Why language data?

In our hyperconnected world, all aspects of our lives are turned into endless streams of data–including language

Yet those of us who study language often find ourselves "on the outside looking in"; we don't get the opportunity to consider our language training as having much to do with data

As **language experts**, we can play an active role in the data-driven technology landscape, bring our expertise to the table, and lead decisions on how language data is used and curated

Why now?

There's much talk about problematic biases in AI and language technologies, and language data plays a big part. Is there anything *we* could do?

The time is ripe for language experts to become a bigger part of the data conversation!

Remarks by Dr. Nora Mattern

Director of the Sara Fine Institute and Assistant Teaching Professor, School of Computing and Information

Chair, **Year of Data and Society** steering committee

Who are you?

Among us today (up here and in the audience) are…

- Current students and Pitt alumni
- Faculty in languages, linguistics, computer & information science, education, English, communication, and the library
- Academic advisors
- Language specialists in industry

We can't wait to hear from you!

Today's symposium

1:15–2:45pm including audience Q&A	**Session 1: Industry Panel** Language professionals in the tech industry: an insider's look
2:45–3:00pm	**Coffee break**
3:00–4:30pm including audience Q&A	**Session 2: Curriculum Panel** Data literacy for language students: curricular recommendations
4:30–5:00pm	**Coffee & Showcase** Industry and community tables, posters
5:00pm	**Wrap**

Industry panel

"A language major?! How is that going to help you get a job??"

This panel brings together experts in language data to help answer that question

In particular, each panelist will discuss:

- **What they do** in their current job
- How language data factors into their **current job** ("language data" is broad!)
- Their **background** in language data
- Perspectives and language data and the hiring process

Then we'll have some Q&A (get your questions ready!)

Panelists

Emily Moline

Curriculum Designer

Duolingo

Justin Glover

Program Manager

Google

Marleigh Bickel

Linguistic Data Specialist

Voci by Medallia

Steve Sloto

Software Engineer

Microsoft

Preview

Here are just a few themes:

- Language data comes in a lot of forms
- Language data is used for different purposes in academia vs. industry
- A lot of learning happens on the job
- You don't have to be a numbers whiz or a computer science expert, but demonstrating your potential for growth makes a big difference
- Industry jobs are pretty cool!

Emily Moline: Intro

- What I do:
  - Curriculum Designer at Duolingo
  - Help drive curriculum/pedagogical improvements to courses big and small
  - Work on nitty-gritty language stuff plus bigger-picture stuff, e.g. training and strategizing
  - Career trajectory: PhD in Linguistics from UC Davis in 2018; first job naming products at a naming agency (really!)
- Use a mix of language data in my current position:
  - Metrics: experiments to test user retention and time spent learning (every change must win!)
  - UX: e.g determining which variety of English to teach with EN<HI
  - Research/curriculum resources that use previous language experts' data to inform decisions

Emily Moline: Background in language data

- Background in language data
  - Qualitative! I don't code!
  - Discourse analysis, community action research, applied sociolinguistics
    - E.g. analyzing the kinds of errors newly literate adults make and the way they talk about literacy itself; conducting interviews with stakeholders
- Language data in academic training vs. in my job
  - Have to focus analyses on things that drive the product vs. getting a detailed look; moves quickly
  - But also go deeper with language data: have to decide on very precise grammatical concepts, functional topics, and vocabulary to teach in a Duolingo lesson—more exacting than classroom

Emily Moline: Language data & hiring

- What Duolingo looks for in hiring language experts
  - Variety of backgrounds; common is knowledge of linguistics and language teaching and experience teaching language
  - Also experimental language data backgrounds in language learning
  - "Soft skills" like project management, ability to apply theoretical knowledge to applied functions, how to communicate to different audiences
- How I translated my skills
  - Non-academic career stuff (just as important, especially for transitioning out of academia!)
    - Fulbright ETA in Spain during PhD and TESL/TEFL teacher at UC Davis most summers. Teaching experience was key for Duolingo job!
    - First job: learned project management and non-academic communication skills

Justin Glover: Intro

- My Role: Program Manager at Google
- What I do:
  - Embedded in a query understanding engineering team
  - Scope and lead cross-functional and cross-organizational projects
  - Partner with engineering manager on team growth and strategy
  - Collaborate with engineers on implementing and improving machine learning models
- How language data factors into my current position:
  - Apply linguistic mindset to analysis of data sets
  - Generate ideas about how to improve losses and initiate projects to address them
  - Evaluate and determine methods for curating labeled data for machine learning models

