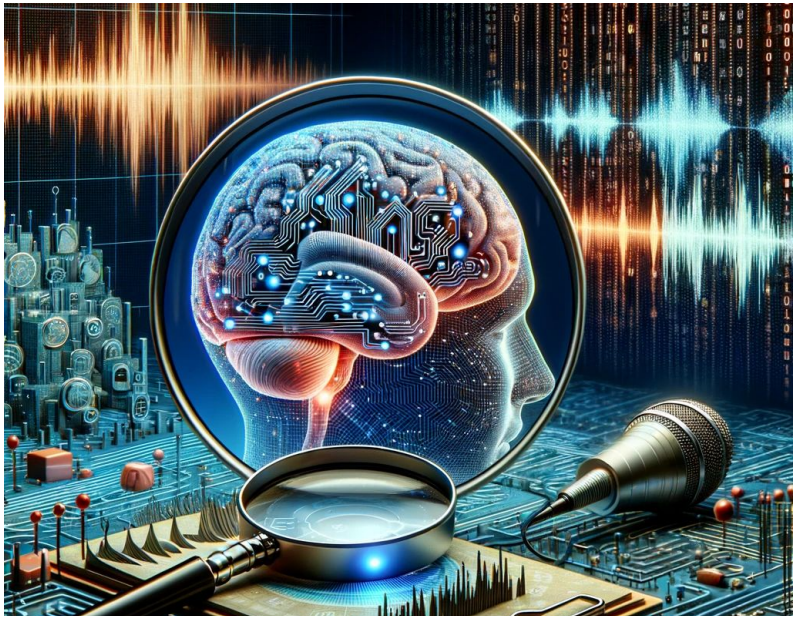
BA/MA projects at the Department of CL



Elisa Pellegrino and Jeannette Roth

December 8, 2025

8.45-10.00

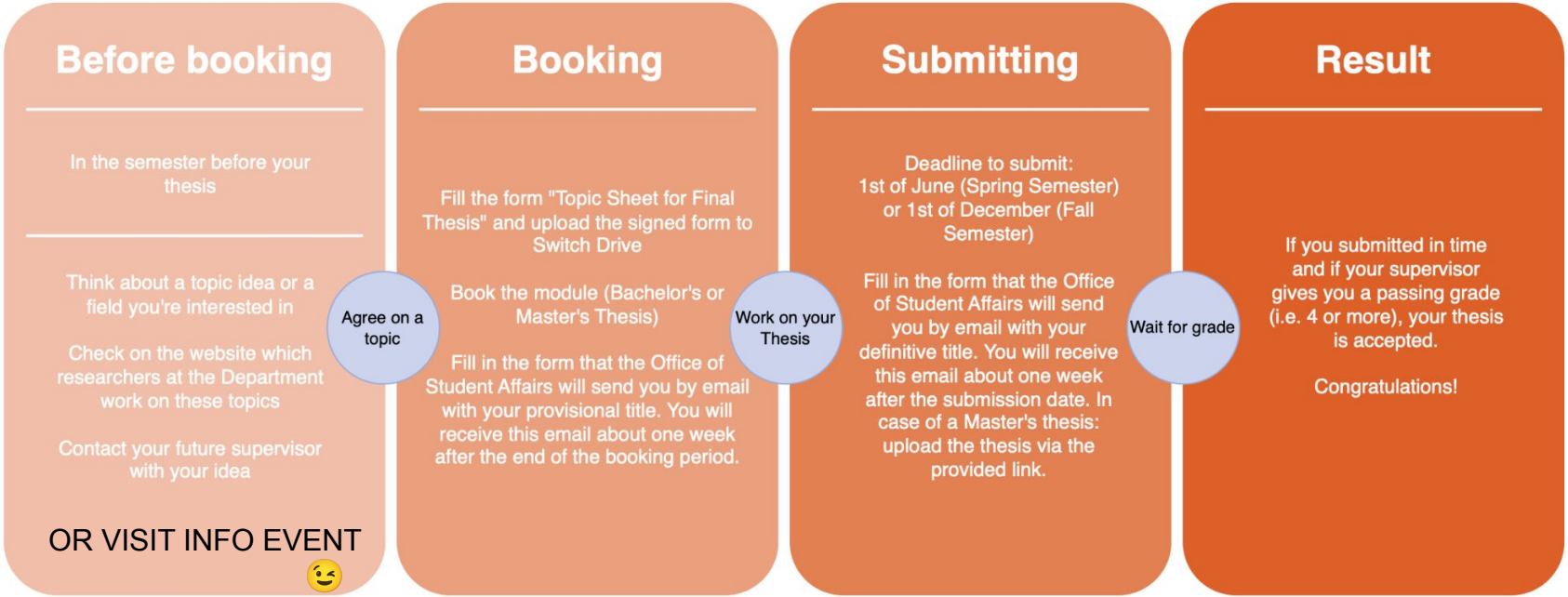
BA Thesis (Study regulations §25 - §28)

- **Credits:** 15 ECTS, compulsory, graded
- **Duration:** 1 semester
- **Submission Deadlines:** June 1 (spring semester) / December 1 (fall semester)
- **Thesis**: Individual, no co-authorship
- **Supervisor:** Master's degree or higher. Commitment for supervision is the prerequisite for booking the module Bachelor's thesis.
- **Booking:** Via Student Portal in the standard booking period

MA Thesis (Study regulations §30 - §33)

