1994.
- Called for
  - better description of research methods,
  - better use of data, and
  - better description of data sources.
- How have we been doing since then?

Today's presentation



I. Linguistic data in fieldwork and language documentation

II. Reproducible Research Movement in science

III. Reproducible Research in linguistics: Where we've come from

Reproducible Research in linguistics: Where we’ve come from

Reproducible Research in linguistics: Where we've come from

- Two studies surveying linguistics publications for
  - Transparency of research methods
  - Transparency of data sources
- Over same ten-year period 2003-2012

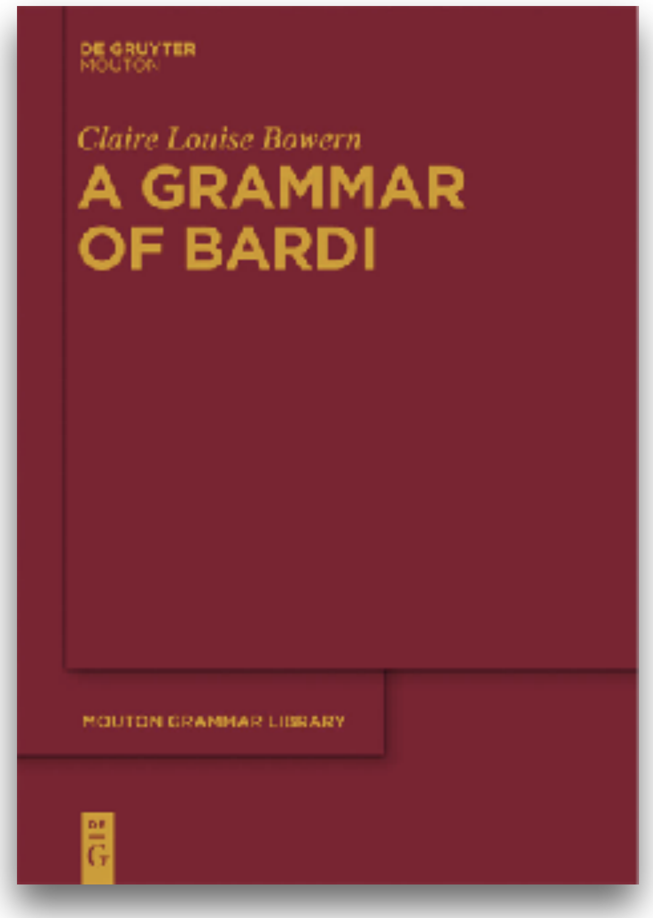
Reproducible Research in linguistics: Where we've come from

LANGUAGE DOCUMENTATION & CONSERVATION

- 100 descriptive grammars

  (Gawne, Kelly, Berez-Kroeker & Heston 2017)

  - 50 published grammars, 50 dissertations
  - Worldwide spread of languages described
  - Dissertations mainly from USA, Australia, Europe

