

Tara (Yi-Fang) Cheng

🏠 yifangcheng.me | 📧 cheng627@purdue.edu | in yifangchengg

Computational linguist with experience designing multilingual data-collection, annotation, and analysis workflows using Python, R, and SQL.

EDUCATION

Purdue University

West Lafayette, IN

Ph.D. in Linguistics (Exp. May 2027); Graduate Certificate in Applied Statistics; MA in Linguistics

- Dissertation: Evaluated bilingual speech patterns using acoustic measures and layer-wise wav2vec 2.0 probing with speaker-independent data splits and cross-validation.

National Taiwan Normal University

Taipei, Taiwan

BA in English, BM in Music

RESEARCH & TECHNICAL EXPERIENCE

Purdue Bilingualism Lab

West Lafayette, IN

Graduate Researcher

Aug. 2022 – Present

- Led end-to-end collection and analysis of original and corpus-based multilingual speech datasets totaling 100+ hours across Mandarin, English, Korean, and Vietnamese.
- Managed the full research lifecycle, including corpus selection, participant recruitment, recording, annotation, feature extraction, statistical modeling, and reporting.
- Designed annotation workflows, label sets, coding guidelines, exclusion criteria, and automated annotation extraction for both self-collected and existing speech corpora.
- Built reproducible Python and R pipelines to clean, join, validate, and analyze multilingual speech and text data, and communicated findings through 4 publications and 10+ conference presentations.

NSF National Innovation Corps

Remote

Entrepreneurial Lead

Jun. 2025 - Jul. 2025

- Led customer discovery for an AI-powered language-learning product (Link), conducting 100+ stakeholder interviews to identify user needs, adoption barriers, bias concerns, and product-market fit.
- Synthesized qualitative interview data into weekly strategic recommendations and presented progress updates to NSF program officers.
- Secured \$50,000 in NSF I-Corps funding.

The Data Mine

West Lafayette, IN

Graduate Data Science Researcher – Knudsen Institute

Aug. 2024 – Dec. 2024

- Project: Training LLMs within Manufacturing (Link)
- Designed data pipelines to scrape, clean, and preprocess manufacturing text data for LLM training workflows.

Graduate Data Science Researcher – C-SPAN

Jan. 2024 – May 2024

- Project: Analysis of Rhetoric on Capitol Hill (Link)
- Built a Bash, Python, Praat, and Whisper API workflow to process congressional speech and extract acoustic measures for quantitative analysis of rhetorical patterns.

WORK EXPERIENCE

VoiceTube Inc.

Taipei, Taiwan

Curriculum and Content Intern

Feb. 2022 – Jul. 2022

- Translated, edited, and quality-checked 60+ English videos for a digital platform serving 4M+ users worldwide.
- Responded to individual learners' language and pronunciation questions, providing clear, personalized explanations and corrective feedback.
- Supported digital product and content development by designing language-learning materials and podcast-based pronunciation content.

SKILLS

Programming: Python, R, SQL, Bash

Tools: Git, GitHub, Hugging Face, Jira, Praat, Montreal Forced Aligner

Languages: Mandarin (Native), English (Full professional), Southern Min (Conversational)