- **Credits:** 30 ECTS, compulsory, graded
- **Duration**: 2 semesters
- **Submission Deadlines:** June 1 (spring semester) / December 1 (fall semester)
- **Thesis:** Individual, no co-authorship
- **Supervisor:** PhD or higher. Commitment for supervision is the prerequisite for booking the module Master's thesis.
- **Booking:** Via Student Portal in the standard booking period

Process and Organisation



You can find more information via these links:

Faculty of Arts and Social Sciences:

https://www.phil.uzh.ch/en/studies/studyessentials/graduation.html

https://www.phil.uzh.ch/dam/jcr:092773b8-9a44-44a4-a666-c81c6c8f8aa1/STO_Allgemeiner_Teil_EN.pdf (study regulations)

Computational Linguistics:

https://www.cl.uzh.ch/en/studies/studies-BA-MA/teaching/bachelor-thesis.html

https://www.cl.uzh.ch/en/studies/studies-BA-MA/teaching/master-thesis.html

https://www.cl.uzh.ch/en/studies/studies-BA-MA/teaching/BaMasterArbeit.html

University of Zurich[UZH]

Department of Computational Linguistics

Gerold Schneider
gerold.schneider@uzh.ch

Yingqiang Gao
yingqiang.gao@uzh.ch

Tilia Ellendorff
tilia.ellenforff@uzh.ch

Three BA / MA opportunities in the ENGAGE project

https://www.engage-dialogue.ch/en/

The Forest Entomology Group of the Swiss Federal Research Institute (WSL) together with the Linguistic Research Infrastructure (LiRI) NLP Group of UZH are offering the opportunity for three exciting Master thesis project to identify influential scientific literature and influential scientific actors in renewable energy and biodiversity research. You can be part of our research team.

Engage, a Joint Initiative of the ETH domain, *evidence-based dialogue on trade-offs in wicked societal problems* is a transdisciplinary research project to further the discussion about trade-offs in complex situations where no single best solution exists. In the era of climate change, the transition to renewable energy use is paramount, while at the same time, the loss of biodiversity threatens the collapse of ecosystems.

1) Identification of influential scientific literature and scientific stakeholders based on network analysis

We conduct a systematic literature review about trade-offs and synergies between renewable energies and biodiversity.

Goal: Based on this literature corpus, a network model shall be advanced to identify influential scientific literature and relevant scientific actors, groups, and topics in this domain. You will program a pipeline that visulaizes authors, institutions, topics, and their interactions, and ideally allows visual exploration.

Required: good programming skills in R or Python, knowledge in NLP and/or language models, and a strong interest in renewable energy production and biodiversity stewardship. Experience with network models is an advantage.

2) Analysis of Swiss media texts about renewable energies and biodiversity, in German, French and/or Italian

We conduct a systematic literature review about trade-offs and synergies between renewable energies and biodiversity. However, scientific evidence can vary from the way this topic is discussed in the media and from public opinion

Goal: Based on our literature corpus and SwissDOX, your goal is to analyze the way in which renewable energy generation, biodiversity stewardship, and their role for society are conveyed in Swiss media texts.

Required: good programming skills in R or Python, knowledge in NLP and/or language models, and a strong interest in renewable energy production and biodiversity stewardship. Very good German, French and/or Italian is needed

3) Advancement and evaluation of RAG system in the context of renewable energies and biodiversity

We conduct a systematic literature review and have a large corpus of articles about trade-offs and synergies between renewable energies and biodiversity. To enable a detailed search and answer very specific questions, we develop a RAG system.

Goal: Based on our literature corpus and existing RAG technology, you help us to advance and evaluate RAG model performance in generating answers about renewable energies or biodiversity, based on scientific evidence.

Required: good programming skills in R or Python, knowledge in NLP and/or language models, and a strong interest in renewable energy production and biodiversity stewardship. Experience with RAG models is an advantage.

BA / MA opportunities in the Sinergia project

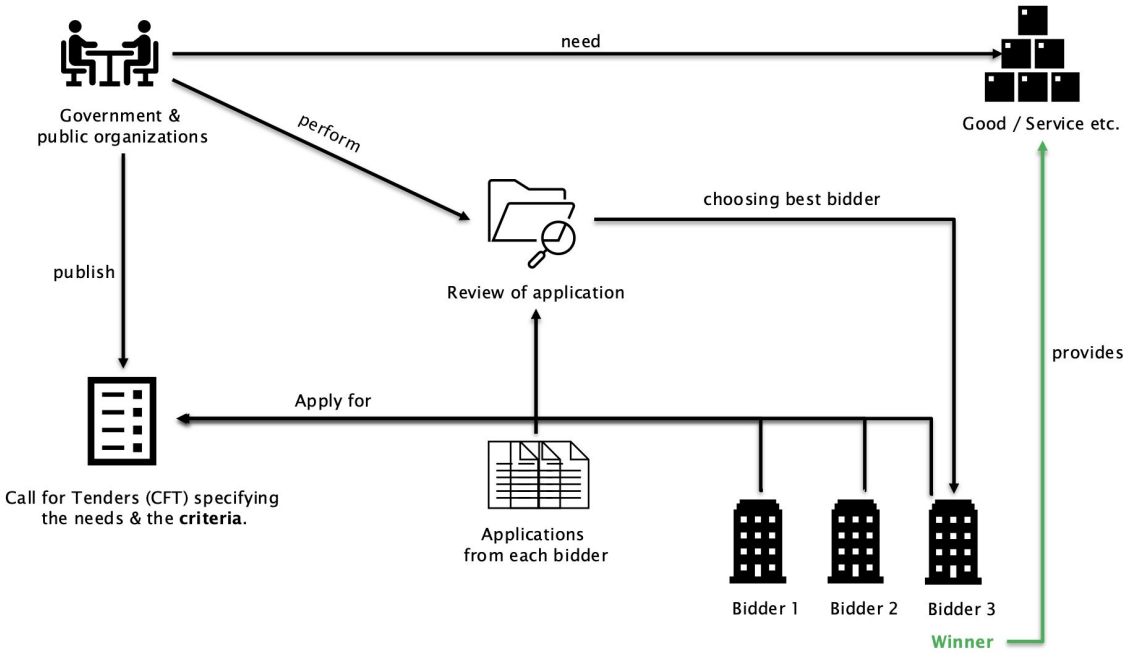
https://data.snf.ch/grants/grant/10000100