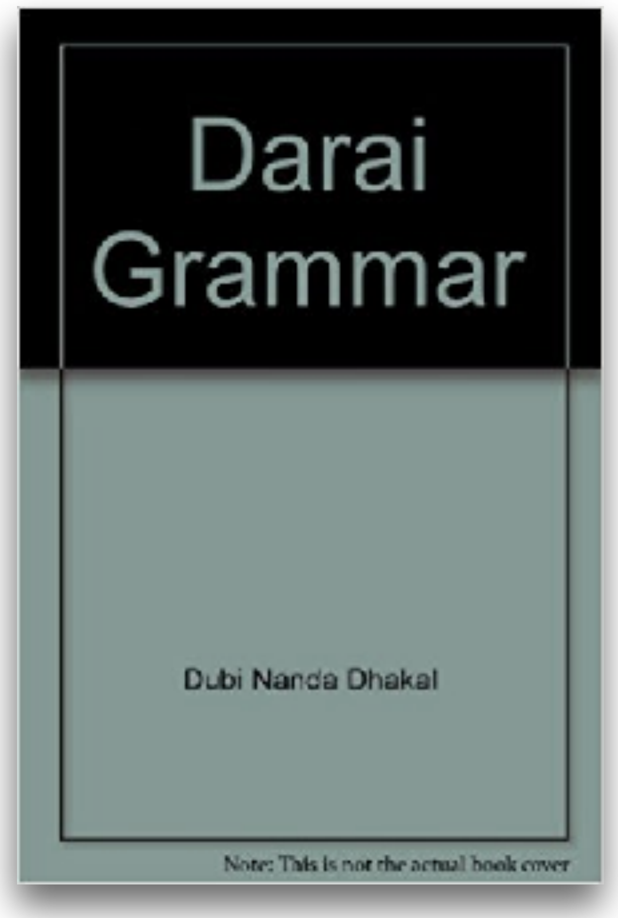


DE GRUYTER MOUTON

Claire Louise Bowern

A GRAMMAR OF BARDI

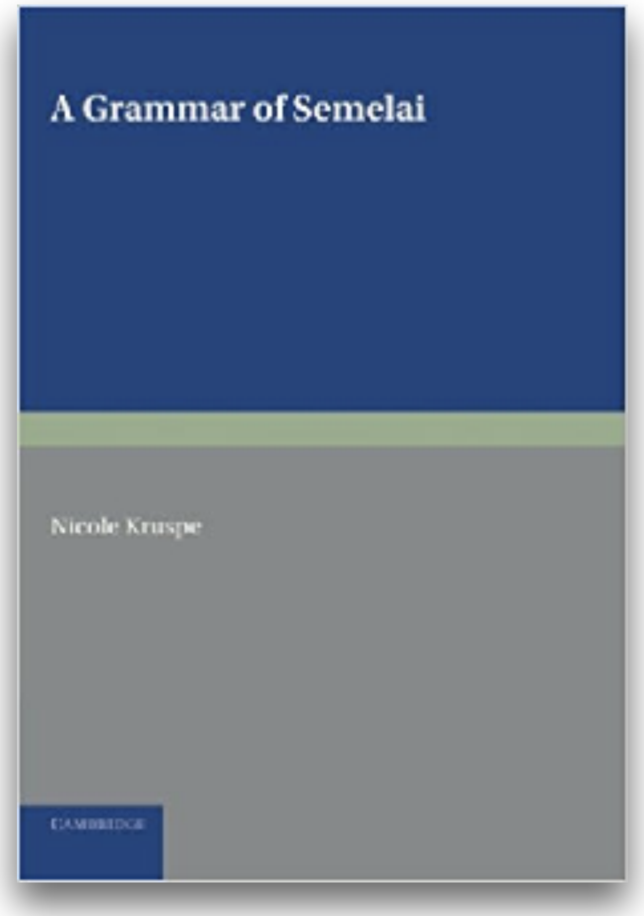
MOUTON GRAMMAR LIBRARY



Darai Grammar

Dubi Nanda Dhakal

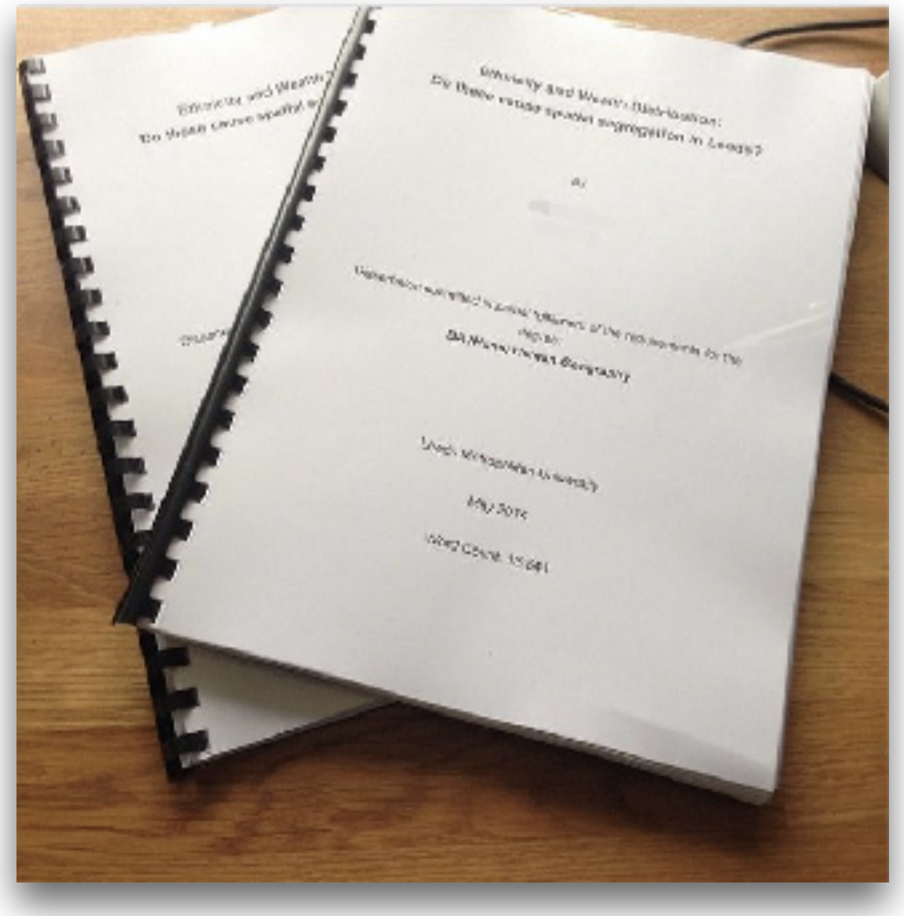
Note: This is not the actual book cover



A Grammar of Semelai

Nicole Kruspe

CAMBRIDGE



Reproducible Research in linguistics: Where we've come from

- 270+ journal articles across 9 journals
  (Berez-Kroeker, Gawne, Kelly & Heston 2017)
  - *International J. of American Linguistics*
    Oceanic Linguistics
    Linguistics of the Tibeto-Burman Area
    J. of African Languages & Linguistics
    J. Second Language Acquisition
    J. Sociolinguistics
    Natural Language and Linguistic Theory
    Studies in Language
    Language

INTERNATIONAL JOURNAL OF AMERICAN LINGUISTICS
Oceanic Linguistics
DE GRUYTER MOUTON
JOURNAL OF AFRICAN LANGUAGES AND LINGUISTICS
Journal of Sociolinguistics
STUDIES IN LANGUAGE
NUMBER 3
LANGUAGE
JOURNAL OF THE LINGUISTIC SOCIETY OF AMERICA

Data coding

- 1. Transparency of *research methods*
  - How well did authors describe their:
    - Research participants?
    - Data collection tools and equipment?
    - Data analysis software?
    - Time collecting data?
    - Speech genres collected?
    - Archiving practices?

Data coding

- 2. Transparency of *data sources*
  - How well did authors:
    - a. Describe the source of the data?
    - b. Describe where their data can be found now?
    - c. Cite numbered examples back to their sources?

8

gi:ligiri bádu [illegible]

mayaŋguldi - greeting
my [illegible] person

burúr' - dwarf wattle
gúrag winbon burúr'
wattle c̄ flowers

daramuŋa - ~~wattle~~ broom-bush
(lilac flowers)

gi:ga - hop-bush
('going away')

ŋi: - back side, [illegible]
murnda - front " "

ŋi:gi dúrŋi múgaru,
[illegible]
its real sore

p8. b5

am
9

ŋaday {dugay
bangan
very sweet (dull to eat)

[illegible] - greeting

wiriwal - heavy
mangkan very

[illegible] diramaŋa
I can't lift it

[illegible] waygindu
[illegible]

[illegible]?

[illegible] gilaywa
time to you left (imp.)
pindal [illegible] gilayway
right

p9. b5

(1)

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

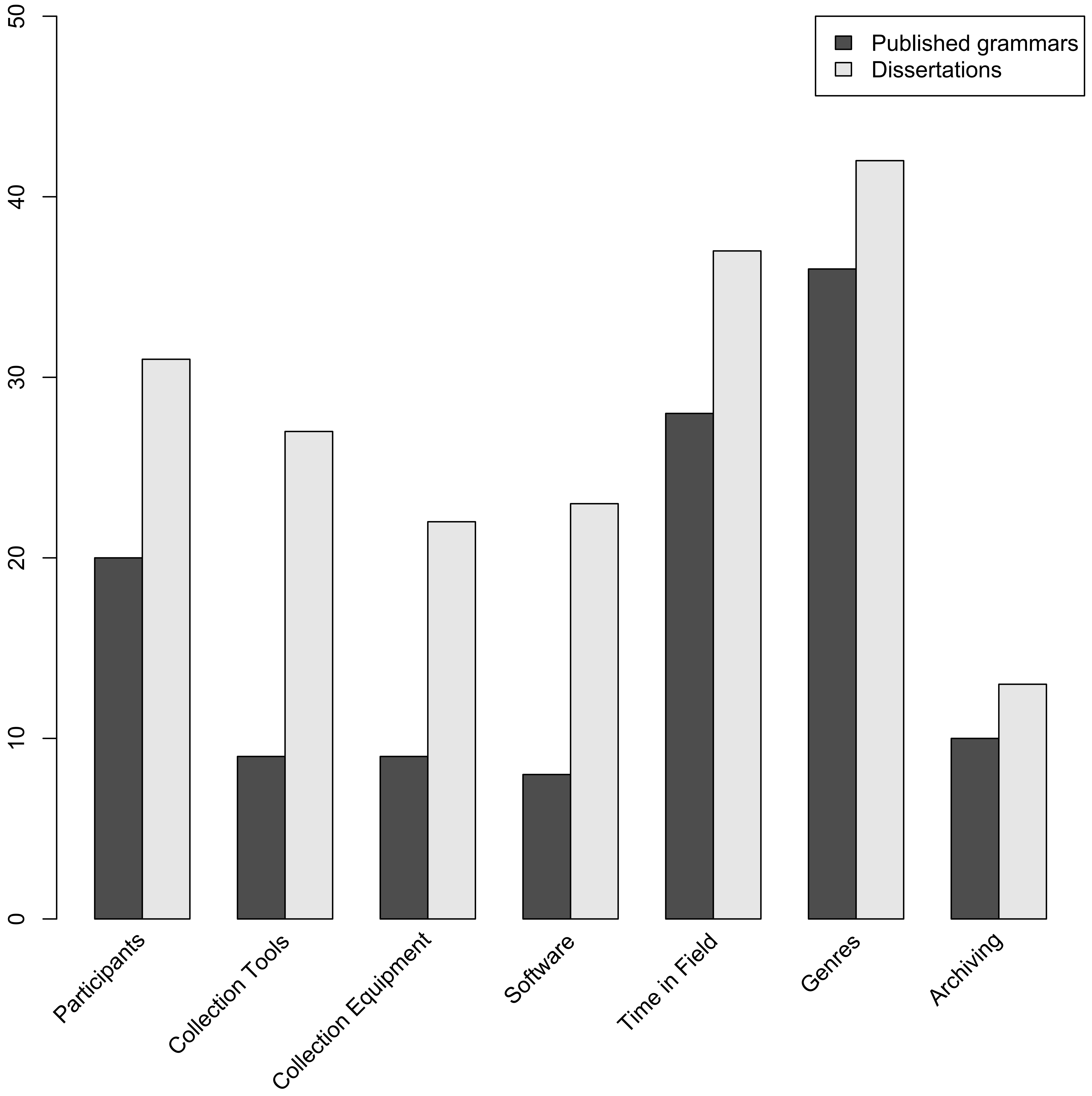
‘Either he will kill fish or he will hunt.’

