Taiqi He

taiqih@andrew.cmu.edu

EDUCATION

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA

Sep 2023-Dec 2026

- Ph.D. in Language and Information Technology

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA

Sep 2021-Aug 2023

- Master of Language Technologies

University of California Davis, Davis, CA

Sep 2015-Jun 2019

- 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

Sep 2022 - present

Supervisor: Lori Levin, Research Professor, Language Technologies Institute

- Creating end-to-end systems that generate morphosyntax analysis (glossing) from acoustic signals on low-resource languages.
- Collecting and normalizing data from existing annotated speech datasets.

Student Researcher, AIDA/OPERA Project, CMU

Oct 2021 - Aug 2022

Supervisor: Yonatan Bisk, Assistant Professor, Language Technologies Institute

- Claim-frame extraction on novel topics.
- Claimers and epistemic information extraction using pretrained language.
- Dockerization of existing pipelines.

Junior Specialist, Computational Cognitive Neuroscience Lab, UC Davis

Nov 2019 - Jul 2021

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

- Adapted DeepLeabra based predictive learning neural model to learn a language model.
- 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.

Undergraduate Research Assistant, Luck Lab, UC Davis

Sep 2016 - Sep 2018

Supervisor: Steve Luck, Professor, Department of Psychology

- Programmed behavioral experiments with Psychtoolbox and conducted experiments.
- Collected and analyzed EEG data with Matlab (EEGLAB, ERPLAB).

RESEARCH EXPERIENCE

Constructions in pretrained language models

Sep 2021 - present

Adviser: Lori Levin, Research Professor, Language Technologies Institute

- Investigating linguistic constructions that are mappings from discontinuous markers to meanings, as encoded by large language models, employing methods that include classification, clustering, and usage-based testing.
- Extracting linguistic strategies from Universal Dependencies treebanks.

Annotation of coreference and bridging in dialogue

Dec 2021 - Jun 2022

Adviser: Lori Levin, Research Professor, Language Technologies Institute

- Annotated parts of the dataset for the CODI-CRAC shared task 2022.

Emergent structures from language models

Nov 2019 - Jul 2021

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

- Used Kohonen models to generate self-organizing maps of words from pretrained word embeddings.
- Compared the topology between the Kohonen maps and fMRI maps.

Correlation analysis between the brain and computational linguistics models

Sep 2018 - Jun 2019

Adviser: 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.
- Currently finishing up the analysis and preparing a paper manuscript for publication.

High dimensional vector representation of languages with unsupervised learning

Jan 2018 - Jun 2019

Adviser: 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 outperformance over the baseline and naïve approaches
- Provided an easier approach for future typological research, especially on lower resource languages
- Explored the distribution of word embeddings derived from different source data
- Explored generating paraphrases with sequence-to-sequence models

PUBLICATIONS

- **SigMoreFun Submission to the SIGMORPHON Shared Task on Interlinear Glossing.**
  Taiqi He, Lindia Tjuatja, Nathaniel Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig, and Lori Levin (2023).
  In Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology.
- **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 (PyTorch, Transformers, ESPNet), C++, Java

[*] Equal Contributions