Bern University of Applied Sciences (BFH),

University of Bern,

Linguistic Research Infrastructure (LiRI) NLP Group

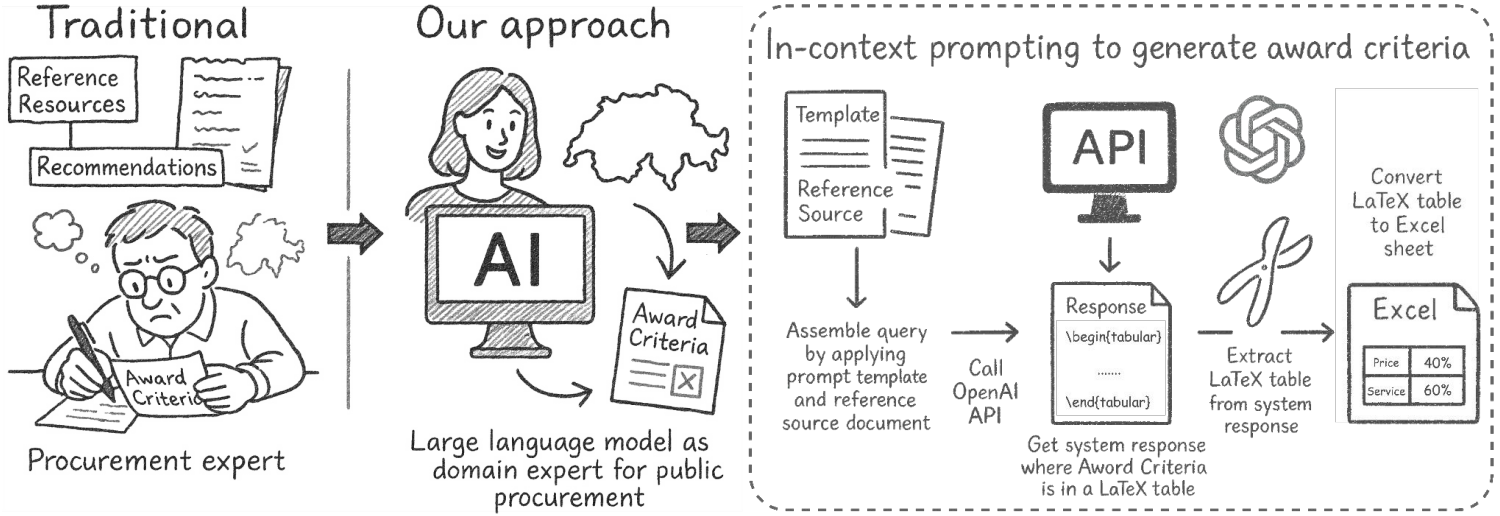
are offering the opportunity for one Master thesis project to develop a multi-agent system used in the public procurement domain.



University of Zurich[UZH]
Department of Computational Linguistics
Berner Fachhochschule
UNIVERSITÄT BERN
IntelliProcure

Multi-Agent System for Generating and Evaluating Award Criteria for Swiss Public Procurement

Award criteria for public procurement are difficult to obtain because 1) lack of human expertise across different public sectors; 2) no gold evaluation standard to compare against.



University of Zurich[UZH]
Department of Computational Linguistics
Berner Fachhochschule
UNIVERSITÄT BERN
IntelliProcure

Multi-Agent System for Generating and Evaluating Award Criteria for Swiss Public Procurement

Goal: Develop a multi-agent system using small language models (<10B) for 1) generating synthetic award criteria for different public procurement sectors; 2) grounding the criteria generation to trustworthy web-sources; 3) developing new methods (human and automatic approaches) for quality assessment; 4) software development for interactive workflow

Required: good programming skills in Python, solid knowledge in NLP and/or large language models, and strong engineering ability in multi-agent system design (ideally frontend and backend). Experience with RAG is a plus.

Contact: yingqiang.gao@uzh.ch , tilia.ellendorff@uzh.ch

University of Zurich[UZH]
Department of Computational Linguistics
Berner Fachhochschule
UNIVERSITÄT BERN
IntelliProcure

VatPub: Investigating Chronological Semantic Shift in Religious Documents using Large Language Models

This project examines how religious concepts evolve in meaning over time and investigates whether LLMs can reliably detect these chronological semantic shifts.

Goal: Tasks include designing prompt-based and embedding-based analyses, multilingual machine translation, evaluating model performance against linguistic baselines, and visualizing diachronic semantic trends in an interpretable manner. The expected outcome is a systematic assessment of how well modern language models capture long-term conceptual change in religious discourse, accompanied by an analytical report and a prototype tool for exploring semantic evolution.

Required: good programming skills in Python, solid knowledge in NLP and/or large language models, experience in machine translation is a plus.

