

# MICHAŁ ADAMCZYK

## SOFTWARE DEVELOPER • RESEARCHER

code.madamczyk@gmail.com • <https://github.com/mdm-code>

I am a software developer with experience in natural language processing and natural language understanding. This is mostly backend and data management/processing. I do research in both contemporary and historical linguistics.

## EXPERIENCE

### Software Developer / Researcher

February 2021 - Present

CLARIN-PL / University of Łódź (Remote / Łódź, Poland)

- Developed text segment parallelization REST API
- Set up a production-ready Solr instance for intent recognition and classification
- Designed and integrated database schemas for spoken language acquisition
- Coauthored tasks for PolEval 2021 and 2022 (<http://poleval.pl/>)
- Coauthored a report evaluating the performance of major ASR engines for Polish
- Added automatic text punctuation with a transformer-based sequential classifier
- Designed and implemented intent classification taxonomy
- Coauthored DiaBiz, an annotated corpus of Polish call center dialogs
- Enhanced existing language models and classifiers with domain-specific data
- Processed large volumes of transcribed speech with forced-alignment
- Systematized data across databases for author authentication experiments
- Coded intent classifier API with FAISS and semantic embeddings

### Software Developer / Researcher

November 2019 - February 2021

VoiceLab (Remote, Norway / Gdańsk, Poland)

- Created English language model with 10% increase in an extrinsic evaluation
- Developed web scrapers to gather medical language material at scale
- Scripted numerous audio processing pipelines in POSIX shell
- Written numerous tools and commands for natural language processing
- Developed vastly faster grapheme-to-phoneme converter for Polish
- Worked with most language modelling toolkits
- Utilized semantic embeddings for intent recognition and clustering
- Embedded forced-alignment and speaker diarization in ASR pipelines

### PhD Fellow

January 2018 - December 2020

University of Stavanger (Stavanger, Norway)

- Delivered papers on international conferences and symposia
- Worked on annotation guidelines and administered transcription process
- Developed data for low-resource, historical languages
- Assisted fellow researchers to utilize VCS in the team's daily work
- Developed extensible data processing pipeline for historical language material
- Written up the whole monograph in LaTeX

## **Program Administrator**

July 2016 - December 2017

TAPFIN/ManpowerGroup Solutions (Łódź, Poland)

- Handled invoicing and business reports for the client and workforce suppliers
- Participated in job interviews for various coding positions
- Supported the client with workforce intelligence

## **Copywriter**

June 2016 - August 2016

Softbox Group (Łódź, Poland, Remote)

- Localized marketing materials for a major hotel brand
- Helped to promote brand with professional Polish to English translations
- Adapted English advertizing materials to the Polish market

## **English Language Instructor**

October 2015 - August 2016

Empik Language School (Łódź, Poland)

- Facilitated English language learning among students
- Developed language learning aids and tutorials
- Motivated students to practice English pronunciation in class

# **PROJECTS**

- Opinionated snippet manager for Unix (<https://github.com/mdm-code/gsnip>)
- XDG base directory specification in Go (<https://github.com/mdm-code/xdg>)
- T5-based historical lemmatizer (<https://github.com/mdm-code/manx>)
- Zettelkasten as a single POSIX shell script (<https://github.com/mdm-code/zet>)
- Polish grapheme-to-phoneme converter (<https://github.com/mdm-code/prg2p>)

# **EDUCATION**

- PhD, Literacy Studies (Linguistics), University of Stavanger, 2018-ongoing
- MA, English (Linguistics), University of Łódź, 2012-2014
- BA, English (Linguistics), University of Łódź, 2009-2012

# **SKILLS**

- UNIX/Linux Systems Architecture and Use
- Software Design and Development
- Python
- Go
- Terminal User Interface
- Test Driven Development Practices
- Domain Driven Design Practices
- Interprocess Communication Methods
- Transport Layer Protocols API Development and Design
- API Protocols
- Docker Containerization
- Scripting and Automation
- Speech and Natural Language Processing
- Language Modeling
- Data Collection and Management

- Solr
- Data Engineering
- Corpus Design
- Data-centric Approach
- Grammars and Parsing
- Scientific Method
- Writing and Research Publication

## KEYWORDS

UNIX • Linux • BSD • Ubuntu • Docker • Containers • Docker-compose • Command Line Interface (CLI) • Terminal User Interface (TUI) • Secure Shell (SSH) • Networking • TCP/IP • Unix Domain Sockets (UDS) • Sockets • HTTP • REST • RPC • JSON-RPC • SQL • PostgreSQL • POSTGRES • Solr • TMUX • Screen • Vim • Vi • Neovim • Git • GitHub • GitLab • Python • Go (golang) • Bash • Zsh • Shell • POSIX • FastAPI • Flask • Pandas • Jupyter • Natural Language Processing (NLP) • Natural Language Understanding (NLU) • UML • YAML • JSON • Markdown • XML • LaTeX • Pandoc • TDD • DDD

## LANGUAGES

- Polish (native)
- British English (native-like spoken and written)