Taiqi He

taiqi.xyz | taiqihe@gmail.com

EDUCATION

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA

2023/08-2025/01

- Ph.D. in Language and Information Technology - Incomplete

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA

2021/08- 2023/08

- Master of Language Technologies

University of California Davis, Davis, CA

2015/09- 2019/06

- Bachelor of Science in Cognitive Science, Computational Emphasis, with Highest Honors
- Bachelor of Art in Linguistics, with Highest Honors

EMPLOYMENT AND INTERSHIPS

Student Researcher, Wav2Gloss Project, CMU 2022/09-2025/01

Supervisor: Lori Levin, Research Professor, Language Technologies Institute

- Created end-to-end multi-modal models that generate four levels of annotations from speech signals on low-resource languages.
- Showed that zero-shot unseen language translation with LLMs can achieve good performance.
- Collected and normalized speech data from existing annotated field linguistics datasets.

Student Researcher Internship, Google 2024/05-2024/08

Supervisors: Sam Nguyen; Sergey Levi

- Created text-to-speech and audio transformation normalizing flow models with control of the accent of the speaker.
- Curated a dataset of diverse spoken English accents from an open source.

Student Researcher, AIDA/OPERA Project, CMU 2021/10-2022/08

Supervisor: Yonatan Bisk, Assistant Professor, Language Technologies Institute

- Created models that extracted structured information (including claims, sources, veracity of information) from news articles.
- Dockerized existing pipelines and data.

Junior Specialist, Computational Cognitive Neuroscience Lab, UC Davis 2019/11-2021/07

Supervisor: Randall O'Reilly, Professor, Center for Neuroscience

- Adapted the DeepLeabra predictive learning neural model for the language modeling task.
- Used self-organizing maps to discover the emergent structures of word clusters through word embeddings and compared the topological organizations of word embeddings to neural data.
- Used embodied language models in 2D grid worlds to model linguistic compositionality.

RESEARCH EXPERIENCE

Correlation analysis between the brain and computational linguistics models

Sep 2018 - Jun 2019

Advisor: Steve Luck, Professor, Department of Psychology

- Adapted representational similarity analysis (RSA) for EEG/ERP data and language embeddings.
- Showed consistent correlation between brain activity and word embeddings, indicating that natural language processing models share structural similarities with the brain without intentional designs.

High dimensional vector representation of languages with unsupervised learning

Jan 2018 - Jun 2019

Advisor: Kenji Sagae, Assistant Professor, Department of Linguistics

- Created language embeddings from plaintext corpora with unsupervised methods.
- Showed validity of the language embeddings in typological and machine translation tasks and demonstrated their superior performance over the baseline and naïve approaches.

SELECTED PUBLICATIONS

- **Wav2Gloss: Generating Interlinear Glossed Text from Speech.**
  Taiqi He, Kwanghee Choi, Lindia Tjuatja, Nathaniel Robinson, Jiatong Shi, Shinji Watanabe, Graham Neubig, David Mortensen, and Lori Levin (2024).
  In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics.
- **Construction Grammar Provides Unique Insight into Neural Language Models.**
  Leonie Weissweiler, *Taiqi He*, Naoki Otani, David R. Mortensen, Lori Levin, and Hinrich Schuetze (2023).
  Georgetown University Round Table on Linguistics 2023.
- **Neural Correlates of Word Representation Vectors in Natural Language Processing Models: Evidence from Representational Similarity Analysis of Event-Related Brain Potentials.**
  Taiqi He, Megan A. Boudewyn, John E. Kiat, Kenji Sagae, and Steven J. Luck (2021).
  Psychophysiology.
- **Language Embeddings for Typology and Cross-lingual Transfer Learning.**
  Yu Dian[*], *Taiqi He*[*], and Kenji Sagae (2021).
  In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.

SKILLS

- Proficient in Python, C++, Java.
- Familiar with common Python machine learning frameworks, including PyTorch, Transformers, ESPNet.
- Contributed to open-source projects (ESPNet, OpenCLIP).

[*] Equal Contributions