(Everett and Kern 1967:162)

(example from SL (Mauri 2008:23)

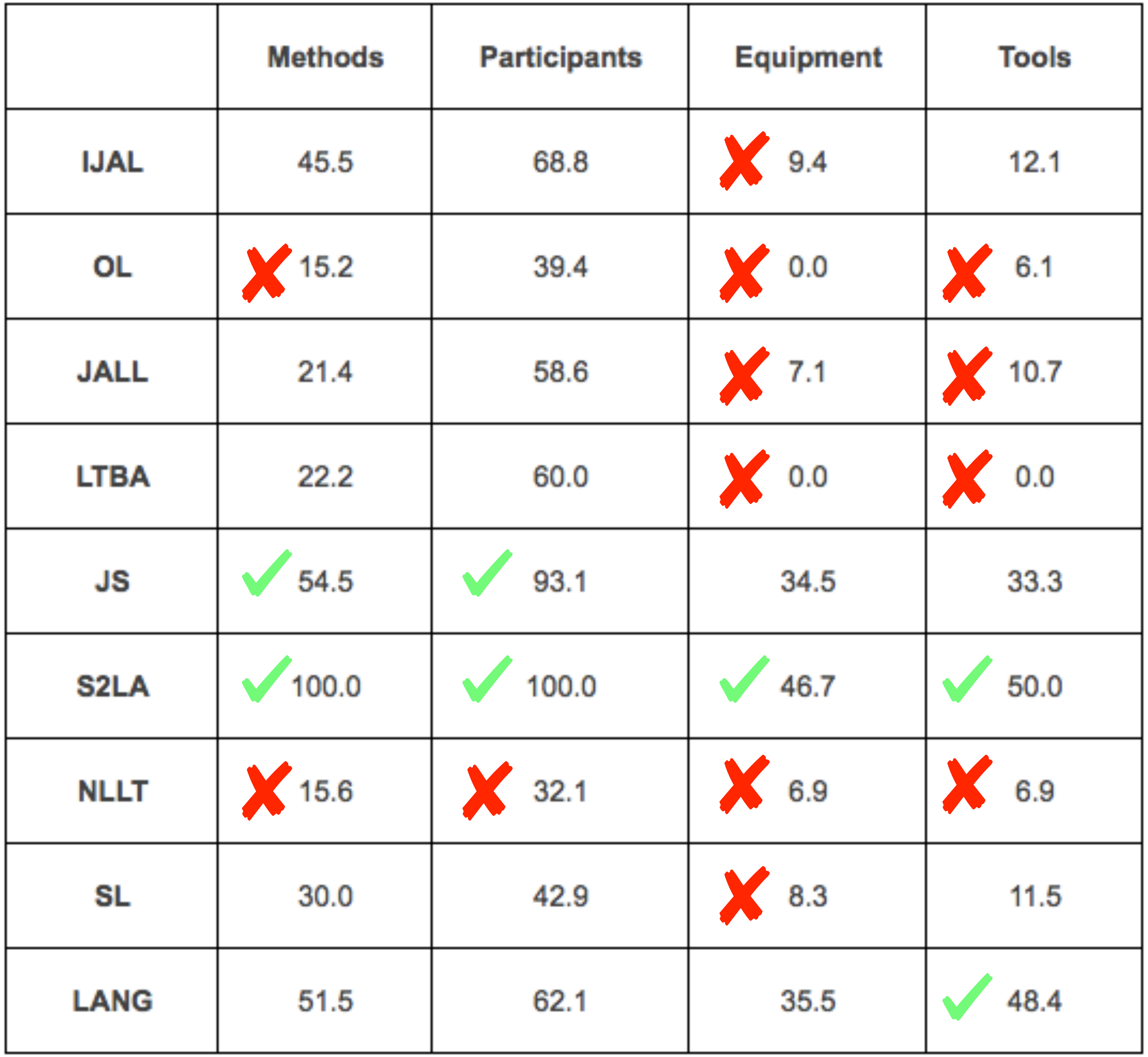
1. Transparency of research methods - Grammars

- Best at describing genres collected, time in field, participants
- Poor description of tools, equipment and software, especially for published grammars
- Dissertation authors outperformed published authors
  - Students need to show mastery of methods



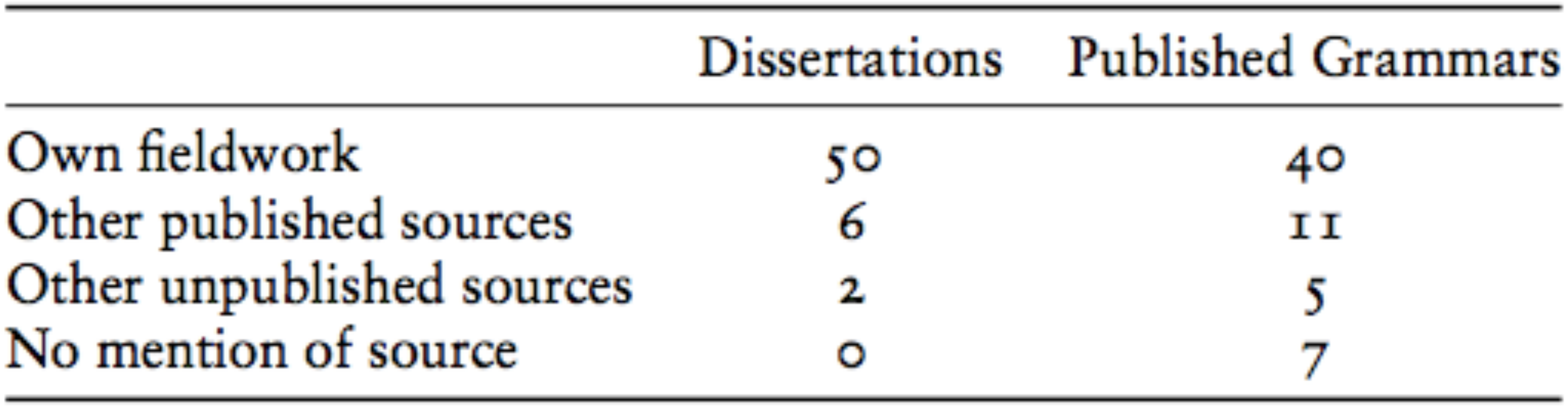
1. Transparency of research methods - Journals

- Space concerns in journals, so even a brief mention counts
- % mention, of non-NA articles



	Methods	Participants	Equipment	Tools
IJAL	45.5	68.8	✗ 9.4	12.1
OL	✗ 15.2	39.4	✗ 0.0	✗ 6.1
JALL	21.4	58.6	✗ 7.1	✗ 10.7
LTBA	22.2	60.0	✗ 0.0	✗ 0.0
JS	✓ 54.5	✓ 93.1	34.5	33.3
S2LA	✓ 100.0	✓ 100.0	✓ 46.7	✓ 50.0
NLLT	✗ 15.6	✗ 32.1	✗ 6.9	✗ 6.9
SL	30.0	42.9	✗ 8.3	11.5
LANG	51.5	62.1	35.5	✓ 48.4

