NHU VO

NLP & LLM Researcher · Computer Science Educator · PhD Student, UTS & VinUniversity

Sydney, NSW, Australia | vodiepnhu@gmail.com | Website | LinkedIn | Google Scholar | Project Page

Work rights: Student visa (subclass 500); no limit on working hours as a PhD student.

7 years
NLP research, language-data and university teaching appointments, 2017–2025

7
international conference papers, incl. EMNLP, LREC-COLING, LREC and Interspeech

8
institutions taught at, incl. English-medium teaching at VinUniversity

PROFILE

NLP researcher and university educator with seven years of language-data and NLP research and teaching appointments spanning 2017–2025. My PhD (UTS–VinUniversity) develops datasets, models and evaluation workflows for Vietnamese–English medical translation and biomedical information extraction: I led **MedEV** (about 360K medical sentence pairs) and co-led **En-ViMedNER** (202,949 aligned entities across 21 UMLS types). Since September 2025 I have also worked in Sydney as a Research Assistant at the University of Technology Sydney, contributing to the En-ViMedNER (*Findings of EMNLP 2026*) and ViMedCSS (*LREC 2026*) papers. I have taught programming, computer science, NLP and information systems at eight institutions, including English-medium teaching at VinUniversity, and managed applied projects with clinicians, government departments and software partners.

Open to: research positions (Research Assistant/Associate, postdoctoral), lecturing roles and NLP/LLM roles in industry.

CORE STRENGTHS

Research	Biomedical and low-resource NLP; NMT and LLM-based translation; cross-lingual NER; LLM evaluation and translation reliability; dataset and benchmark construction.
Engineering	Python, PyTorch, Hugging Face Transformers; fine-tuning, prompting and evaluation; corpus construction, Label Studio and LLM-assisted annotation; hands-on use of LangChain, LangGraph, Docker and AWS in prototypes.
Teaching	Programming, computer science, NLP and information systems; curriculum and syllabus design; blended and online delivery (Moodle, Canvas, Google Classroom, Articulate Storyline 360, SCORM); student research supervision.
Delivery	Project coordination with clinicians, government departments and software vendors; quality assurance; technical writing.

EDUCATION

PhD, Computer Science (Collaborative Research Degree)
Apr 2023 – present
University of Technology Sydney & VinUniversity

- **Topic:** NLP and LLMs for biomedical and low-resource machine translation.
- **Supervisors:** Prof. Wray Buntine (Monash University); Asst Prof. Dung D. Le (VinUniversity); Prof. Massimo Piccardi and Dr Iñigo Jauregi Unanue (University of Technology Sydney).

MSc, Computer Science
2018 – 2020
VNU-HCM University of Science (VNU-HCMUS)

- Thesis: *Vietnamese Text Corpus Mining* (10/10).

BEd, Computer Science Teacher Education
2013 – 2017
HCMC University of Education

- Thesis graded 9.5/10.

Teaching qualification: Certificate of Pedagogical Training for College and University Lecturers, Saigon University (Dec 2020).

RESEARCH & PROFESSIONAL EXPERIENCE

Research Assistant – NLP & LLMs
Sep 2025 – present

University of Technology Sydney · Sydney, Australia

Medical NLP research in a UTS–VinUniversity collaboration, working with UTS researchers Prof. Massimo Piccardi and Dr Iñigo Jauregi Unanue.

- Co-led **En-ViMedNER**: 44,892 sentence pairs from 4,392 PubMed abstracts and 202,949 aligned entity mentions across 21 UMLS types. Translations post-edited by doctors and medical students; LLM-projected entity spans human-checked, reaching **98.32% entity exact match** on a 450-sentence reviewed mini-test (κ = 0.83). Co-first author, *Findings of EMNLP 2026*.
- Co-developed **ViMedCSS**, a Vietnamese medical code-switching speech dataset and benchmark (34 hours; 16,576 utterances). Co-author, *LREC 2026*.

Research Assistant & Project Coordinator
Apr 2023 – present

VinUniversity, College of Engineering & Computer Science · Hanoi, Vietnam

Concurrent appointment within the UTS–VinUniversity collaborative PhD program; medical MT group with Prof. Wray Buntine and Asst Prof. Dung D. Le, in collaboration with VinUniversity College of Health Sciences.

- Built **MedEV** (about 360K English–Vietnamese medical sentence pairs); fine-tuned and benchmarked Transformer translation models in both directions. Co-first-author paper, LREC-COLING 2024.
- Contributed to two *Interspeech 2026* papers on robust speech translation and code-switching ASR.
- Coordinate the *Robust Vietnamese–English Clinical and Educational Medical Translation* project (Oct 2024 – present), including a real-time medical speech-translation prototype.