Justin Glover: Background in language data

- Background:
  - PhD Germanic Linguistics (phonology) from Indiana University
  - Self-taught NLP fundamentals while writing dissertation
- Academic interaction with language data:
  - Theoretical
  - Small data sets and case studies
  - Success measured in terms of advancing the field
- Professional interaction with language data:
  - Applied to solve user problems
  - Large, noisy data sets
  - Success measured in terms of improved user experience

Justin Glover: Language data & hiring

- Important background:
  - Linguistics training (academic or practical)
  - Some technical experience
  - Working in a collaborative environment
  - Functioning as a leader
- How I translated my skills:
  - Demonstrated technical experience: toy projects with Python on GitHub
  - Working in a collaborative environment: joint research work with other grad students and faculty
  - Leadership: led a team of graduate instructors for elementary German courses

Marleigh Bickel: Intro

- Linguistic Data Specialist for Voci by Medallia
- What I do
  - QA of data, transcriptions, & automatic transcription models
  - Writing language rules to be used in machine learning
  - Normalizing data for transcription models
- How I use language Data
  - QA language data to be used in the transcription models
  - Language data is used to train computer models for automatic transcription

Marleigh Bickel: Background in language data

- Academic Background
  - Bachelor's degree in Linguistics and Political Science from the University of Pittsburgh
    - Interned with Voci through Linguistic Dept internship program
  - Certificate in German for Professional purposes from Pitt
- Academic vs Industry differences working with language data
  - Academia
    - more controlled environment
    - Goal is to study certain phenomenon
  - Industry
    - Data has varying levels of quality and validity for technical use
    - Goal is to advance customer experience

Marleigh Bickel: Language data & hiring

- Desired qualifications/skills
  - Linguistics knowledge
  - L2 proficiency
  - Ability to learn new technical skills
- Translating my skills
  - Applying German and linguistics knowledge to real-world examples
  - Showing initiative in new projects
  - Applying previous technical knowledge and building off of it

Steve Sloto: Intro

- My Role: Software Engineer at Microsoft Translator
- What I do:
  - Figure out how to find the best training data possible for machine learning models from web sources.
  - Work on making these techniques reliable and scalable.
  - Improve old, high-impact tools that we use to process language data.
- How language data factors into my current position:
  - Understanding how different languages 'encode' information is useful for anticipating problems.
  - All my programming is focused on working with text data and understanding algorithms and tools for doing so.

Steve Sloto: Background in language data

- Academic Background:
  - Bachelor's Degree in Linguistics and Anthropology from the University of Pittsburgh (2015)
    - No classes from the computer science department 😳
    - Intro to Computational Linguistics, Directed Research, Internships…
- Career Since Then:
  - Fulbright ETA Fellow in Turkey 2015-2016
  - Various roles at Amazon, Microsoft
    - Lots of self-teaching, on-the-job learning, and great mentorship
- Industry vs. Academia
  - Different focuses, different mistakes
    - Linguistic phenomena as an object of study vs. a problem to solve
    - Less focus on understanding nuances of language in industry
    - Industry work is for a product, academic work can be for a project
  - Different tools
    - Larger datasets, bigger computers in industry

Steve Sloto: Language data & hiring

- Desirable Skills:
  - Computer Programming
  - Academic Knowledge
  - Linguistic Knowledge
  - Research Experience
  - Soft skills: inclusivity, growth mindset, leadership
- Getting Hired:
  - Ability to discuss past projects and your specific impact on them.
  - Technical skills: the dreaded coding interview.
  - Talking through problems you might face on the job.
  - Everyone loves an applicant who with knowledge of a less common language!

Panel discussion + audience Q&A

- How does your current job allow you to "scratch the curiosity itch" that draws so many people to studying language?
- What's an example of a skill you've developed since starting your current job?
- What do you wish you'd known about careers in language data when you were a student?
- What do you wish people at your company who make hiring decisions knew about language majors (and what they bring to the table)?
- **For the audience: what questions do you have?**

Contact us

- Dan Villarreal: d.vill@pitt.edu, https://www.linguistics.pitt.edu/people/dan-villarreal
- Emily Moline: moline@duolingo.com, https://www.linkedin.com/in/emily-moline/
- Justin Glover: justin.glover01@gmail.com, linkedin.com/in/justinglover-pgh
- Marleigh Bickel: mbickel@medallia.com, https://www.linkedin.com/in/marleigh-bickel/
- Steve Sloto: ssloto@gmail.com, https://www.linkedin.com/in/ssloto/