2. Transparency of data sources:
a. What is the source of the data? - Grammars

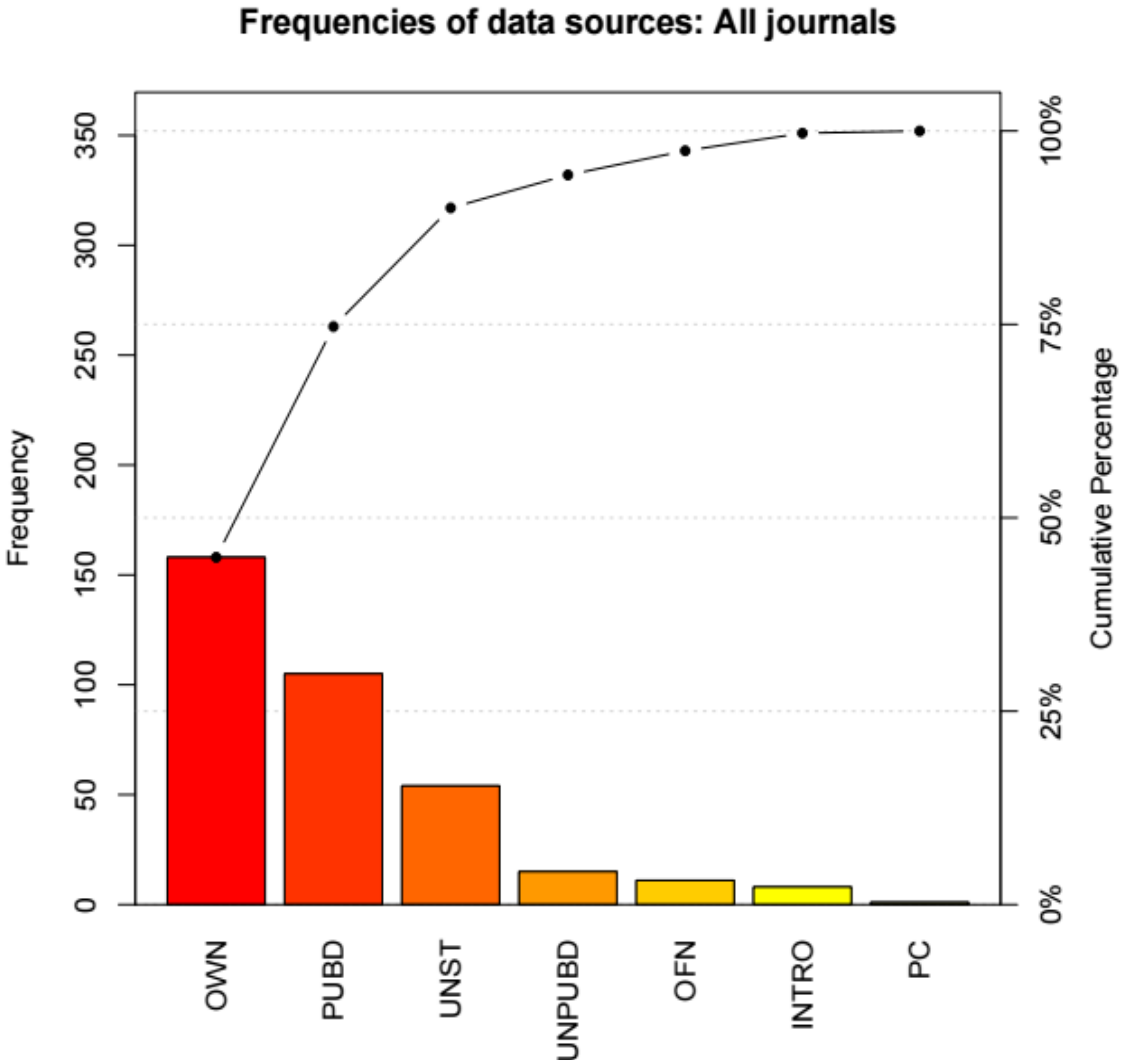


	Dissertations	Published Grammars
Own fieldwork	50	40
Other published sources	6	11
Other unpublished sources	2	5
No mention of source	0	7

- Grammars are overwhelmingly based on author’s fieldwork
- 7 authors did not say anything about source of data

2. Transparency of data sources:
a. What is the source of the data? - Journals

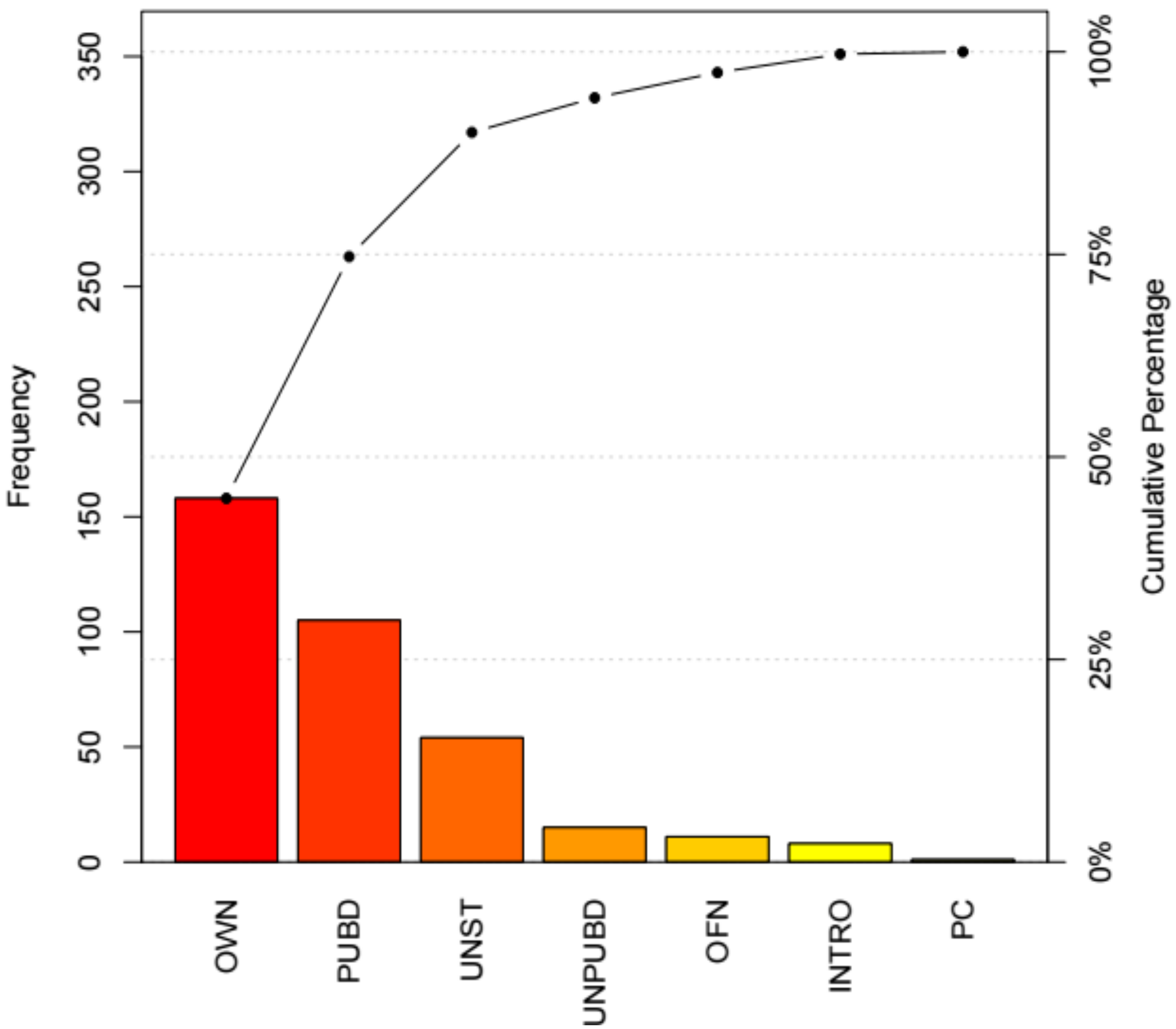
- OWN: data collected by author
- PUBD: published data
- UNPUBD: unpublished data collected by someone other than the author (excluding fieldnotes)
- INTRO: introspection
- OFN: other person’s fieldnotes
- UNST: source of data unstated



