Data management in the job market & graduate education

Meagan Dailey
medailey@hawaii.edu
University of Hawaiʻi

Ryan Henke
rhenke@hawaii.edu
University of Hawaiʻi
BIG PICTURE: How does data relate to *employability*?

Our barometers

- Job postings
- Course descriptions
1

Job postings

What do employers say they want?
Source: The LINGUIST List

- Faculty positions in US
- Any discipline/field
Numbers over the years

2006
307 posts
-110 Non-faculty
-37 PostDocs
197 total posts

2011
173 posts
-86
-10
87 total

2015
228
-61
-8
159 total

443 total
Trends every year

Qualifications
- PhD
- Published research
- Teaching experience
- Grant success
- Subfield specialization

Job duties
- Publish research
- Develop/teach courses
- Advise students
- Get grants
- Do service
What about DATA?

2006: Specifically addressed in 1/197 posts

- University of Hawaiʻi: Looking for “expertise in computational methods of data manipulation, storage, and dissemination”
What about DATA?

2011: Addressed in 0/87 posts

- Not a good year for faculty jobs in general ...
What about DATA?

2015: Addressed in 6/159 posts

- **Four** institutions
- Not deep details, but there are good signs
- Let’s see what they do say ...
3 linguistics jobs, w/ similar template

“The recent move of Linguistics to the College of Information has positioned the Linguistics Program in a collaborative environment for developing cutting-edge curricula that incorporates research on technology and language learning, language structure and use with database design and management, text mining, and computational and quantitative methods ...”
General call for tenure-track positions, including a spot in computational linguistics

“We are seeking applicants for tenure track positions in interdisciplinary research in data science”

“The University of Rochester has made data science a centerpiece of its 5-year strategic plan ...”
• “The internet and the availability of massive data repositories (e.g., language corpora, large-scale neuroimaging studies, educational data logs) are changing research paradigms in the allied fields of cognitive science. We seek an individual who leverages these resources and/or their own data collection efforts to advance our understanding of a substantive topic within cognitive science …”
• Tenure-track position with a specialization “in very large data resources (so-called ‘Big Data’).”

• Seeking expertise in:
  ▪ “Big Data NLP for social media and the Web”
  ▪ “Multilingual big data for statistical machine translation and cross-language information retrieval”
THE TAKEAWAY

Few posts are talking about data

Not much about reproducible research

But what we do see is positive
Course descriptions

What are we teaching future faculty?
Phase 1: Course Catalogs

- Top 25 Quacquarelli Symonds ranked Linguistics programs
  - 20 US programs
  - 5 international programs

- Surveyed course descriptions

- Any discipline/field
Quick caveat

- Top-25 lists are arbitrary and might not include less visible programs

- Ex) Computational, corpus, language documentation focused programs
QS Universities

- MIT
- UCLA
- Harvard
- Stanford
- UMass
- UC Berkeley
- U Penn
- OSU
- U of Chicago
- Maryland
- NYU
- UC San Diego
- UC Santa Cruz
- UT Austin
- Cornell
- UIUC
- Georgetown
- Boston U
- Brown
- Carnegie Mellon
- Edinburgh
- Cambridge
- Oxford
- National U of Singapore
- U of Hong Kong
What sorts of courses are offered?

Courses dealing with data tended to be:
- Field Methods
- Computational Methods
- Sociolinguistics

No data for 6 universities surveyed
US versus Non-US Universities

Only 9 of 20 US universities have training.

None of the international universities surveyed offer data collection, management, or dissemination training.
Breakdown: 19 schools

Collection:
- 9 universities offer training on data collection

Analysis/Management:
- 7 offer training in annotation and transcription
- 4 offer training in data management, visualization or database design

Preservation/Dissemination:
- 5 offer training in data preservation
  - 1 specifically mentions archiving
- 2 offer training in data dissemination or publication
‘Language Documentation [...] Issues in documenting languages, including collection of primary data using linguistic field methods, organizing data into documents (annotated texts, dictionaries, multimedia presentations, technical articles), audiences for language documents (speakers of target languages, linguists, scholars outside linguistics, general public), presentation and storage of documents (paper publication, online publication, electronic and physical archives), documenting endangered languages, and organizations and initiatives for documenting endangered languages.’
‘Endangered Languages and Language Revitalization [...] will investigate how [trends in language loss] can be reversed by methods of language documentation and description, the use of innovative technologies, multimodal fieldwork, writing dictionaries and grammars for different audiences, language planning, and data creation, annotation, preservation, and dissemination.’
Phase 2: Syllabi

Requested syllabi on data related courses from the top rated Linguistics departments

- Very similar course descriptions
THE TAKEAWAY

Programs offering **field methods** courses addressed issues of data collection and analysis.

With few exceptions, only **computational methods** courses described long-term data management or dissemination.
CONCLUSION

We don’t see a lot about data right now.

But what we do see is headed in the right direction.