Project Manager & Research Assistant
Apr 2022 – present

Smart Green Transformation Center (GREEN-X), VinUniversity

- Managed a web/mobile monitoring and evaluation platform with Bamboo Software; coordinated requirements and quality assurance with **10 Khanh Hoa provincial departments** (Nov 2024 – Nov 2025).
- Lead domain-specific Vietnamese corpora for computational social science: anti-corruption discourse (180K+ news texts, 50K+ social-media comments) and the electric-vehicle transition; managed a 1,622-talk business-speech corpus (12.5M+ words).

Remote AI Trainer – Vietnamese Writing
Apr 2024 – Sep 2025

Outlier · also Remote Vietnamese Transcriber, Appen (Feb 2022 – Sep 2025)

- Wrote Vietnamese prompts and evaluated generative-AI responses against quality guidelines; transcribed and annotated Vietnamese audio for speech-data projects.

Researcher, Computational Linguistics Center
Nov 2019 – Dec 2020

VNU-HCM University of Science

- Built **VnC 1.0**, a balanced 100M-word Vietnamese corpus (10% spoken, 90% written) with word segmentation, POS and NER annotations; worked on English–Vietnamese MT frameworks and corpus-based authorship and genre analysis.

Open resources: En-ViMedNER, MedEV and ViMedCSS on Hugging Face; VnC 1.0, ViLEC and the Business Speeches Corpus (12.5M+ words) available on request.

TEACHING EXPERIENCE

Visiting Faculty (CBM) & Teaching Assistant (CECS)
Apr 2023 – Sep 2025

VinUniversity, Hanoi · English-medium teaching

- **Visiting Faculty**, College of Business and Management (Oct 2024 – Sep 2025): taught *Management Information Systems* and *Spreadsheet Modelling for Business* to undergraduates.
- **Teaching Assistant**, College of Engineering & Computer Science (Apr 2023 – Sep 2025): tutorials and assessment for *Natural Language Processing; Big Ideas: AI and the Knowledge Revolution; Research Communication; Health Informatics; Introduction to Programming with Python; Computational & Algorithmic Thinking*.

Lecturer, Faculty of Information Technology
Sep 2022 – Apr 2023

HCMC University of Foreign Languages – Information Technology (HUFLIT)

- Taught *Programming Fundamentals*, *Programming Techniques*, *Applied Informatics in Tourism & Travel* and *Computer Skills*; mean student evaluation **4.72/5**.

Lecturer, Department of English Teaching Methodology — Apr 2021 – Feb 2022

Nguyen Tat Thanh University (NTTU), Faculty of Foreign Languages

- Designed syllabi and learning resources for *E-learning in English Language Teaching* and *Instructional Technology in English Language Teaching*; taught both units in English.
- Guided **65 student research projects** on online course design; supervised two undergraduate theses (2021, 2022) and Business English internships. Led **ViLEC**, a learner corpus of 258 student documents (267,512 words; 15,129 sentences).

Visiting Lecturer, Faculties of Information Technology — May 2020 – Apr 2023

HCMC University of Economics and Finance (UEF), 2021–2023 · VNU-HCM University of Science, 2020–2021

- **UEF:** taught *Introductory Computer Science* and coordinated online delivery through the university LMS.
- **VNU-HCMUS:** practice lecturer for *Introduction to NLP* (High-Quality Program) and tutor for *Computer Hardware* (Advanced Program, in English); developed flipped-classroom syllabi for *Statistical Linguistics* and *Text Mining and Applications*.

Lecturer (Team Leader from Aug 2019) — May 2017 – Feb 2020

HCMC University of Education, Information Technology Center

- Taught *Basic Informatics*; developed online curricula, syllabi and e-learning scripts aligned with approved course outlines.
- Led the university team building and rolling out **VLE.HCMUE** with ACP Singapore, funded by the World Bank and Vietnam's Ministry of Education and Training. Developed Moodle courses for **3,000+ high-school teachers**; trainer in ETEP and the National Foreign Language 2020 Project.

Other teaching

- Visiting Lecturer, *Web Interface Design*, Ly Tu Trong College and District 12 Technical-Economic College (2022).
- Head of Kits-based Department, SUMATO Academy (Jan – Jun 2020): Scratch, LEGO Mindstorms EV3 and Python curricula; trained teachers.