Contact: yingqiang.gao@uzh.ch

University of Zurich[UZH]

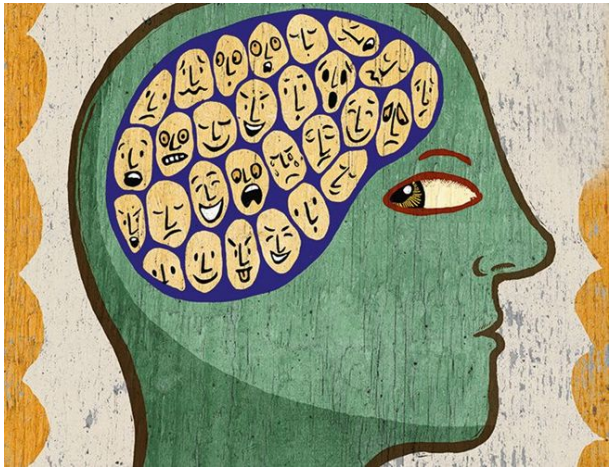
Department of Computational Linguistics

Pavo Orepic
pavo.orepic@uzh.ch

University of Zurich[UZH]

Department of Computational Linguistics

pavo.orepic@uzh.ch



How does my voice sound like?

Hearing voices in schizophrenia?

University of Zurich[UZH]

Department of Computational Linguistics

Rico Sennrich

rico.sennrich@uzh.ch

Reproduction Studies in Machine Translation and NLP

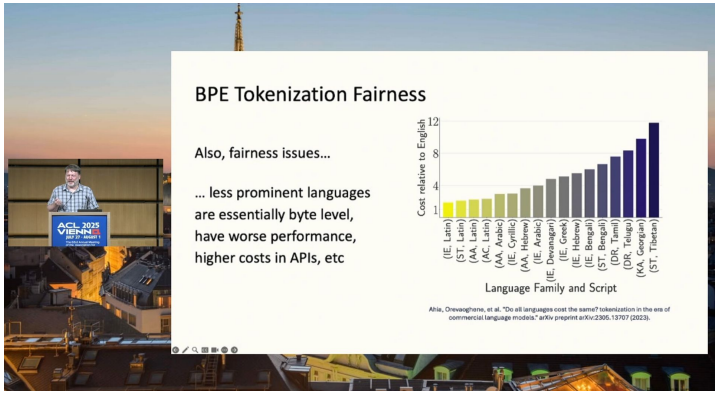
goal:

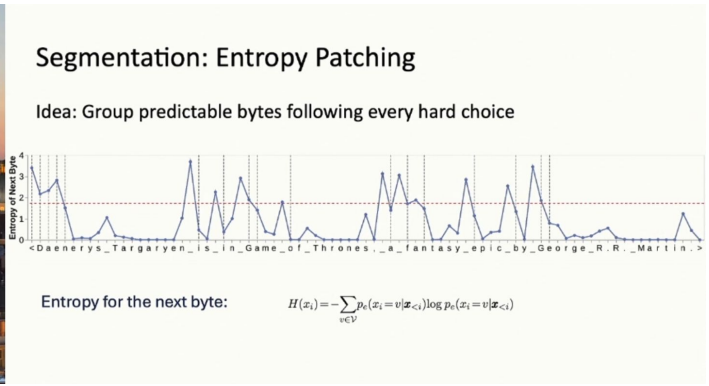
- pick recent publication and attempt reproduction
- gain novel insights about method: understand limitations, generalization to new settings, perform error analysis, …

requirements:

- machine learning and Python skills
- have taken "Machine Translation" or "Advanced Techniques in Machine Translation" for MT project

Multilingual Latent Tokenization





YES
traditional tokenizers have fairness issues

BUT
byte latent transformers have been prominently pushed by META, but only tested on English, and entropy as decision criterion is likely affected by language and resourcedness

University of Zurich[UZH]

Department of Computational Linguistics

Elisa Pellegrino
elisa.pellegrino@uzh.ch

The Role of Expressive Audio-Visual Information on Face-Voice and Voice-Face Identity Matching

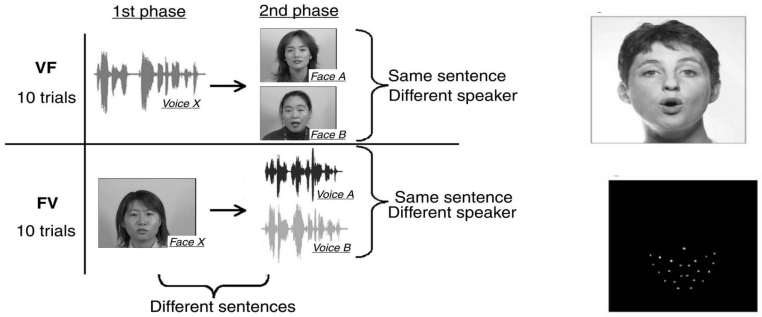
Elisa Pellegrino (elisa.pellegrino@uzh.ch)

Previous MA projects (submitted in Dec 2024 and under review)

Corpus

Less expressive (ADS) — More expressive (IDS)

Perceptual Testing



Current available projects

(BA Thesis)
Testing the effect of language familiarity on cross-modal identity matching

(MA Theses)
Acoustic analysis/perceptual evaluations/classification of the visual and auditory expressiveness of IDS and

