Sonal Sannigrahi

+351 915309433 | sannigrahisonal@gmail.com| sonalsannigrahi.github.io | **GitHub**: sonalsannigrahi

EDUCATION

Instituto Superior Técnico, Lisbon, Portugal **November 2023-present**
Ph.D candidate on "Multimodal Natural Language Processing" advised by Dr. André Martins

University of Saarland, Saarbrücken, Germany **October 2021-September 2023**
MSc. in Language Science and Technology | **GPA: 1.6/5** (1.0 highest, lower is better)

- **Thesis:** Robustness in Lexical Sharing of Multilingual Language Models for Indian Languages

École Polytechnique, Paris, France **August 2018-June 2021**
BSc. in Mathematics and Computer Science | **GPA: 3.89/4.0**

- **Thesis:** Low Resource Machine Translation with Cross-Lingual Mapping
- **Relevant Coursework:** Design and Analysis of Algorithms, Machine Learning, Computer Architecture, Deep Learning in Computer Vision (Graduate Course)

TECHNICAL EXPERIENCE

IST, Lisbon, Portugal- **Graduate Research Assistant** **11/23-Present**

- Adapting text LLMs to handle multimodal inputs in the form of speech and vision

Amazon, Cambridge, UK- **Applied Scientist in Alexa Conversational Speech** **10/24-02/25**

- Semantically-meaningful speech representations for better downstream alignment with text-based models at Amazon Alexa Conversational Speech.
- Designed a novel modular codec architecture to perform voice conversion and waveform reconstruction while maintaining speaker privacy and low-latency.
- Patent resulting from this work

Apple, Aachen, Germany- **Research Intern in AI/ML** **05/23- 10/23**

- Synthetic Data Generation using Large Language Models (LLM) for Siri Speech Recognition (ASR).

DFKI , Saarbrücken, Germany- **Graduate Research Assistant** **10/21-05/23**

- Machine Translation group under Dr. Cristina España-Bonet to research multilingual document embeddings
- Designed comprehensive experiments to verify performance in the wild performance of document-level embeddings across language models and statistical methods
- Applied my findings in automated document alignment and multilingual document classification

INRIA, Paris, France- **NLP Research Intern** **06/21-12/22**

- Joined ALMAnaCH team under Dr. Rachel Bawden to research interpretability of large multilingual language models and to improve current techniques in low resource translation.
- Trained and evaluated several Transformer-based translation models to run holistic analysis

RaspberryPi Translation , London, UK- **Volunteer Translator** **06/2019-Present**

- Translated over 100 coding tutorials from English to Hindi on the RaspeberryPi Projects page in collaboration with CoderDojo and Code Club
- Led the organisation of 3 hackathons to speed-translate across more than 12 languages engaging 60+ volunteers
- Awarded Volunteer of the Month multiple times for my work

PUBLICATIONS

TowerVision: Understanding and Improving Multilinguality in Vision-Language Models
A. Viveiros*, P. Fernandes*, S. Santos, **S. Sannigrahi,** E. Zaranis, N. M.G. Guerreiro, A. Farajian, P. Colombo, G. Neubig, A. FT Matins
In Submission

Aligning Small Scale Speech-Text Language Models for Speech-Text Learning
G. Attanasio, **S. Sannigrahi,** B. Peters, A.FT Martins
In Proceedings of the 22nd International Conference on Spoken Language Translation 2025 (IWSLT co-located at ACL)

Movie Facts and Fibs (MF2): A Benchmark for Long Movie Understanding
M. Zaranis, A. Farinhas, …, **S. Sannigrahi**, …, A. FT Martins
In Submission

From TOWER to SPIRE: Adding the Speech Modality to a Text-Only LLM
K. Ambilduke*, B. Peters*, **S. Sannigrahi*,** A. Keshwani, T.K Lam, B. Martins, M. Z. Boito, A.FT Martins
In Proceedings of the 30th Conference of Empirical Methods for Natural Language Processing 2025 (EMNLP)

Semantically-meaningful Speech Representations
S. Sannigrahi, B.T Vecino, I. Vallés-Pérez, C. Papayiannis
Patent (Accepted) 2025

Synthetic Query Generation using Large Language Models for Virtual Assistants
S. Sannigrahi*, T. Fraga Silva, Y. Oualil, C. Van Gysel*
In Proceedings of the 47th International ACM SIGIR Conference on Information Retrieval 2024

Are the Best Multilingual Document Embeddings simply Based on Sentence Embeddings?
S. Sannigrahi, J. van Genabith, C. España-Bonet
In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics 2023 (EACL)

Investigating Lexical Sharing in Multilingual Machine Translation of Indian Languages
S. Sannigrahi, R. Bawden
In Proceedings of the 24th Conference of the European Association for Machine Translation 2023 (EAMT)

Isomorphic Cross-Lingual Word Embeddings for Low-Resource Languages
S. Sannigrahi, J. Read
In Proceedings of the 7th Workshop on Representation Learning for NLP at the 60th Annual Meeting of the Association for Computational Linguistics 2022 (ACL)

* equal contribution

SELECT AWARDS

Saarland Stipend for Academic Excellence 2023
ACM SIGHPC Computational and Data Science Fellowship 2022
Palantir Women in Technology Europe Scholar 2022
Bloomberg Women in Technology Insights Scholar 2022
Google Generation Scholarship for Excellence in Computer Science 2020
Honours Grant and Excellence Scholarship at École Polytechnique

SKILLS AND ADDITIONAL INFORMATION

Programming: Python, Bash/Shell, C++, PyTorch
Teaching: Deep Structured Learning (PhD course) Spring 2025
Languages: English, Hindi, Bengali (native), French (B2/C1)