PUBLICATIONS

Peer-reviewed international conference papers

1. **Nhu Vo**[*], Phuong Nguyen[*], Nu Uyen Phuong Le, Iñigo Jauregi Unanue, Dung D. Le, Massimo Piccardi, and Wray Buntine (2026). En-ViMedNER: An English–Vietnamese parallel biomedical corpus with UMLS semantic type annotations. *Findings of EMNLP 2026*.
2. Giang Son Nguyen, Tung X. Nguyen, Hieu Minh Truong, **Nhu Vo**, Wray Buntine, and Dung D. Le (2026). PiDA: Phonetically-informed data augmentation for robust Vietnamese speech translation. *Interspeech 2026*.
3. Tung X. Nguyen[*], Hieu Minh Truong[*], Giang Son Nguyen, **Nhu Vo**, Wray Buntine, and Dung D. Le (2026). Contrastive training with LLM-generated near-misses for robust code-switching speech recognition. *Interspeech 2026*.
4. Tung X. Nguyen, **Nhu Vo**, Giang-Son Nguyen, Duy Mai Hoang, Chien Dinh Huynh, Iñigo Jauregi Unanue, Massimo Piccardi, Wray Buntine, and Dung D. Le (2026). ViMedCSS: A Vietnamese medical code-switching speech dataset & benchmark. *LREC 2026*.
5. **Nhu Vo**[*], Dat Quoc Nguyen[*], Dung D. Le, Massimo Piccardi, and Wray Buntine (2024). Improving Vietnamese–English medical machine translation. *LREC-COLING 2024*, 8955–8962.
6. **Nhu Vo**, Thuy Bui, and Dien Dinh (2020). Text-type variation in Vietnamese: Corpus mining for linguistic features in narrative and non-narrative genres. *NICS 2020* (IEEE), 221–225.
7. Lam Ho, **Nhu Vo**, Dien Dinh, and Nhung Nguyen (2020). Identifying authors based on stylometric measures of Vietnamese texts. *PACLIC 34*, 447–452.

Journal articles

8. **Nhu Vo**, Thuy Bui, Tran Phan, and Dien Dinh (2022). Vietnamese corpus construction for teaching Vietnamese to foreigners. *Linguistic Journal* (Vietnam Institute of Linguistics), 375(1), 13–22.
9. Long Le and **Nhu Vo** (2019). Pedagogical script: Challenges to be resolved in e-learning. *HCMC University of Education Journal of Science*, 16(12), 947–960.

National conference proceedings (Vietnam)

10. **Nhu Vo** and Dien Dinh (2020). Language sampling criteria for Vietnamese corpus construction. *FAIR 2020*, 26–32.

11. **Nhu Vo** and Dien Dinh (2020). Building a Vietnamese corpus for dictionary compilation. *Vietnamese Dictionary and Encyclopedia: Theory and Practice*, 423–433.

Manuscripts

- **First author.** *ViLEC: Design, Annotation, and Pedagogical Applications of a Vietnamese Learner English Corpus*. Under review, *SAGE Open*.
- **Co-author.** *Data-Driven Equity-Sensitive Urban Mobility Planning: LLM-Based Policy Appraisal of Ho Chi Minh City's Two-Wheeler Electrification*. Under review, *Journal of Urban Mobility*.
- **Co-author.** *Anti-corruption discourse across news and social media in Vietnam (2017–2025): A corpus-assisted critical analysis*. Ready for submission.

RESEARCH PROJECTS

VinUniversity

- *Robust Vietnamese–English Clinical and Educational Medical Translation* — **Project Coordinator** 2024 – present
- *Domain-specific Vietnamese Corpora: Anti-Corruption; EV Shift* — **Co-investigator** 2025 – present
- *Green Transformation M&E System, Khanh Hoa* — **Project Manager** 2024 – 2025
- *Business Speeches Corpus* — **Co-investigator** 2022 – 2023

Earlier

- *ViLEC*, NTTU — **Lead** 2021 – 2022
- *VnC 1.0*, VNU-HCMUS — **Lead** 2020 – 2021
- *NLP Engine*, HUFLIT 2022 – 2023

PRESENTATIONS

2024	**SHARE 2024**, Monash University: Vietnamese–English medical translation AI for healthcare communication (Buntine et al.).
2022 – 2023	Guest Speaker, Google Data Analytics (Module 8), NIC & Google.
2022	*Spoken corpus construction*, FIT-HUFLIT Seminar Series.
2020	Five co-authored talks, Linguistics and Application Areas Conference, VNU-HCMUSSH.

AWARDS & SCHOLARSHIPS

2025	Vingroup Science & Technology Scholarship for Overseas Doctoral Study
2025	UTS Collaborative Research Degree Scholarship
2023	VinUniversity Full PhD Scholarship
2021	VNU-HCM Postgraduate Scholarship
2020	ACP (Singapore) Certificate of Appreciation

CERTIFICATIONS & LANGUAGES

DataCamp	Introduction to Python; Intermediate Python; Introduction to NLP in Python; Data Visualization with Matplotlib (2022).
IBM	Python 101 for Data Science.
Languages	Vietnamese (native); English (professional working proficiency).

REFEREES

Prof. Wray Buntine
Department of Data Science & AI
Monash University
wray.buntine@monash.edu

Prof. Massimo Piccardi
Faculty of Engineering & IT
University of Technology Sydney
Massimo.Piccardi@uts.edu.au

Dr Thuy Bui
Project Director, GREEN-X
VinUniversity
thuy.btt@vinuni.edu.vn