

# Caterina Bonan

PhD in Linguistics | Junior NLP Engineer

 OrcID  CaterinaBi  LinkedIn  caterina.bonan@outlook.com  07983818663

## SELECTED EXPERIENCE

### AICORE | MACHINE LEARNING AND DATA ENGINEERING TRAINEE

May 2022 – Present | London, UK

Training in AI & Data using industry-standard tools. Working on industry projects throughout.

- Skills: Web scraping, SQL; Data science (Data cleaning & visualization), Machine learning (Regression, random forest, SVMs, PCA, t-SNE), Deep learning (Pytorch, neural networks, CNNs, RNNs, Autoencoders, GANs), NLP (LSTMs, transformers, attention mechanism, BERT, HuggingFace), Training in the cloud (AWS EC2), Deployment (Python Flask, Cron, API creation, AWS Lambda, Docker, ONNX, TensorFlow.JS, GCP AI services).

### UNIVERSITY OF CAMBRIDGE | POSTDOCTORAL RESEARCHER IN THEORETICAL LINGUISTICS

September 2019 – Current | Cambridge, UK

Worked on 30+ varieties throughout. Published 8 peer-reviewed research articles, 2 edited volumes and 1 research monograph. Awarded a SNSF mobility scholarship twice.

- Tools: Excel, Python, Matplotlib, Selenium, LaTeX, VS code, Word and Power Point.
- Links to scientific writing samples: **Cadernos**, **IOS Press**, **Glossa**, **Isogloss**.

### UNIVERSITÉ DE GENÈVE | RESEARCH ASSISTANT IN THEORETICAL LINGUISTICS

September 2015 – August 2019 | Geneva, CH

Fully-funded doctoral research as part of a SNSF lab. Worked on standard and non-standard French, and 50+ northern Italian varieties. Published 3 research articles, 1 co-edited volume, and a PhD dissertation.

- Tools: Excel, Python, Praat, LaTeX, Word and Power Point.

## EXPERIENCE THROUGH PROJECTS

### POLYGLOT | PYTHON, SELENIUM, NLTK, PYTORCH.

Repo: [eleutherAI-polyglot](#)

- Creation of multilingual models that show better non-English performance than mBERT, BLOOM, and XGLM.
- Currently working towards the establishment of a large Romance dataset (with a 3T target).

### FB MARKETPLACE RECOMMENDATION RANKING SYSTEM | PYTHON, PYTORCH, AWS, EKS.

Repo: [aicore-recommendation-ranking-system](#)

- Currently training a model that works with images and text to make recommendations to potential buyers based on their demographic information;
- Will host the app on EKS to orchestrate the containers that maintain/retrain the model.

### NLP CLASSIFIER TO STOP ONLINE VIOLENCE AGAINST CHILDREN | PYTHON, SELENIUM.

Repo: [nlp-classifier-omdena-marseille](#)

- Data annotation, data scraping, language model creation;
- Deployment as an API, execution of the API as a stand-alone Chrome extension.

### WEB SCRAPING PIPELINE | PYTHON, SELENIUM, DOCKER, GITHUB ACTIONS.

Repo: [aicore-web-scraping-pipeline](#)

- Developed a module that scrapes data to upsert information to a database;
- Pushed an image of the containerised application in a CI/CD pipeline that performs unit testing and deploys the image to an EC2 instance.

### IMAGE CLASSIFIER FOR FLOWER RECOGNITION | PYTHON, TENSORFLOW.

Repo: [udacity-image-classifier](#)

- Python image classification application that runs from the command line.
- The application trains a deep learning model on the 102 Category Flower Dataset, and uses it to classify images of flowers.

### COMPUTER VISION APPLICATION | PYTHON, TEACHABLE MACHINE, OPENCV

Repo: [aicore-computer-vision](#)

- Trained a computer vision model and used Tensorflow to detect whether Rock, Paper or Scissors is shown to the camera in real-time and with a high accuracy
- Used the OpenCV library to access the webcam and play Rock Paper Scissors with the computer using the image from the camera

## SKILLS

### PROGRAMMING

Intermediate:

Python (>2y) • LaTeX (>5y)

Familiar:

Java (>1y) • HTML (<1y)

### LIBRARIES/Frameworks

OpenCV • TensorFlow • Selenium  
PyTorch • pandas • Matplotlib  
NLTK • SpeechRecognition •  
PyDub • spaCy • polyglot

### TOOLS/PLATFORMS

Git • GitHub • Docker • SQL

### LINGUISTIC RESEARCH

Syntax, Documentation, Data  
modelling, IPA, Praat.

## EDUCATION

### PHD IN LINGUISTICS

UNIVERSITÉ DE GENÈVE

December 2015 - May 2019

### MA IN SCIENCE OF LANGUAGE

UNIVERSITÀ CA' FOSCARI VENEZIA

October 2011 - June 2013

### BA IN SCIENCE OF LANGUAGE

UNIVERSITÀ CA' FOSCARI

September 2008 - October 2011

## CERTIFICATES

AiCore

September 2022 - November 2022

Software Engineer Certificate.

Udacity

March 2022 - October 2022

Programming in Python for AI.

DATA CAMP

October 2022 - Present

NLP in Python Skill Track.

AI CORE

May 2022 - September 2022

Python Programmer Certificate.

UNIVERSITÉ DE GENÈVE

December 2015 - December 2017

Specialisation in Linguistics.

## CURIOSITIES

I'm a retired pro volleyball player, and since 2015 I have attracted research funds for more than 400.000CHF (about £375.000).