# PhD in Data Science: Modeling Digital Language Diversity
# TU Wien, Research Unit Data Science

## Project Context

Languages are not just a means of communication – they are central to identity, culture, and cognition. Yet, in our increasingly digitized world, many of the world's ∼7,000 languages remain almost invisible online. This project investigates how digitization, through infrastructure, platforms, and language technologies, affects the representation and survival of languages in the digital domain.

The PhD position is based at TU Wien and is part of a larger interdisciplinary research project that combines data science, natural language processing (NLP), and sociolinguistics to study the global dynamics of linguistic diversity in digital environments (see: https://digiling.univie.ac.at/digilingdiv/).

This position offers the opportunity to develop and apply computational methods for analyzing large-scale multilingual data, modeling digital language distributions, and investigating the technological and socio-economic factors that shape linguistic diversity online.

## Your Focus

You will contribute to developing scalable methods and models to analyze linguistic diversity in digital spaces. Your work will include:

- Extracting and analyzing multilingual data from web archives (Common Crawl) and social media (e.g., Twitter/X)
- Applying and fine-tuning language identification models across hundreds of languages
- Measuring the digital representation of languages across countries and time periods
- Modeling how digitization factors (infrastructure, language technology, access) relate to linguistic visibility online
- Contributing to reproducible pipelines and interactive visualizations (e.g., dashboard tools)

## Your Profile

We are looking for a candidate with the following qualifications:

- A completed Master's degree in Computer Science, Data Science, Computational Linguistics, or a closely related field
- Strong skills in Python (preferred), with experience in data processing, NLP, and machine learning
- Familiarity with large-scale text data, language modeling, or language identification tasks
- Experience with data analysis and statistical modeling
- Experience with web data (e.g., crawling, Common Crawl, APIs) is a plus
- Interest in digital inequality, language technology, or cultural analytics
- Excellent communication and collaboration skills in English

## We Offer

- A 30-hour/week, 30-month funded PhD position at TU Wien in a highly interdisciplinary and innovative project (with the possibility of contract extension)
- Joint supervision from TU Wien (Computer Science) and University of Vienna (Linguistics)
- Participation in the Vienna Doctoral College on Digital Humanism
- A highly collaborative and flexible working environment in Europe's most livable city

## How to Apply (by May 16, 2025)

Send a single PDF to: `digilingdiv-application[ät]ds-ifs.tuwien.ac.at`, including:

- Motivation letter (1–2 pages)
- CV
- Academic transcripts
- List of publications (including your Master's thesis)
- (Optional) Link to code samples or GitHub, past and/or current projects

  Interviews will be held in May 2025.

## Start Date

Summer 2025.

We look forward to your application!