

EDUCATION

University of Edinburgh*PhD in Designing Responsible NLP*

Edinburgh, UK

Sep 2025 – Aug 2029

- **Supervision:** Dr. Emily Allaway
- Fully funded by the **G-Research PhD Scholarship** for outstanding academic merit.

Master of Informatics [\[transcript\]](#) **(First Class Honours)***Sep 2019 – May 2024*

- **Dissertation supervision:** Prof. Mirella Lapata
- **Courses:** Natural Language Processing, Machine Learning, Reinforcement Learning, Calculus, Linear Algebra

Abbey Christian Brothers' Grammar School*Secondary school (AAA in A-Level Mathematics, Physics and Computer Science)*

Newry, UK

*Sep 2012 – May 2019*PUBLICATIONS

MatheMagic: A Dynamic Benchmark for Detecting Memorization in Mathematics

Dayyán O'Brien, Pinzhen Chen, Barry Haddow, Emily Allaway

*Under review for ACL Rolling Review (ARR). 2025.**MathNLP Workshop, Empirical Methods in Natural Language Processing (EMNLP) (non-archival). 2025.***[DocHPLT: A Massively Multilingual Document-Level Translation Dataset](#)**

Dayyán O'Brien, Bhavitvya Malik, Ona De Gibert Bonet, Pinzhen Chen, Barry Haddow, Jörg Tiedemann

*Proceedings of the Conference on Machine Translation (WMT). 2025.***[An Expanded Massive Multilingual Dataset for High-Performance Language Technologies](#)**Laurie Burchell, Ona de Gibert, *et al.* (incl. **Dayyán O'Brien**)*Proceedings of the Association for Computational Linguistics (ACL). 2025.***[Mind the Gap: Diverse NMT Models for Resource-Constrained Environments](#)**Ona De Gibert Bonet, **Dayyán O'Brien**, Dušan Variš, Jörg Tiedemann*Proceedings of the Nordic Conference on Computational Linguistics (NoDaLiDa). 2025.***[EMMA-500: Enhancing Massively Multilingual Adaptation of Large Language Models](#)**Shaoxiong Ji, Zihao Li, *et al.* (incl. **Dayyán O'Brien**)*arXiv:2409.17892. Under review for the Journal of Data-centric Machine Learning Research (DMLR). 2024.***[Prompting Numerical Commonsense Reasoning across Languages](#)****Dayyán O'Brien***Outstanding Honours Thesis, School of Informatics, University of Edinburgh. 2024.***[Numerical Commonsense Reasoning across Languages](#)****Dayyán O'Brien***Outstanding Honours Thesis, School of Informatics, University of Edinburgh. 2023.*RESEARCH EXPERIENCE

University of Edinburgh*Doctoral Researcher, supervised by Emily Allaway*

Edinburgh, UK

Sep 2025 – Present

- Improving the compositionality of language models.

*Research Assistant, supervised by Barry Haddow**Aug 2024 – Aug 2025*

- Co-developed the HPLT 2.0 bitexting pipeline to mine parallel data.

*Junior Research Assistant, supervised by Pinzhen Chen**May 2024 – Jul 2024*

- Cleaned and processed data for EMMA-500, a 7B parameter LLM for over 500 languages.

*Junior Research Assistant, supervised by Mirella Lapata**Jun 2022 – Apr 2024*

- Curated and evaluated mNumerSense: 36k+ Arabic, Chinese, Russian numeric commonsense sentences.

AWARDS & ACHIEVEMENTS

- G-Research PhD Scholarship (2025 – 2029)
- Outstanding Honours Project (2023 & 2024)
- Runner-up Best Coursework for Reasoning and Agents (2021)
- Exemplary Project for Foundations of Data Science (2021)

MAJOR CONTRIBUTIONS & OPEN SOURCE

- Leader, Document-Level HPLT Corpus:** Large-scale, document-level parallel corpus (huggingface.co/datasets/HPLT/DocHPLT).
- Contributor, EMMA-500:** An open-source LLM for 500+ languages (github.com/MaLA-LM/emma-500).
- Contributor, HPLT 2.0 Bitexting Pipeline:** Parallel data mining pipeline (hplt-project.org/datasets/v2.0).

TEACHING EXPERIENCE

- | | |
|------------------------------------------------------------------------------------------------|----------------------------|
| University of Edinburgh | Edinburgh, UK |
| <i>Demonstrator for Accelerated Natural Language Processing (INFR11125)</i> | <i>Sep 2023 – Nov 2023</i> |
| <i>Tutor for Foundations of Natural Language Processing (IINFR10078)</i> | <i>Jan 2023 – May 2023</i> |
| <i>Tutor for Foundations of Data Science (INFR08030)</i> | <i>Sep 2022 – May 2023</i> |
| <i>Demonstrator for Foundations of Natural Language Processing (INFR10078)</i> | <i>Jan 2024 – Mar 2024</i> |
| Self-employed | Edinburgh, UK |
| <i>Private tutor</i> | <i>Sep 2020 – Mar 2024</i> |

INDUSTRY EXPERIENCE

- | | |
|----------------------------|-----------------|
| Kainos | Belfast, UK |
| <i>Software Engineer</i> | <i>Mar 2018</i> |
| Bombardier | Belfast, UK |
| <i>Engineer</i> | <i>Sep 2017</i> |
| Computer Hospital | Newry, UK |
| <i>Computer technician</i> | <i>Oct 2017</i> |

LEADERSHIP & OUTREACH

- | | |
|------------------------------------------------------------------|----------------------------|
| Brand Ambassador | Edinburgh, UK |
| <i>G-Research</i> | <i>Sep 2025 – Present</i> |
| Committee Member, Community Events Planning | Edinburgh, UK |
| <i>Edinburgh Bahá'í Community</i> | <i>Mar 2025 – Present</i> |
| Primary School Micro:bit Programming Workshop Facilitator | Bathgate, UK |
| <i>University of Edinburgh</i> | <i>Jan 2024 – Feb 2024</i> |

SKILLS

- Programming:** Python, Java, Haskell, \LaTeX , SQL
- Libraries:** PyTorch, Tensorflow, HuggingFace, NLTK, NumPy, Slurm, Kubernetes, Docker, MTurk, pandas, sklearn, statsmodels, Matplotlib, Festival, HTK, Tkinter, Google Cloud VM, Weather & Maps API, Seaborn, Kivy, JUnit, Maven
- Languages:** English (Native), German (Professional)