

# Proyag Pal

Edinburgh, UK

✉ [proyag.pal@ed.ac.uk](mailto:proyag.pal@ed.ac.uk)

📁 [proyag.github.io](https://proyag.github.io)

in [www.linkedin.com/in/proyag-pal](https://www.linkedin.com/in/proyag-pal)

🐙 [github.com/Proyag](https://github.com/Proyag)

## Interests

Natural language processing (NLP), large language models, large-scale and high-quality text datasets, neural machine translation – especially multilingual and document-level

## Education

- 2020 – 2024 **Ph.D. in Informatics, University of Edinburgh (ILCC)**  
Edinburgh Ph.D. research in machine translation. Supervised by Kenneth Heafield and Alexandra Birch.
- 2016 – 2017 **M.Sc. in Informatics, University of Edinburgh, with Distinction**  
Edinburgh *Selected Courses:* Machine Translation, Accelerated Natural Language Processing
- 2011 – 2016 **B.Sc. & M.Sc. in Computer Science, St. Xavier's College**  
Kolkata *Selected Courses:* Artificial Intelligence, Data Mining & Warehousing, Computer Architecture

## Experience

### Professional Experience

- Aug 2024 – Present **Senior NLP Engineer, Aveni**  
Edinburgh Building FinLLM - LLMs for NLP applications in the financial services industry.
- Dec 2023 – Apr 2024 **Deep Learning Engineer, Efficient Translation Limited, part-time**  
Edinburgh Corpus extraction and efficient low-resource machine translation.
  - Trained efficient machine translation and corpus cleaning models for low-resource language pairs.
  - Ran and optimised an efficient scalable parallel corpus extraction pipeline on web-scale data.
  - Delivered datasets and models to customers on time and meeting requirements.
- Nov 2022 – Feb 2023 **Applied Scientist Intern, Amazon AWS AI, internship**  
Santa Clara Four-month internship working on improving isochronous machine translation for automatic dubbing. Co-organised the automatic dubbing track at IWSLT 2023.
- Jun 2020 – Oct 2020 **Data Engineer, TAUS**  
Amsterdam Worked on the EU-funded ParaCrawl project to collect parallel corpora from large-scale web crawls.
  - Optimised, maintained, and ran a highly scalable processing pipeline to extract, translate, align, clean, and release parallel corpora from web crawling data.
- Feb 2020 – Apr 2020 **Junior AI Researcher, Unbabel**  
Lisbon Machine translation and quality estimation for customer-facing products.
  - Built domain-specific production machine translation models and quality estimation models.
- Feb 2018 – Jan 2020 **Fellow in Neural Machine Translation, World Intellectual Property Organization (WIPO), Advanced Technology Applications Center**  
Geneva Development and maintenance of WIPO Translate and related NLP tools and technologies.
  - WIPO Translate*: Built, improved, evaluated and deployed domain-specific neural and statistical machine translation models using the Marian and Moses toolkits.
  - IPCCAT*: Developed neural text classification systems for patent categorisation.
  - Developed a system to retrieve semantically similar content from large collections of text using sentence embeddings and Faiss indexes.
  - Instrumental in the training and deployment of neural MT systems at several other international organisations and patent offices including IMF, OECD, WTO, IAEA, and KIPO.

## Academic Research Experience

- Nov 2020 – Dec 2024     **Ph.D. Student**, *University of Edinburgh (ILCC)*, School of Informatics  
Doctoral research in machine translation. Supervised by Kenneth Heafield and Alexandra Birch.
- Research on analysing and incorporating extra information required by neural machine translation models in addition to source text to produce accurate translations.
  - Introduced “cheat codes” – providing compressed target-side information to models – as a method to analyse additional information required by the models.
  - Created large-scale document-level translation corpora in several language pairs based on ParaCrawl and built and analysed context-aware translation models.
  - General research interests mainly in analysis of machine translation models, multilingual and document-level machine translation.
- Mar 2023 – May 2023     **Visiting Researcher**, *University of Zurich*, Department of Computational Linguistics  
Three-month visit, conducting research on detection and analysis of underspecification of the source sentence in machine translation. Supervised by Rico Sennrich.
- Sep 2017 – Dec 2017     **Research Assistant**, *University of Edinburgh (ILCC)*, School of Informatics  
Low-resource domain-specific machine translation research on the MeMaT project. Supervised by Kenneth Heafield and Alexandra Birch.
- Worked on developing isiXhosa-English medical-domain machine translation to facilitate doctor-patient communication in health centres in South Africa.
  - Collected corpora released as a public resource.

---

## Selected Publications

Full list of publications at <https://proyag.github.io/publications>

- ACL 2024     **Document-Level Machine Translation with Large-Scale Public Parallel Corpora**, *Proyag Pal, Alexandra Birch, and Kenneth Heafield*
- Interspeech 2023     **Improving Isochronous Machine Translation with Target Factors and Auxiliary Counters**, *Proyag Pal, Brian Thompson, Yogesh Virkar, Prashant Mathur, Alexandra Chronopoulou, and Marcello Federico*
- EACL 2023 (Findings)     **Cheating to Identify Hard Problems for Neural Machine Translation**, *Proyag Pal and Kenneth Heafield*
- NAACL 2022     **Cheat Codes to Quantify Missing Source Information in Neural Machine Translation**, *Proyag Pal and Kenneth Heafield*

---

## Master's Projects

- Jun 2017 – Aug 2017     **Reward Augmented Maximum Likelihood to Improve Neural Machine Translation Training**, *University of Edinburgh*, supervised by Kenneth Heafield
- Used reinforcement learning-inspired task rewards to augment the training objective.
  - Improved upon a strong baseline by 1.07 BLEU.
  - Re-implemented and integrated into the then Theano-based Nematus framework.
- Aug 2015 – May 2016     **Permutation Flow Shop Scheduling using Natural Algorithms**, *St. Xavier's College, Kolkata*, supervised by Siladitya Mukherjee
- Optimization of makespan in permutation flow shop scheduling, using genetic algorithms.

---

## Programming

**Python**, with PyTorch, NumPy, sklearn, etc.

**C++**, Marian toolkit for MT

**Bash, Docker, L<sup>A</sup>T<sub>E</sub>X**