Intro/Purpose of this panel

We've brought together faculty and experts in diverse disciplines to put together a set of curricular recommendations:

- Claude Mauk (Pitt Linguistics, Less-Commonly-Taught Languages Center)
- Abdesalam Soudi (Pitt Linguistics)
- David J. Birnbaum (Pitt Slavic)
- Dmitriy Babichenko (Pitt SCI: Department of Informatics and Networked Systems)
- Laura Dickey (Amazon Pittsburgh)

Preview

We've asked each panelist to discuss:

- Courses they teach / offered in their unit that are relevant to data skills building
- Examples of students they mentored who successfully went on to a career in language data, what aspects of their training have been relevant
- Examples of projects that build on language and data skills
- Additional opportunities/gaps they can identify

A view from:
Linguistics, Less-Commonly-Taught-Languages (1/2)

- LCTL Center offers instruction in 14 languages, certificates or minors in 12 of them.
- The majority of people in the world do not speak one of the US's commonly taught languages, but the majority of our resources goes towards teaching them.
- Scholarship opportunities for language learning: Foreign Language Area Studies (FLAS), Critical Languages (CLS), Pitt's Summer Language Institute (SLI)

Claude Mauk
Pitt Linguistics, LCTL

A view from:
Linguistics, Less-Commonly-Taught-Languages (2/2)

- **LING 1000 Intro to Linguistics**: an overview course that covers the basics of how languages work, both structurally and socially. Great for anyone interested in language.
- Core areas: Phonetics, Syntax, Morphology, Semantics, Sociolinguistics
- Courses we offer also cater to students with an interest in computational & data methods:
  - LING 1330 Computational Linguistics
  - LING 1340 Data Science for Linguists
  - Also: Statistics for Linguistics, Research Methods
  - Also: LING 1050 Computational Methods in the Humanities

A view from:
Linguistics, Internship (1/2)

- **Ling Courses:**
  - Ling 1263, Cross-Cultural Communication; Ling 1267, Aspects of Sociolinguistics; Ling 1930 Applications of Linguistics, Ling 1903, Directed Research
- *Overall goals*
  - Accommodate Linguistic and cultural diversity
  - understand & appreciate differences in how people use language
  - understand how linguistic resources create meaning
  - How social relationships influence language and vice versa
  - What drives people to make linguistic choices
  - Linguistic variation
  - Practical applications
- **Working Across Disciplines** (E.g. Humanities in Health @ Pitt)

Abdesalam Soudi
Pitt Linguistics,
Internship

A view from:
Linguistics, Internship (2/2)

- **Ling 1900 (Internship Program, Humanities @ Work) in operation since 2013**
  - Preparing students for versatile careers
  - Fostering cross-disciplinary Collaborations
  - Diverse engagement with language data, Weekly seminar
- **First ever seed grant (2018)**: Engagement Platform for Humanities @ Work
- **Industry Partners**: M*Modal/3M, Voci Tech, Astrata Inc., Magee Women's Hospital, Semantic Compaction Systems, Wikitongues, HCL Google…
- **Interns**: N= 146, 15 Rotations
- **Outcomes**: Job placements in diverse industries, skill-building
- **Internship Grant Program**: Established in 2022
  - 4 awards will be administered in AY 2022-2023
- **Overall goals & future plans**
  - Career pathways for the humanities (new models)
  - Hub at Pitt: Humanities @ work in the community, health & tech industries

A view from:
Slavic, Digital Humanities (1/2)

- Course: "Computational methods in the humanities"
  - Ling 1050 (and 7 other cross-listings; offered every semester)
  - Open-access course site (description, tutorials, activities, projects): http://dh.obdurodon.org
  - For humanists: No prior computing experience required or expected
    - But not only for humanists
  - Honors listing = counts toward an honors degree
  - Graduate cross-listing: Slav 2050 (same course, graduate credit)
  - Outstanding students may be invited to return as undergraduate teaching assistants for academic credit
- Alumni go on to
  - Graduate study in Classics, Computer Science, English, History, Linguistics
  - Faculty positions (English, Slavic) with DH teaching duties
  - Industry positions: Data science, media analytics, web development

David J. Birnbaum
Pitt Slavic Department

A view from:
Slavic, Digital Humanities (2/2)