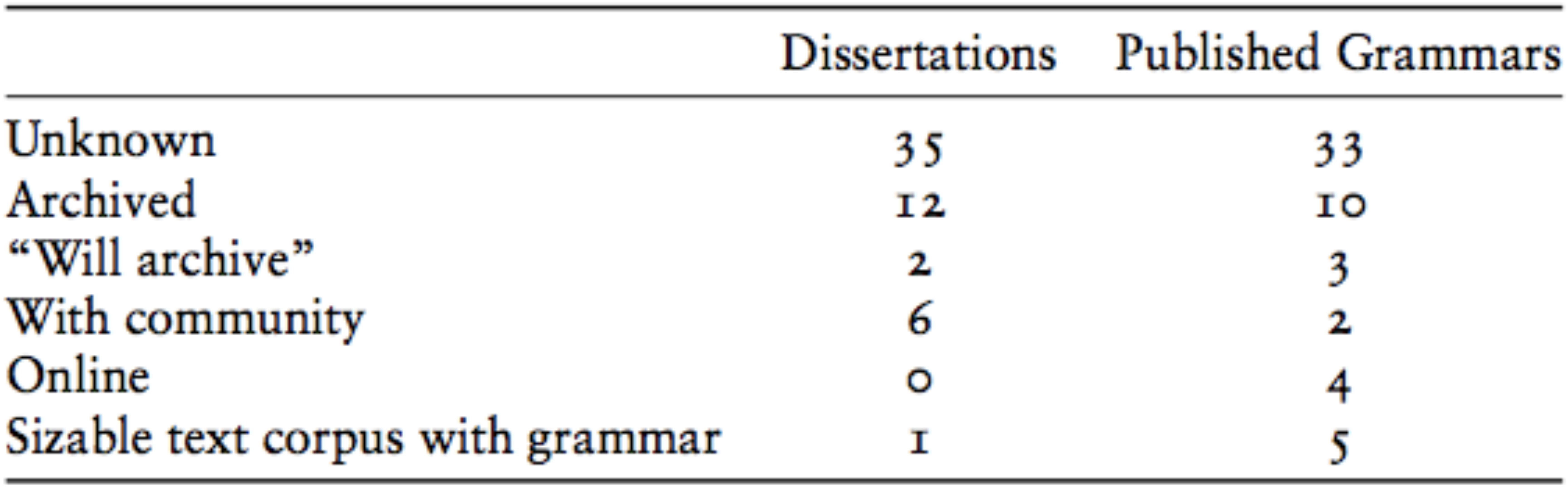
2. Transparency of data sources:
a. What is the source of the data? - Journals

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

Frequencies of data sources: All journals



2. Transparency of data sources:
b. Where is the data now? - Grammars

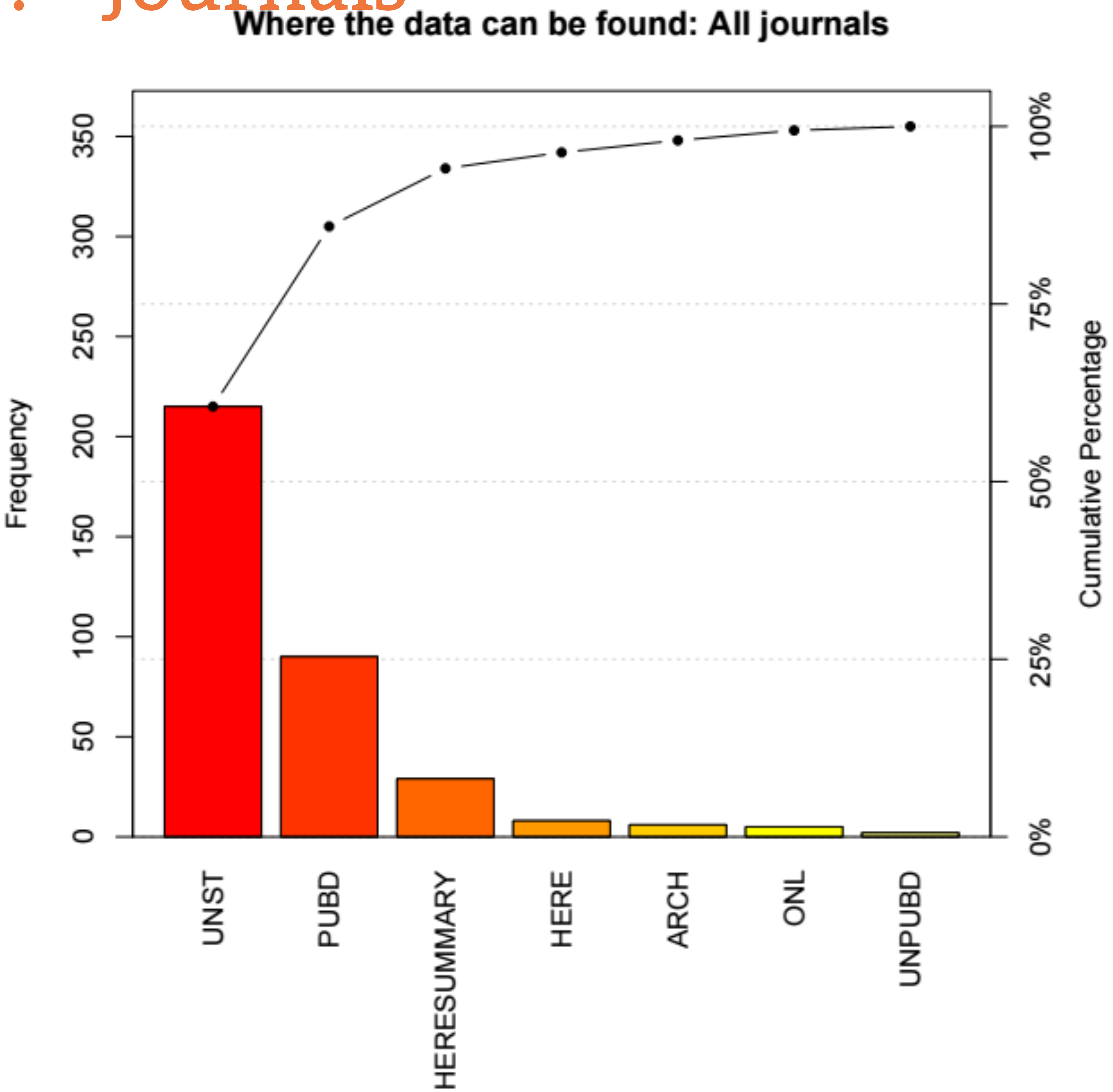


	Dissertations	Published Grammars
Unknown	35	33
Archived	12	10
“Will archive”	2	3
With community	6	2
Online	0	4
Sizable text corpus with grammar	1	5