Skills/Software

Acoustic analysis

Perception tests

Prolific

Data analysis

R

The Role of Speech Acoustics to Detect Deep Fake Voices

Elisa Pellegrino, Timothy Tianze Xu, Volker Dellwo
in collaboration with aurigin.ai

(elisa.pellegrino@uzh.ch, tianze.xu@uzh.ch)

Previous project

VOICE CONVERSION DATASET

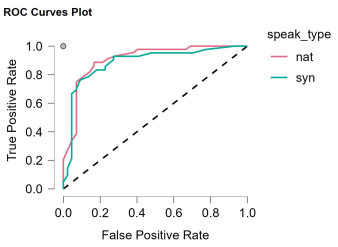
SPEECH MATERIAL

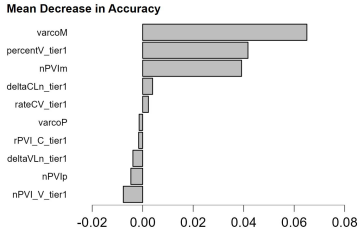
- **685 natural utterances**
  - 137 utterances * 5 speakers (1 source, 4 targets)
    - **44** statements (SVO; 5 words): **LONG UTTERANCE**
    - **10** y/n questions (based on 5-word statement)
    - **83** statements (SV; 2 words): **SHORT UTTERANCE**
- **548 voice converted utterances**
  - 137 converted utterances by source speaker * 4 target speakers

SPEAKERS

- **Target speakers**
  - 4 male speakers of Stand. German
  - Master/Phd students
  - 22-34 y. o.
  - Previously screened for no audible regional accent
- **Source speaker**
  - Professional speaker
  - Recruited at ZHDK

Preliminary results based on prosodic features





Current available projects

Project 1 (TTX): How do listeners recognize deepfake versus real voices? Testing whether familiarization with artificial or authentic voices affects recognition

Project 2 (EP): Investigate the acoustic features that distinguish human voice from AI-generated voices (adaptable for BA theses)

Project 3 (EP): Compare the capabilities of humans and machines to rely on these features to detect natural vs AI-generated content.

Dataset

– Real speech includes a combination of **publicly available datasets** for real speech and **proprietary recordings**

- AI: existing, proprietary **data creation pipeline**, which creates voice clones of the real speech samples from various leading voice-cloning models

Skills/Software

Prolific

HUMAN MACHINE INTERACTION

Elisa Pellegrino with **David Gruenert (ZHAW, PhD Candidate at UZH)**

(elisa.pellegrino@uzh.ch)

Background

- The rise of multimodal large language models can transform the way we interact with technology
  - From task-oriented interactions to more natural and fluid conversations with users
    - The ontological boundary between AI-powered voice assistants and humans as conversational partners is challenged

Projects

To understand

- speech and language adaptations and accommodation toward both humans and conversational AI,
- factors shaping these emerging forms of digital communications (digital competence, age group, dialects, multi vs mono-modality)

Skills/Software

Acoustic analysis

Perception tests

Data analysis

Acoustic Signatures of Prosody–Emotion Integration in Spoken AI

Elisa Pellegrino with **David Gruenert (ZHAW, PhD Candidate at UZH)**

(elisa.pellegrino@uzh.ch)

ZHAW project

Background

Large Language Models (LLMs) like ChatGPT or Claude are optimized for text, but miss crucial paralinguistic and extralinguistic information conveyed through speech segmentals and suprasegmentals

- Speaker specific information
- Mental and physical states

Goal

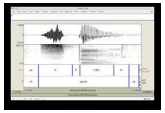
- Design data and interaction model that integrates emotional and prosodic information into the communicative process
- Implement a prototype system (e.g., using OpenAI, Whisper, SpeechBrain, etc.) in which the use of this supplementary information can be selectively enabled or disabled.

Our (UZH) contribution

- Examine the changes in speech acoustics in the performance of the system when these information are included
  - Acoustic Markers Prosody–Emotion Integration in Spoken AI
- Understand how human interlocutors perceive the resulting differences in naturalness, empathy, and intelligibility.

Skills/Software

Acoustic analysis



Perception tests

Prolific

Data analysis

The production, Perception and Classification of L2 accented english

Elisa Pellegrino with **Rabeah Kiani**
(elisa.pellegrino@uzh.ch)

Previous project

Corpus Collection

600 sentences in Pakistani-accented English (20 sentences *30 speakers)
600 sentences in L1 English (30 sentences *20 speakers)

Preliminary results based on prosodic features

Evaluation Metrics

	ENG L1	ENG L2	Average / Total
Accuracy	0.825	0.825	**0.825**
Precision (Positive Predictive Value)	0.822	0.829	0.825
Recall (True Positive Rate)	0.848	0.800	0.825
F1 Score	0.835	0.814	**0.825**
Area Under Curve (AUC)	0.877	0.875	**0.876**

Note. All metrics are calculated for every class against all other classes.

Rate differences and the timing of consonants

Available projects

BA Thesis
Investigate the segmental differences between Pakistani accented English and L1 English (pronunciation difference) and their weight for accuracy of classification

MA Theses
Identify the segmental features of Pakistani accented English (pronunciation difference) and test their relevance for L2 speech comprehension by humans and machines

Skills/Software

Acoustic analysis

Perception tests

Prolific

Data analysis

University of Zurich[UZH]

Department of Computational Linguistics

Anna Bondar
anna.bondar@uzh.ch

Explainability of Cognitively-Enhanced NLP Models

Look into the inner working mechanisms of the cognitively-enhanced LLMs/reward models for the LLMs, investigate why they achieve a better performance than the ones not enhanced by gaze.

