Natt KORAT

Phum Phsar Touch, Sangkat Toul Sangke, Khan Russey Keo, Phnom Penh 12105, Cambodia

natt.korat@cadt.edu.kh +855-68-355-899

GitHub

About Me

I am a lecturer and researcher at the Cambodia Academy of Digital Technology (CADT) working on speech and language technologies for the Khmer language. My research focuses on automatic speech recognition, natural language processing, and information extraction for low-resource languages. I have experience developing end-to-end AI research pipelines including data collection, dataset annotation, model training, and evaluation. My recent work focuses on building Khmer speech recognition systems and multilingual event extraction pipelines for epidemic surveillance using news and social media data. I am particularly interested in designing robust NLP and speech systems for multilingual and under-resourced language environments. Alongside research, I teach undergraduate computer science courses including Object-Oriented Programming, Algorithms, and Data Structures.

Research Interests

Speech Recognition for Low-Resource Languages, Natural Language Processing, Information Extraction and Event Extraction, Multilingual Language Models, Machine Learning for Social Good

Education

Master of Computer Science *May 2024 – Feb 2026*
Cambodia Academy of Digital Technology (CADT)
Relevant Research: AI-driven Event Extraction for Epidemic Surveillance

Bachelor of Computer Science *Aug 2020 – Jul 2024*
AGA Institute (AI)

Bachelor of Psychology *Sep 2018 – Jul 2022*
Royal University of Phnom Penh (RUPP)

Professional Experience

Lecturer–Researcher *Jan 2026 – Present*
Cambodia Academy of Digital Technology, Cambodia

- Conduct research on Khmer speech and language technologies, focusing on automatic speech recognition and NLP for low-resource languages.
- Develop datasets, experimental pipelines, and machine learning models for speech and text processing.
- Design and evaluate transformer-based models for multilingual language processing tasks.
- Teach undergraduate courses including Object-Oriented Programming, Algorithms, and Data Structures.

Research Exchange Intern *Jun – Dec 2025*
Burapha University, Thailand

- Conducted research on epidemic event extraction using large language models for public health surveillance.
- Designed multilingual information extraction pipelines for analyzing health-related news and social media data.
- Evaluated large language models using zero-shot and few-shot prompting strategies for event extraction.

Lecturer–Research Assistant *Dec 2022 – Jun 2025*
Cambodia Academy of Digital Technology, Cambodia

- Contributed to multiple research projects including Khmer License Plate Recognition and Khmer Large Language Models.
- Assisted in teaching courses including Object-Oriented Programming, Algorithms, and Data Structures.
- Participated in dataset development, model training, and research experimentation.

Data Engineer Intern *Aug – Nov 2023*

Z1 Data Co., LTD, Cambodia

- Developed data collection pipelines and performed annotation for satellite imagery in the Cambodia Building Footprint Project.

Research Projects

Khmer Automatic Speech Recognition *Jan 2026 – Present*

- Constructed large-scale Khmer speech datasets from multiple audio sources and performed data cleaning and normalization.
- Fine-tuned and evaluated Wav2Vec2 XLS-R models for Khmer automatic speech recognition.
- Conducted experiments on dataset expansion and text normalization to improve recognition performance.

AI-driven Epidemic Event Extraction for Surveillance *Jun 2025 – Dec 2025*

- Designed a multilingual event extraction pipeline for epidemic surveillance using news and social media sources.
- Built annotated datasets for Named Entity Recognition and event extraction focusing on disease, pathogen, and symptom entities.
- Evaluated large language models for event extraction using zero-shot and few-shot prompting strategies.
- Fine-tuned transformer models (XLM-RoBERTa) for trigger detection and argument extraction.
- Developed a human-in-the-loop evaluation interface for validating and correcting extracted events.

Khmer Natural Language Processing Resources *Mar 2025 – Aug 2025*

- Developed datasets for Khmer Named Entity Recognition in the health domain.
- Created annotation guidelines for low-resource language NLP tasks.
- Applied transformer-based models for multilingual NLP tasks.

Khmer Large Language Models *Jan 2024 – Dec 2025*

- Developed a custom pipeline to extract only Khmer data from Common Crawl corpus.
- Developed a crawler pipeline using Scrapy to scrape target Khmer websites.
- Developed a data pre-processing pipeline to clean collected data.
- Produced statistics for the collected data.

Khmer License Plate Recognition System *Jan 2023 – Dec 2023*

- Developed datasets for Khmer Khmer License Plate Recognition.
- Created annotation guidelines for hiring students to annotate the license plate data using CVAT annotation tool.
- Fine-tuned YOLOv5 for location detection and serial number segmentation on license plate and used Parseq-OCR for extracting serial numbers.

Publications

Journal Articles

- **Korat, N.**, Heang, S., & Lay, V. (2025). "Khmer News Classification in Low-Resource Settings: A Comparative Analysis of Embedding Methods." *Journal on Information Technologies & Communications.*

Conference Papers

- Chiep, C., **Korat, N.**, Lay, V., & Ly, R. (2025). "Building a Khmer NER Benchmark from Health News Data Towards Event Extraction." *ASEAN Conference on Emerging Technology (ACET 2025).*
- **Korat, N.**, Kak, S., Lay, V., Uraiwan, B., & Waranrach, V. (2025). "Epidemic Event Extraction from News Media Using Large Language Models." *International Conference on Digital Economy and Fintech Innovation (DEFI 2025).*

Posters

- Kong, P., **Korat, N.**, and Veng, P. (2024). "Robust and Efficient Recognition of Khmer License Plates Using YOLOv5 and Parseq OCR." *CV4DC Workshop, Asian Conference on Computer Vision (ACCV).*
- **Korat, N.**, Waranrach, V., & Kak, S. (2025). "Toward Multilingual Epidemic Event Extraction for Low-Resource Languages." *UEC International Seminar Poster.*

Technical Skills

Programming: Python, C++, C, Java, SQL, JavaScript

Machine Learning/NLP: PyTorch, TensorFlow, HuggingFace Transformers, Scikit-learn

Framework: FastAPI, Django, Node.js, Laravel

Tools: Git, Linux, Docker, Scrapy, Label Studio

Languages

- **Khmer**: Native
- **English**: Upper-Intermediate (B2)
- **French**: Beginner (A1)