- Most grammar authors make no mention of where data is

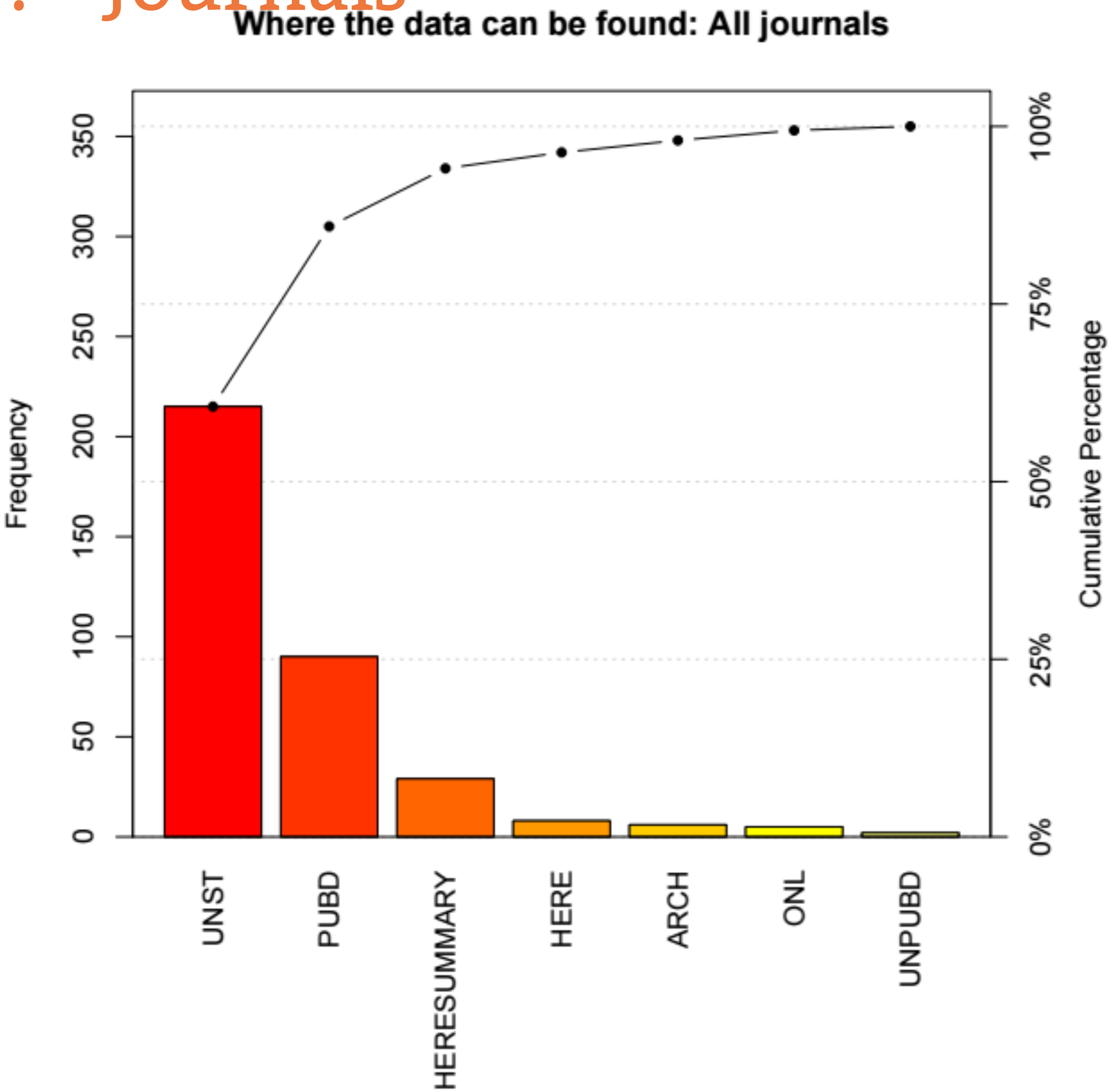
2. Transparency of data sources:
b. Where is the data now? - Journals

- ARCH: archived in proper 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



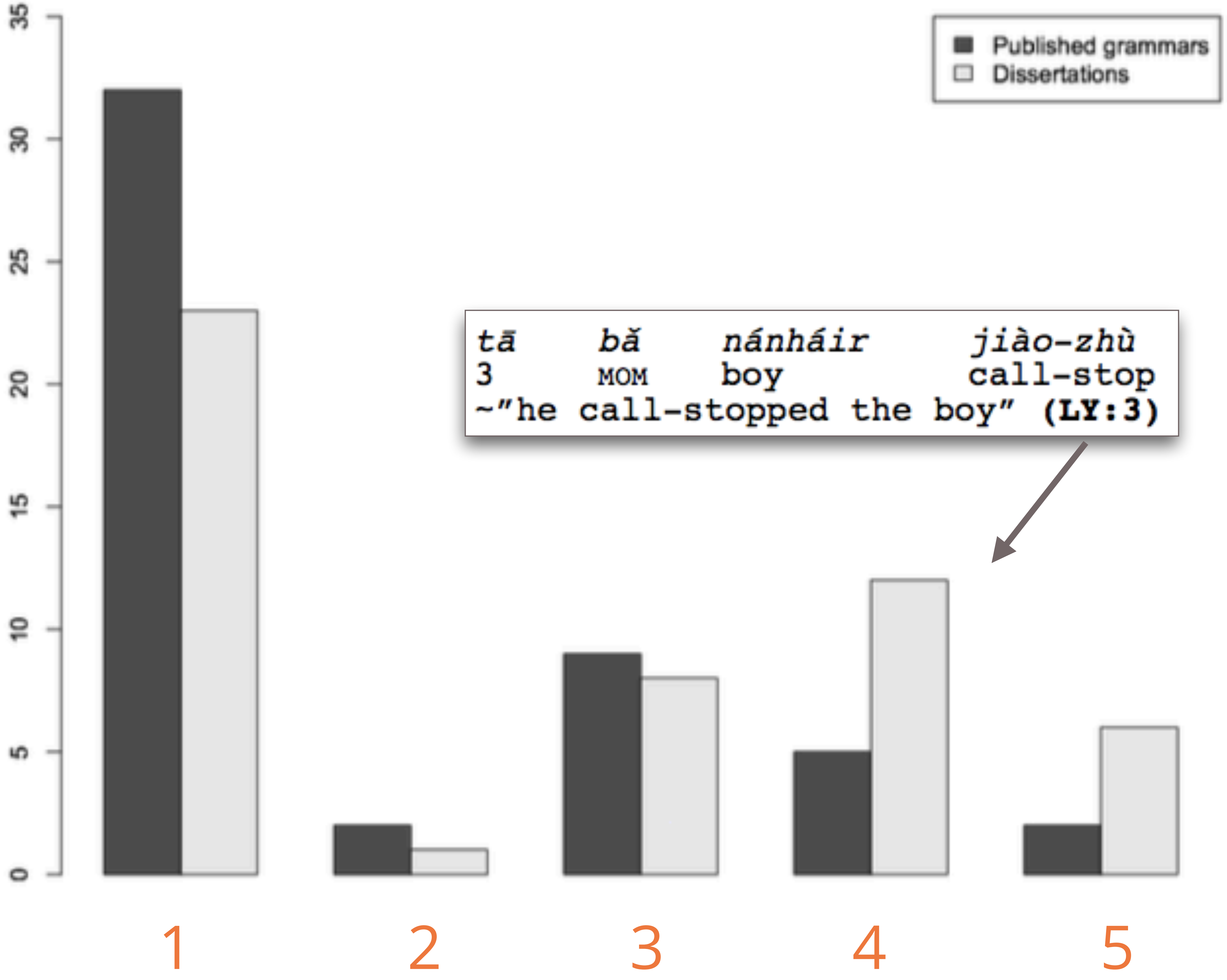
2. Transparency of data sources:
b. Where is the data now? - Journals

- Mostly we don't know!
- "Published" a distant 2nd



2. Transparency of data sources:
c. How are examples cited back to their source data? - Grammars

- 5 point Likert:
  - 1: No citation
  - 2: Minimal reference to speaker *or* title of text
  - 3: Minimal reference to speaker *and* title of text
  - 4: “Resolvable” to section of corpus, but no corpus location
  - 5: Fully resolvable in locatable corpus