References:

Fine-Tuning Pre-Trained Language Models with Gaze Supervision

Pre-Trained Language Models Augmented with Synthetic Scanpaths for Natural Language Understanding

Seeing Eye to AI: Human Alignment via Gaze-Based Response Rewards for Large Language Models

From Human Reading to NLM Understanding: Evaluating the Role of Eye-Tracking Data in Encoder-Based Models

University of Zurich[UZH]

Department of Computational Linguistics

Jan Brasser
brasser@cl.uzh.ch

Influence of Dialectal Spelling Variation on Language Processing (Jan Brasser)

Motivation

- Swiss German has no standardized spelling → strong individual variation
- Such variation may influence reading ease and comprehension

Project

- Design & run an **eye-tracking experiment** on how readers process dialectal spellings
- Focus on **non-standardized written language** in Swiss German (or similar)

Potential Research Goals/Ideas

- Investigate effects of specific spelling variants on **eye-movement measures**
- Use eye-tracking data to **identify spelling choices that make dialect texts easier to read**

Linguistic Data Annotation in Dialectal Texts Using LLMs (Jan Brasser)

Motivation

- Dialects like Swiss German lack a **standard orthography**, making automatic annotation difficult
- Traditional NLP tools struggle with **non-standard spelling**
- Modern LLMs show strong improvements in understanding dialectal text

Project

- Explore how **LLMs can be used to automatically annotate dialectal texts** (e.g., PoS tags, morphological information)

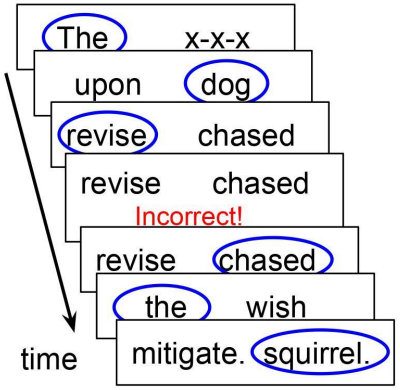
Potential Research Goals/Ideas

- Evaluate LLM performance on **linguistic annotation tasks** for dialects
- Develop strategies to **improve reliability and consistency** of LLM-based annotations

University of Zurich[UZH]

Department of Computational Linguistics

Cui Ding & Junlin Li
cui.ding@uzh.ch



1. Multilingual A-Maze

Background & Problem:

- Maze task: select between a target word and a distractor to continue sentence reading
- A-Maze: automatically generates distractors with LMs, initially only for English
- When it comes to other languages, we need to find a language-specific LM that does not use a subword tokenizer.
- Works ridiculously poorly for Chinese and other languages, and is very slow (4 mins per sentence).

Our idea:

- Develop a new generation of Multilingual A-Maze using **AI agents that is more flexible and supports** many languages.
- We have written a demo for Chinese. The results are much better and faster in the text we tested.

Your task:

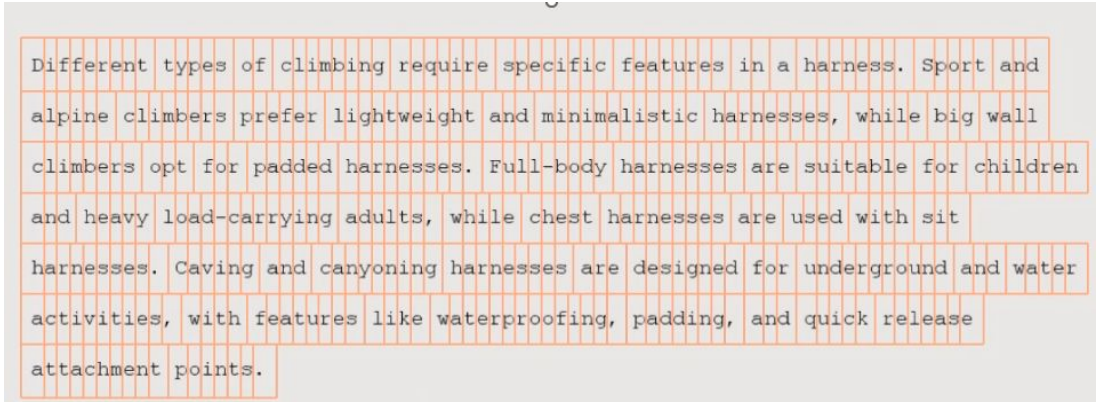
- Follow our demo to further develop it and test it across multiple languages.
- Replicate two classic studies (controlled experiment and naturalistic reading) to validate the pipeline.

University of Zurich[UZH]

Department of Computational Linguistics

David R. Reich
david.reich@uni-potsdam.de

Eye movement generation: Eye movements in reading as multi-modal modeling challenge

$f_\theta($  $)$

Different types of climbing require specific features in a harness. Sport and alpine climbers prefer lightweight and minimalistic harnesses, while big wall climbers opt for padded harnesses. Full-body harnesses are suitable for children and heavy load-carrying adults, while chest harnesses are used with sit harnesses. Caving and canyoning harnesses are designed for underground and water activities, with features like waterproofing, padding, and quick release attachment points.