- Sample projects
  - Developing digital editions and other humanities research-oriented online resources
  - See sample student course projects linked from course website (http://dh.obdurodon.org/)
- Skills acquired
  - Formal modeling (schema languages)
  - Functional and declarative programming paradigms
  - XML technologies: XML, XPath, XSLT, XQuery, SVG, Relax NG, Schematron
  - Web technologies: HTML, CSS
  - Development environment: command line, collaborative project management with Git

A view from: Industry (Professional Paths)

- Courses/job aspirations can inform each other
  - taking a range of classes might spark interests and suggest new job paths
  - an eye on a job path can inform which classes to take
- Are there certain (language-adjacent) things that really excite you?
  - Naming/marketing
  - Coding/machine learning
  - Localization
  - Lab/experimental work
  - Language acquisition/language teaching
  - Corpus work
  - Translation
- I've hired a lot of language analysts - analytical skills are just as important as language skills. I've found that native speakers aren't always sensitive to the patterns and issues in language data.

Laura Dickey
Amazon Pittsburgh

A view from: Industry (Courses)

- Linguistics
  - At least Intro Linguistics for all!
  - Morphology is broadly applicable across industry fields (phonetics and syntax also great)
  - Language acquisition, particularly second language acquisition
- Engineering/design
  - Python
  - SQL
  - User experience design (design thinking, human-computer interaction, web/mobile design)
- Language education
  - Language teaching
  - Online learning

A view from: Industry (Courses)

- Data structures/handling
  - Library science
  - Statistics
  - Data visualization
  - Experimentation (A/B testing, experiment design)
- Business
  - project management: lots of translation, localization, and language software work which needs people who can organize people and projects, but who also understand the problem space
  - product management covering the business side of language products
  - marketing

Summary: Data skills for language students

In addition to their language specialization, students will want:

✅ Generalist knowledge of language (= linguistics!)

✅ Computational thinking

✅ Experiential learning

These skillsets help transfer **language competency** into **data domain**

Also: offer a **competitive edge** over native speakers without such skills

Recommendations for students: **courses**

Start with:

- LING 1000: Intro to Linguistics
- CS 12: Intro to Computing for the Humanities

The earlier the better!
Freshman or sophomore

Further:

- Linguistics:
  - LING 1330 Computational Linguistics
- Computer & Information Science:
  - INFSCI 210: Intermediate Programming with Python
- Digital Humanities:
  - LING 1050/SLAV 1050… Computational Methods for the Humanities
- Also: statistics, business marketing

⚠ You do *not* need *every* course!
Explore a path based on your interest

Recommendations for students: **academic programs**

In addition to their language study, students should consider adding:
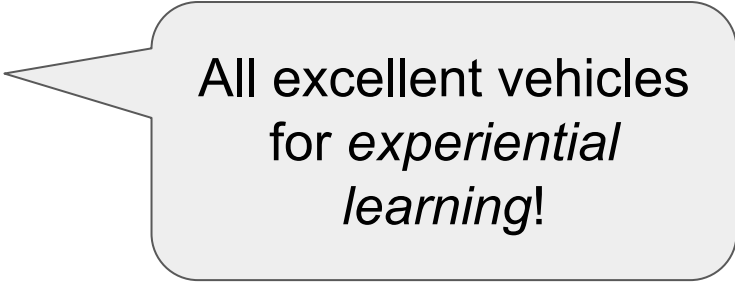
- Linguistics minor, major
- Computer Science minor (cf. CS + language double degree)
- Undergrad joint major in Data Science
- Computational Linguistics undergrad certificate (IN THE WORKS!)

Grad school:

- Linguistics MA with statistics/data science focus
- Computational linguistics graduate programs (list at bottom of this page)

Recommendations for students: **engagement**

- Internship
- Directed research
- Project participation
- UG teaching assistantship
- Communities: PyLing

All excellent vehicles for *experiential learning*!

Recommendations for language programs:
pedagogy & curriculum, advising

- Curriculum development:
  - Language lessons designed to spark students' interest in language as data?
  - What are good data-oriented lessons/modules that language programs can incorporate in their curriculum?

- Upcoming: Robert Henderson Language Media Center will offer a workshop on designing data-themed language modules!
  - Corpus-based lessons
  - Review of multilingual technologies, peek under-the-hood
  - Tour of multilingual datasets (e.g., Universal Dependencies)

Language faculty: watch for invitation!

Universal Dependencies

Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 300 contributors producing nearly 200 treebanks in over 100 languages. If you're new to UD, you should start by reading the first part of the Short Introduction and then browsing the annotation guidelines.