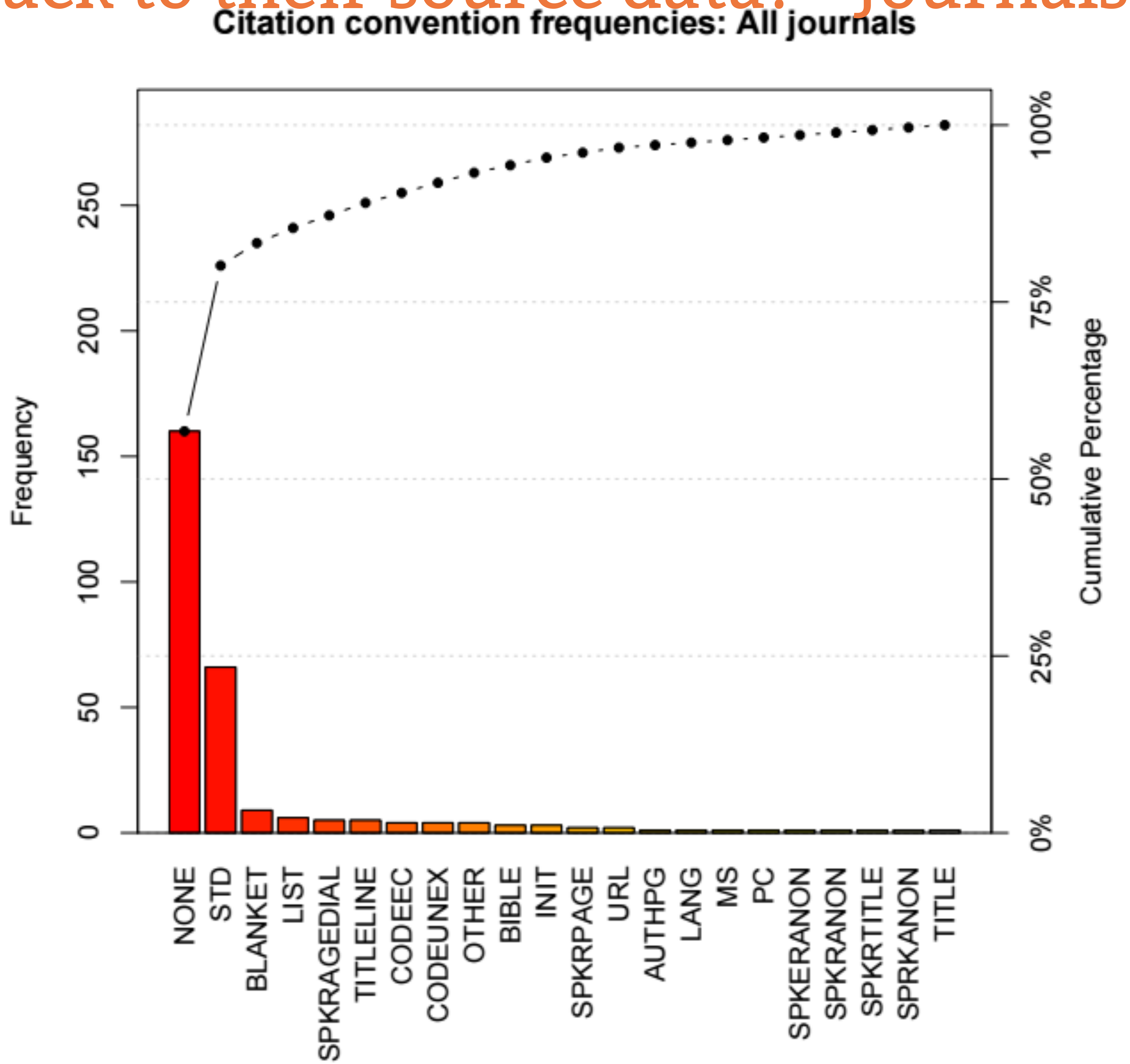


tā *bǎ* *nánháir* *jiào-zhù*
3 MOM boy call-stop
~”he call-stopped the boy” **(LY:3)**

2. Transparency of data sources:
c. How are examples cited back to their source data? - Journals

- Very wide range of possible citation formats (see handout)

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



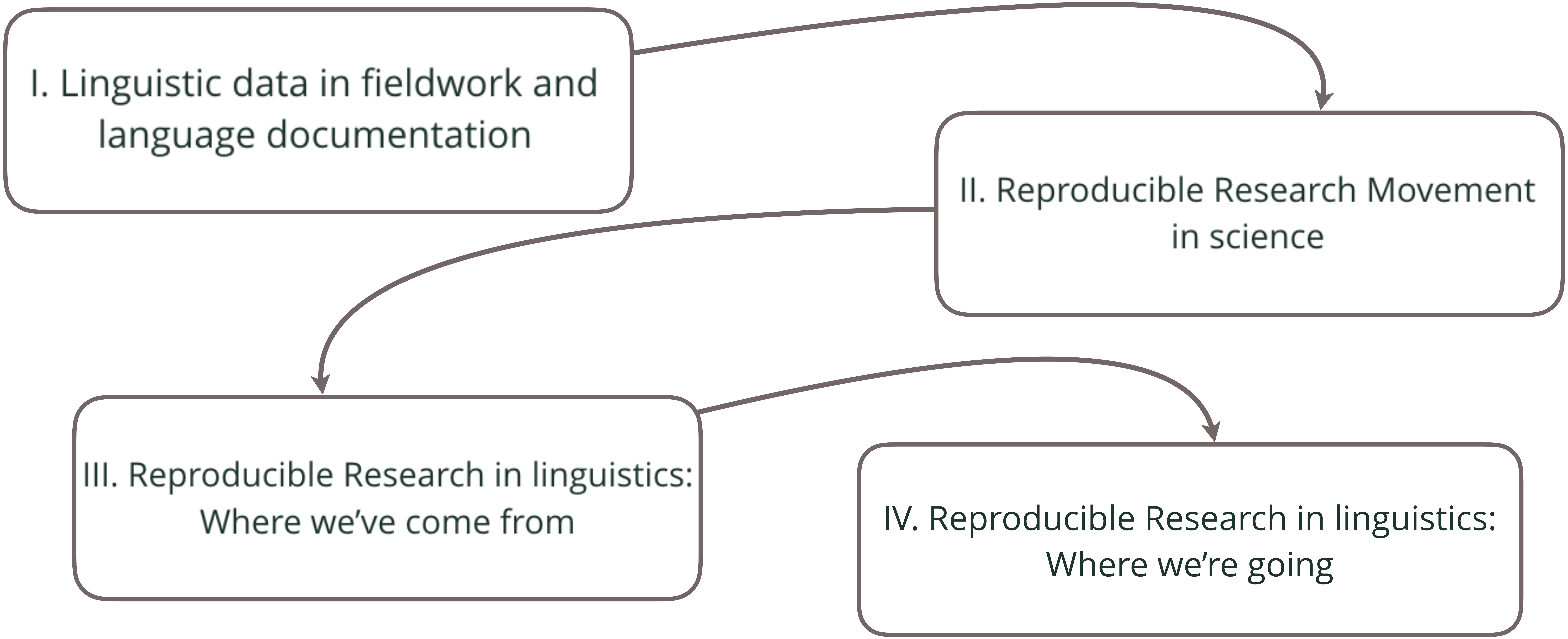
Overall results

- Inadequate description of research methods in publications
- Inadequate description and citation of data sources in publications
- Most authors do not cite data sources
- Except from published paper sources

Overall results

- We have a disciplinary expectation to cite published sources
- And an accepted format for doing so (Everett and Kern 1967:162)
- Data we authors create ourselves is the most common source
    - Not archived!
    - Not cited!
    - Maybe…not valued?
    - Maybe…we don't know how?

