Bach Phan-Tat

phantatbach.github.io github.com/phantatbach phantatbachATgmail.com

Data scientist with a background in finance and linguistics, currently a PhD candidate in computational linguistics. Skilled in quantitative analysis and machine learning, seeking opportunities to grow in AI and data science and apply data-driven methods to real-world problems.

WORK EXPERIENCE

Vbee (an AI company specialising in speech technology and conversational chatbot) Feb 2023 – Aug 2024

Junior Research Engineer *Hanoi, Vietnam*

Project: HUST Chatbot (a conversational AI for Hanoi University of Science and Technology)

- Built machine learning (Random Forest), language models (BERT-based) for classification tasks (Gender Classification of Name, Chat Domain Classification, Chat Toxicity Identification), achieved accuracies of 85%.
- Design statistical tests for data sampling and human evaluation.
- Developed a chatbot evalutation pipeline that utilises commercial LLMs.
- Crawled, extracted, generated (using LLM APIs), (pre-)processed, annotated and reviewed textual and audio data.
- Conducted literature review for optimal solutions.
- Collaborated in the development of linguistic data, prompts, and linguistic rules for various tasks and projects (e.g. Speech to Text, Text to Speech, Automatic Speech Recognition, parts of speech parsing).

what3words (an innovative location-mapping technology company) Jul 2023 – Aug 2023

Remote Research Consultant *London, England*

Project: Vietnamese Automatic Speech Recognition system (in collaboration with VinFast)

- Identified systematic biases of failed results through thorough phonetics and semantic analysis.
- Proposed effective solutions by leveraging knowledge of Linguistics and Machine Learning.

PUBLICATIONS

Book chapters (edited volumes)

- **Bach Phan-Tat**[*], Sofía Aguilar-Valdez[*], Stefania Degaetano-Ortlieb, Dirk Geeraerts, & Dirk Speelman. (2025). Discursive parallels of the chemical revolution: topic modelling and distributional analysis. In *Large Language Models for the History, Philosophy, and Sociology of Science: Reflections from a Field in Motion* (open access, transcript). (Book chapter / edited volume). [*]Equal contribution.

Conference & workshop proceedings (peer-reviewed)

- Khai Le-Duc, Tuyen Tran, **Bach Phan-Tat**, et al. (2025). MultiMed-ST: Large-scale Many-to-many Multilingual Medical Speech Translation. (EMNLP'25; Main Conference).
- Khai Le-Duc, Phuc Phan, Tan-Hanh Pham, **Bach Phan-Tat**, et al. (2025). MultiMed: Multilingual Medical Speech Recognition via Attention Encoder Decoder. (ACL 2025, Industry Track; Top 15% of accepted papers).

- Khai Le-Duc, Khai-Nguyen Nguyen, **Bach Phan-Tat**, et al. (2025). Sentiment Reasoning for Healthcare. (ACL 2025, Industry Track; Oral Presentation).
- Duc Cao-Dinh, Khai Le-Duc, Anh Dao, **Bach Phan-Tat**, et al. (2025). Audio-3DVG: Unified Audio - Point Cloud Fusion for 3D Visual Grounding. (Under review).

PRESENTATIONS

Conference/workshop presentations

- **Bach Phan-Tat**, Kris Heylen, & Stefano De Pascale. (2026). SynFlow: Continuous Semantics Change Analysis via Dependency Co-occurrences. Paper presented at the ICAME annual conference.
- **Bach Phan-Tat**, Dirk Geeraerts, & Dirk Speelman. (2026). Conceptual Change during the Chemical Revolution: Air, Acid and Water in the Royal Society Corpus. Paper presented at the ICAME annual conference.
- **Bach Phan-Tat**, Kris Heylen, Dirk Geeraerts, Stefano De Pascale, & Dirk Speelman. (2026). From Parsers to Prompts: Combining AI and Linguistics for Interpretable Language Evolution. Workshop presentation at AI in Language Evolution (Evolang 2026 workshop).
- **Bach Phan-Tat**. (2026). Slot-Filler Distributional Changes in Times of Scientific Debate: A case study of air. Talk at Quantitative Diachronic Linguistics and Cultural Analytics: Data-Driven Insights into Language and Cultural Change, King's College London, 15–16 Jan 2026.
- Khai Le-Duc, Khai-Nguyen Nguyen, **Bach Phan-Tat**, et al. (2024). Sentiment Reasoning for Healthcare. (AIM-FM Workshop @ NeurIPS 2024).

Invited talks & seminars

- **Bach Phan-Tat**. (2026). SynFlow: Continuous Semantics Change Analysis via Dependency Co-occurrences. Talk at Data in Historical Linguistics seminar series.