Several topics on BSc and MSc level

Reqs:

- (Advanced) Machine Learning
- Python

Explainability of eye movement generation: giving back to fundamental research

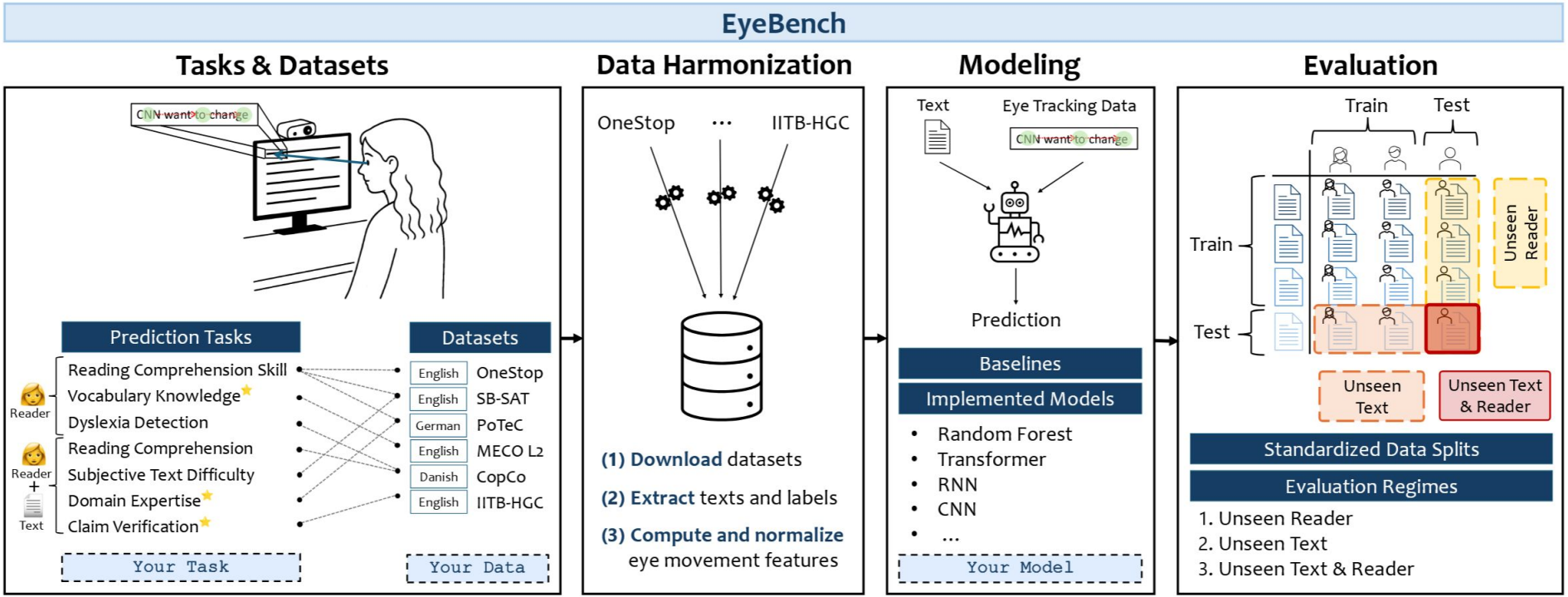
$f_\theta($ Different types of climbing require specific features in a harness. Sport and alpine climbers prefer lightweight and minimalistic harnesses, while big wall climbers opt for padded harnesses. Full-body harnesses are suitable for children and heavy load-carrying adults, while chest harnesses are used with sit harnesses. Caving and canyoning harnesses are designed for underground and water activities, with features like waterproofing, padding, and quick release attachment points. $)$

Topics on MSc level
Reqs:

- (Advanced) Machine Learning
- Python
- Background in reading research
- Pref:
- Familiarity with XAI toolkits

Inferences from eye movements

Several topics on BSc and MSc level
Reqs:
- (Advanced) Machine Learning
- Python



B.Sc.-level

M.Sc.-level

University of Zurich[UZH]

Department of Computational Linguistics

Marius Huber, Juri Opitz & Michelle Wastl
{marius.huber, jurialexander.opitz, michelle.wastl}@uzh.ch

Metric Space Magnitude for analysis of (L)LMs

Background & Problem:

Metric space magnitude is a geometric tool that, similarly to entropy, allows one to measure diversity of point clouds. In NLP, it has been used, e.g., to “fingerprint” LLMs through analysis of their latent spaces.

Idea/task:

The purpose of this project is to develop the necessary understanding of magnitude as a tool, and, subsequently, to investigate the value of magnitude for NLP by applying it to problems such as, e.g.,

- translation (direction) detection;
- uniformity/alignment tradeoff in contrastive learning;
- explainability of (L)LMs, such as, e.g., effects of training and/or fine tuning on magnitude of latent spaces;
- …

Prerequisites:

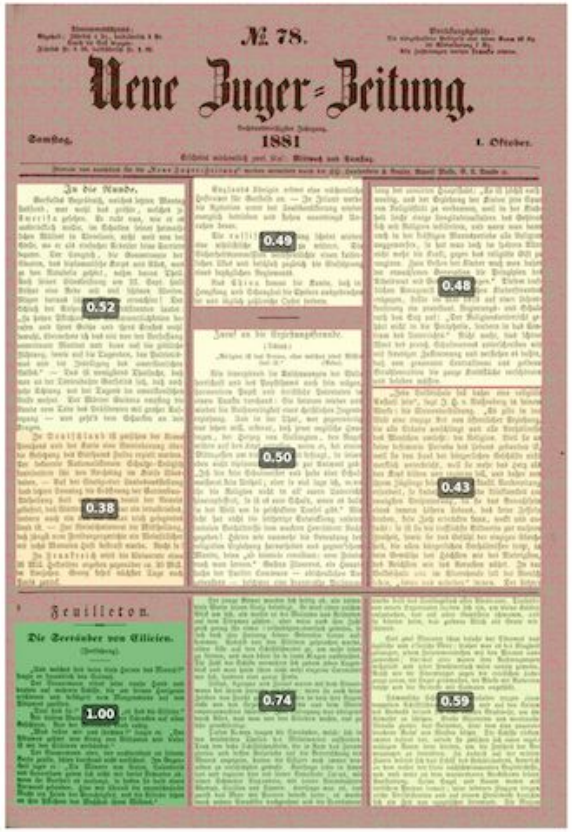
Python, NLP, “Mathematical Foundations of Computational Linguistics 2” course or equivalent

