Massimo Daul

M.Sc. Student in Data Science and Artificial Intelligence
Trieste, Italy | mmd9604@nyu.edu | https://massimodaul.github.io/

Summary

I'm a master's student at the University of Trieste working on speech and language technology for low-resource languages. Most of my research has focused on ASR, forced alignment, phonological representation, and tools for building spoken language corpora.

Education

University of Trieste — Trieste, Italy
M.Sc. in Data Science and Artificial Intelligence — *September 2026 – Present*

New York University — New York City, USA
B.A. in Mathematics and Computer Science — *September 2022 – May 2026*

United World College of the Adriatic — Duino, Italy
International Baccalaureate — *September 2021 – May 2022*

Research Experience

Computational Research Intern — June 2024 – September 2026
Endangered Languages in Contact in Istria and Kvarner — *New York City, USA*

- Headed project for fine-tuned custom Wav2Vec2-Bert models for ASR and Forced-Alignment of low-resource languages
- Developed machine learning approaches for part-of-speech tagging, speaker diarization, and segmentation to be used in low-resource corpus creation

Computational Intern – Addiction Research — June 2019 – December 2023
The Jackson Laboratory — *Bar Harbor, ME, USA*

- Developed neural-network-based approaches for gait detection and pose analysis in mice for behavioral and addiction research.

Computational Intern – Small Business Development — June 2023 – August 2023
Profit Decoder — *New York City, USA*

- Developed data-analysis tools for a startup working with economic and financial data for small businesses.

Publications & Presentations

Linguistically Informed Tokenization Improves ASR for Underresourced Languages
Field Matters, 2026 — First Author

NLP Tools for Constructing a Spoken Corpus of Endangered Language Varieties: A Case Study of the ELIC Corpus
Presented at AACL, 2026 — First Author

Comparing Phonological Feature Sets for Low-Resource ASR
Proceedings of the 2025 Annual Meeting on Phonology — Second Author

Forced Alignment of Istro-Venetian Speech via Neural versus Non-Neural Models
The Journal of the Acoustical Society of America, 159, A161, 2026 — Second Author

Technical Skills

Programming: Python, C++, Bash
Machine Learning: PyTorch, Wav2Vec2-BERT, HuBERT, scikit-learn
Speech & Language: ASR, forced alignment, diarization, phonological feature modeling, TextGrid/Praat, ELAN
Data & Tools: Pandas, Git, SLURM/HPC

Languages

English, Italian: Fluent **Spanish, Portuguese**: Conversational