Today’s presentation



I. Linguistic data in fieldwork and language documentation

II. Reproducible Research Movement in science

III. Reproducible Research in linguistics: Where we’ve come from

IV. Reproducible Research in linguistics: Where we’re going

Reproducible Research in linguistics: Where we’re going

Reproducible Research in linguistics: Where we're going

- A few recent and ongoing projects to address needs in ling:
  - NSF-funded *Data Citation* project
  - Research Data Alliance group
  - Austin Principles of Data Citation in Linguistics

Data Citation project 2014-2017

NSF *Data Citation* grant 2014-2017

- "Developing Standards for Data Citation and Attribution for Reproducible Research in Linguistics" (SMA-1447886)
- Three workshops - 40+ participants
- Deliverables
  - LSA Panel/Poster, 2017
  - Position paper, to appear in *Linguistics* in 2018
  - Identification of 3 key needs

Need 1: Data citation formats

- Editors need to agree upon formats for citing data
  - Add to the Unified Stylesheet
- Journals need data policies

- Crucially: Persistent identifiers (PID), Granularity, Contributor roles
- For example:

Need 1: Data citation formats

People (roles). Date. *Title*. Repository, granularity. PID.

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

For reference section

People (roles). PID, granularity.

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

For in-text citations

Need 2: Standards for evaluating data work

- How can we incorporate into hires and promotions?
- Metrics for assessing value of data work
- Guidelines to empower applicants, T&P committees

LSA 2018, Friday morning:
Open meeting on evaluating "non-traditional scholarship"

Need 3: Education and Outreach

- Spread the word about Reproducible Research
- Educate ourselves about data management
- Encourage a culture of responsible data sharing

- Cultural change needs international buy-in, not just USA

Research Data Alliance
Linguistics Data Interest Group

RDA

RESEARCH DATA ALLIANCE

RDA Linguistics Data Interest Group

- “The RDA builds the social and technical bridges that enable open sharing of data.”
- You can join the RDA for free!
- Linguistics Data Interest Group - 2017
- First product: “The Austin Principles of Data Citation in Linguistics”

Austin Principles of Data Citation in Linguistics

Austin Principles of Data Citation in Linguistics

- Aims to help linguists understand why and how to cite data
- Annotates the *FORCE11 Joint Declaration of Data Citation Principles* (https://www.force11.org/group/joint-declaration-data-citation-principles-final)
- Iterative input from the linguistics community
- (Very!) beta website: linguisticsdatacitation.org
- Aiming for endorsement by the RDA, linguistic societies
- See handout - I'll give TL;DR

FORCE11
The Future of Research Communications and e-Scholarship

1. Importance

Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

Linguistic data form not only a record of scholarship, but also of cultural heritage, societal evolution, and human potential. Because of this, the data on which linguistic analyses are based are of fundamental importance to the field and should be treated as such. Linguistic data should be citable and cited, and these citations should be accorded the same importance as citations of other, more recognizable products of linguistic research like publications.

1. Importance

In other words…

Linguistic data are important!

Linguistic data are scholarly output.

Linguistic data should be cited like other forms of scholarly output.

2. Credit and Attribution

Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

In linguistics, citations should facilitate readers retrieving information about who contributed to the data, and how they contributed, when it is appropriate to do so. One way to do this is through citations that list individual contributors and their roles. Another way is by using citations that link to metadata about contributors and their roles.

2. Credit and Attribution

In other words…

All contributors to a data set should be recognized when it is ethical to do so.

This can be done in two ways:

1. Use a citation format that lists all contributors and their roles, or
2. Use citations that link to metadata about contributors and their roles.

3. Evidence

In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

Linguists should cite the data upon which scholarly claims are based.

In order for data to be citable, it should be stored in an accessible location, preferably a data archive or other trusted repository. Authors should ensure that data collection and processing methods are transparent, either through links to metadata or a direct statement in the text, to make clear the relationship between the data and the scholarly claims based on it.

3. Evidence

In other words…

Linguistic claims based on data should cite that data!

This implies a data preservation strategy.

Authors should make clear the relationship between the data and the linguistic claims.

4. Unique Identification

A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

When selecting a data repository or other resources for housing and providing access to linguistic data, linguists should look for services that provide the means for identification in the form of a Persistent Identifier (PID). For digital data, examples of these include Digital Object Identifiers (DOI) and Handles.

4. Unique Identification

In other words…

Citations to linguistic data should use persistent identifiers.

These include DOIs or Handles.

Make sure your archive uses persistent identifiers.

5. Access

Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

Linguistic data should be as open as possible, in order to facilitate reproducibility; and as closed as necessary, to honor relevant ethical, legal and speaker community constraints.

5. Access

In other words…

Linguistic data should be

as open-access as possible (sharing is good!)

and as closed as necessary (protecting confidentiality is good!)

6. Persistence

Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.

Linguists should confirm that the archives or repositories where they are storing their data have written policies pertaining to persistence of data, metadata, and identifiers.

6. Persistence

In other words…

Use a real data archive,

one with an institutional commitment to long-term preservation of your data and metadata

7. Specificity and Verifiability

Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

Data citations should make it easy for a curious reader to find the specific datum or subset of data within the larger dataset that support a claim. For data uses that require a fine-grained citation for clarity, a systematic method of identification for the data should be used.

Many data sets are not static; rather researchers add to them all the time. Citations should specify which version of the data is being referenced.

7. Specificity and Verifiability

In other words…

Make reproducible research possible by using citations that

point to a particular subset of data

or

point to a particular version of data

8. Interoperability and Flexibility

Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

Linguists work with a wide range of data, addressing a variety of questions. Citation standards developed for linguistics need to meet the needs of the research community, while also meeting the principles described above.

We encourage linguistics publishers to make data citation easier for their authors by developing data citation formats and to develop clear data policies based on this document.

8. Interoperability and Flexibility

In other words…

Linguists work with a wide range of data types.

Citation formats should accommodate as many types as possible.

Publishers should develop clear data policies and stylesheets for citation formats.

In conclusion

We can learn from each other

- Good news! Different subfields *already* do some things well:
  - **Second-language acquisitionists** describe research methods very well
  - **Sociolinguists, field linguists, and second-language acquisitionists** describe research participants very well
  - **Phoneticians** describe tools, hardware and software very well
  - **Field linguists** describe their fieldwork time very well
  - **Everyone** cites published data sources in the very well

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?
  - 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)

What can you do?

- Join the conversation at the RDA LDIG.
- Read, discuss, comment on, and endorse the Austin Principles.
- Make your data citable…and then cite it.

References

- Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. To appear. Reproducible research in linguistics: A position statement on data citation and attribution in our field. Linguistics.
- Berez-Kroeker, A. L., Andreassen, H. N., Gawne, L., Holton, G., Kung, S. S., Pulsifer, P., Collister, L. B., The Data Citation and Attribution in Linguistics Group, & the Linguistic Data Interest Group. 2018. The Austin Principles of Data Citation in Linguistics (Version 1.0). http://linguisticsdatacitation.org/
- Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker, & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. Language Documentation & Conservation 11: 157-189.
- 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!

www.linguisticsdatacitation.org

bit.ly/LinguisticsDataCitation

Special thanks to **Lauren Gawne, Helene N. Andreassen, Meagan Dailey, Ryan Henke, Gary Holton, Kavon Hooshiar, Susan Kung, Peter Pulsifer,** the participants in the Workshops on Data Citation & Attribution in Linguistics, and the members of the Research Data Alliance Linguistics Data Interest Group. 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.