- Short introduction to UD
- UD annotation guidelines
- More information on UD:
  - How to contribute to UD
  - Tools for working with UD

Current UD Languages

Information about language families (and genera for families with multiple branches) is mostly taken from WALS Online (IE = Indo-European).

▶	Afrikaans	1	49K		IE, Germanic
▶	Akkadian	2	25K		Afro-Asiatic, Semitic
▶	Akuntsu	1	<1K		Tupian, Tupari
▶	Albanian	1	<1K	W	IE, Albanian
▶	Amharic	1	10K		Afro-Asiatic, Semitic
▶	Ancient Greek	2	416K		IE, Greek
▶	Apurina	1	<1K		Arawakan
▶	Arabic	3	1,042K	W	Afro-Asiatic, Semitic
▶	Armenian	2	55K		IE, Armenian
▶	Assyrian	1	<1K		Afro-Asiatic, Semitic
▶	Bambara	1	13K		Mande
▶	Basque	1	121K		Basque
▶	Beja	1	1K		Afro-Asiatic, Cushitic
▶	Belarusian	1	305K	W	IE, Slavic
▶	Bengali	1	<1K		IE, Indic

> An example of a massively multilingual dataset used in language technologies!

45

```
1    # sent_id = dev-s1
2    # text = 同樣，施力的大小不同，引起的加速度不同，最終的結果也不一樣，亦可以從向量的加成性來看。
3    1    同樣    同樣    ADV     RB      _       12      advmod      _           SpaceAfter=No
4    2    ，      ，      PUNCT   ，      _       12      punct       _           SpaceAfter=No
5    3    施力    施力    VERB    VV      _       5       acl:relcl   _           SpaceAfter=No
6    4    的      的      PART    DEC     _       3       mark:rel    _           SpaceAfter=No
7    5    大小    大小    NOUN    NN      _       6       nsubj       _           SpaceAfter=No
8    6    不同    不同    ADJ     JJ      _       8       csubj       _           SpaceAfter=No
9    7    ，      ，      PUNCT   ，      _       8       punct       _           SpaceAfter=No
10   8    引起    引起    VERB    VV      _       11      acl:relcl   _           SpaceAfter=No
11   9    的      的      PART    DEC     _       8       mark:rel    _           SpaceAfter=No
12   10   加速    加速    VERB    VV      _       11      compound    _           SpaceAfter=No
13   11   度      度      PART    SFN     _       12      nsubj       _           SpaceAfter=No
14   12   不同    不同    ADJ     JJ      _       0       root        _           SpaceAfter=No
15   13   ，      ，      PUNCT   ，      _       12      punct       _           SpaceAfter=No
16   14   最終    最終    NOUN    NN      _       16      nmod        _           SpaceAfter=No
17   15   的      的      PART    DEC     Case=Gen        14      case        _           SpaceAfter=No
18   16   結果    結果    NOUN    NN      _       29      nsubj       _           SpaceAfter=No
19   17   也      也      ADV     RB      _       19      mark        _           SpaceAfter=No
20   18   不      不      ADV     RB      Polarity=Neg    19      advmod      _           SpaceAfter=No
21   19   一樣    一樣    ADJ     JJ      _       29      advcl       _           SpaceAfter=No
22   20   ，      ，      PUNCT   ，      _       29      punct       _           SpaceAfter=No
23   21   亦      亦      ADV     RB      _       29      mark        _           SpaceAfter=No
24   22   可以    可以    AUX     MD      _       29      aux         _           SpaceAfter=No
25   23   從      從      ADP     IN      _       27      case        _           SpaceAfter=No
26   24   向量    向量    NOUN    NN      _       27      nmod        _           SpaceAfter=No
```

One of many forms language data can take!

Panel discussions + audience Q&A

Q: What are biggest barriers for language students to trying technical (=CS, programming) courses? There's often reluctance, how to overcome it?

Q: How long does it really take to acquire sufficient data skills?

Contact us

- Na-Rae Han: https://sites.pitt.edu/~naraehan/, naraehan@pitt.edu
- Claude Mauk: https://www.linguistics.pitt.edu/people/claude-e-mauk
- Abdesalam Soudi: https://www.linguistics.pitt.edu/people/abdesalam-soudi
- David J. Birnbaum: http://www.obdurodon.org, djbpitt@gmail.com
- Dmitriy Babichenko: https://www.sci.pitt.edu/people/dmitriy-babichenko
- Laura Dickey: https://www.linkedin.com/in/lauradickey/