EDUCATION

KU Leuven **Sep 2024 – Now**
PhD, Computational Linguistics *Leuven, Belgium*
Thesis: Underlying dimensions in conceptual change
Achievements

- Developed SynFlow, a package for analysing semantic change using syntactic co-occurrences and Jensen–Shannon Divergence.
  - Implemented efficient pattern extraction and structuring using BFS/DFS, n-ary syntactic tree search, and filtering.
- Automated LLM linguistic annotation pipelines with local LLMs.
- Cleaned, parsed and lemmatised different text corpora.
- Optimised resource allocation and utilisation (parallel processing, batch processing, distributed processing) for concurrent computing jobs, reducing processing time by 80%.
- Maintained and updated Nephological Semantics, the previous project of QLVL research unit.
  - Modified the original pipeline to use different types of word embeddings as the backbone for further analysis.
- Set up different environments on SSH server.

Thesis supervisors: Prof. Dr. Dirk Speelman, Prof. Dr. Dirk Geeraerts, Dr. Kris Heylen

WorldQuant University **Oct 2025 – Now**
MSc, Financial Engineering *Remote*
Finished modules: Financial Market

University of Stirling **Sep 2021 – Sep 2022**
MSc, English Language and Linguistics *Stirling, Scotland*
Relevant modules: Linguistic Structures, Language and Cognition, Discourse Analysis, Historical Linguistics and the History of English.
Thesis: The structure of *thời gian*: A semantic analysis of the Vietnamese lexeme *thời gian* based on Principled Polysemy
Achievements

- Compiled a specialised Vietnamese news corpus using #LancsBox and Python.
- Identified 7 senses of the Vietnamese lexeme *thòi gian* (time) and their trajectories of development using the Principle Polysemy framework.

Thesis supervisor: Dr. Andrew Smith
Thesis grade: 80% (Distinction – class rank: 1st)
Average grade: 78% (Distinction – class rank: 1st)

Academy of Finance **Aug 2016 – Aug 2020**
BA, Corporate Finance *Hanoi, Vietnam*
Relevant modules: Advanced Math 1, 2 (Linear Algebra, Calculus); Fundamentals of Probability and Mathematical Statistics; Principles of Statistics
Thesis grade: 9.2/10
Average grade: 7.15/10

TECHNICAL SKILLS

Computer/Data Science:

- Exploratory Data Analysis, Data Cleaning and (Interactive) Data Visualisation
- Machine Learning / Deep Learning algorithms (Tabular, Text, Time-series, Visual)
- Programming languages (ordered by proficiency): Python (Numpy, Matplotlib, Pandas, SpaCy, Sklearn, Pytorch, Hugging Face, etc.), R, Bash, LaTeX, Matlab, Java
- GPU computing, Parallel Processing, Batch Processing
- Mathematics (Linear Algebra, Multivariate Calculus, Statistics)
- Databases: MongoDB, SQL
- Quantitative / Experimental Research Design / A/B Testing
- Git/Github
- Prompt Engineering

Linguistics:

- Corpus Tools (CQPWeb, Lancsbox, Sketch Engine, AntConc)
- Linguistic Analysis (Phonetics, Morphology, Syntax, Semantics)
- Data Annotation, Transcription

ACHIEVEMENTS AND AWARDS

- 2024: Marie Skłodowska-Curie Actions 3-year Doctoral Fellowship, Horizon Europe.
- 2022: Research Based Learning Prize for the best Master Dissertation in Literature and Languages (225 GBP), University of Stirling. The first student in the history of the Linguistics Department to be awarded the prize.

- 2021: Post Graduate Scholarship (4000 GBP), University of Stirling.

ADDITIONAL TRAININGS (with embedded credential URL)

Data Science

- Math for Machine Learning (AI4E - Content)
- Inferential Statistics (The University of Amsterdam - Credential ID:WD6E96392LT4)
- Applied Data Science Lab (WorldQuant University - Credential URL)
- Machine Learning (AI4E - Content)
- Deep Learning (Neuromatch Academy - Credential URL)
- Applied AI Lab (WorldQuant University - Credential URL)
- NeuroAI (Neuromatch Academy - Credential URL)

Linguistics

- Corpus Linguistics: Method, Analysis, Interpretation (Lancaster University - FutureLearn Credential ID: i3rvtp4)
- Corpus Linguistics Summer School (Birmingham University - Credential URL)

Neuroscience

- Medical Neuroscience (Duke University - Credential ID:E5AEQREWYNUV)
- Summer School: Computational Neuroscience (Neuromatch Academy - Credential URL)

Philosophy

- Philosophy and the Sciences: Introduction to the Philosophy of Cognitive Sciences (The University of Edinburgh - Credential ID:PYLP69Y3VEC8)
- Philosophy and the Sciences: Introduction to the Philosophy of Physical Sciences (The University of Edinburgh - Credential ID: L46X374LF4XD)

Research Methodology

- Quantitative Method (The University of Amsterdam - Credential ID: TJSJVDB5HZWF)
- Qualitative Research Methods (The University of Amsterdam - Credential ID: 8Z9G4364Q335)
- Research Writing in the Social Sciences (2022) (INASP - Credential URL)
- Open Science 101 (Neuromatch Academy x NASA - Credential URL)

LANGUAGES

- Vietnamese: Native
- English: IELTS 8.5 – C2
- Dutch: A2