University of Zurich[UZH]

Department of Computational Linguistics

Andrianos Michail, Juri Opitz
& Simon Clematite
{andrianos.michail, jurialexander.opitz,
simon.clematide}@uzh.ch

Impresso II UZH CL Student Projects



Andrianos Michail

Juri Opitz

Simon Clematide

impresso Media Monitoring of the Past

Compare the passage on the left side with passage **#2** of **3** in the same cluster

SORT BY DATE (DESC)

« ‹ 1 2 3 › »

Passage in: NaN
L'Impartial Fri, Jan 12, 2018 — p.1

Tous les membres de la Grappe Groupement d'entente communale à Milvignes ont appris avec tristesse le décès de Madame Huguette LAURENT maman de M. Frédéric Laurent, membre fondateur et ancien conseiller communal à Milvignes Nous lui adressons, ainsi qu'à sa famille, l'expression de notre profonde sympathie.

Passage in: NaN
L'Express Sat, Sep 2, 2017 — p.35

Tous les membres de la Grappe Groupement d'entente communale à Milvignes ont appris avec tristesse le décès de Monsieur Eric LAURENT papa de M. Frédéric Laurent, membre fondateur et ancien conseiller communal à Milvignes Nous lui adressons, ainsi qu'à sa famille, l'expression de notre profonde sympathie. 028-802038

Visual Alignment for OCR Robust Text Embeddings

Research problem:

- **Semantic search performance** in historical document collections is significantly **reduced by OCR errors** introduced during the digitization process.

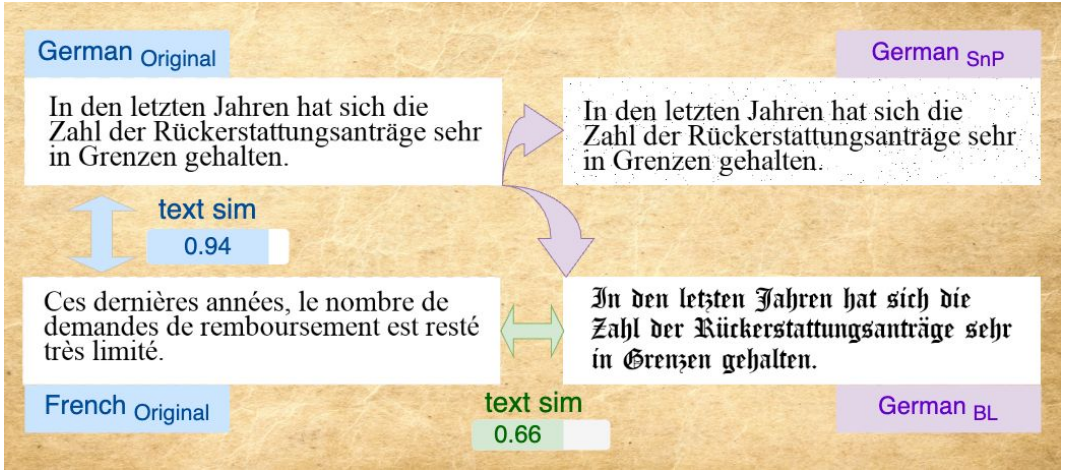
RQ: 1. Can **post-alignment** to **visual encoders** (e.g., **DeepSeekOCR**) effectively mitigate OCR-induced performance degradation in historical text embeddings?
2. What is the optimal strategy to **integrate** this **visual alignment** approach with existing **denoising training** methods that we have previously developed?

Expected outcome:

- **Visually aligned** text embedding models
- Quantitatively evaluated
- Qualitatively examined

Suitable for **a BA/MA** (IFI/CL) Thesis

Contact: **Andrianos Michail**



Evaluating Embedding Models In The Historic Domain

Research problem:

Semantic search performance in historical document collections might be different than in contemporary texts.

Embedding Models are important NLP tools

- They give us a “Similarity” for two texts (→backbone of document retrieval)

The “accuracy” of these models is evaluated on large scale benchmarks

- How trustworthy are results of such large-scale evaluations for the historic domain?

To test this:

- Investigate in-domain evaluations of embedding models on historic newspaper texts
  - E.g., Matching newspaper titles against newspaper texts.
- Will the best models from the benchmark also be the best for historic in-domain tasks?

Results will have implications on the reliability of benchmarks and the recommended use of embedding models in real world projects (e.g Impresso)

Contact: Juri Opitz and Andrianos Michail

Other Topics in Impresso II

We are also interested in setting up student projects in other topics that have to do with NLP/text mining of historic newspapers.

Feel free to contact us with an interest of yours in the NLP topics of:

- Evaluation and Adaptation of Embedding Models
- Keyphrase Extraction / Abstraction
- Topic Modelling
- Post-OCR Correction of historical texts
- [Your idea re processing and mining historic texts]

Contact (send to all three): Andrianos Michail, Juri Opitz, Simon Clematide

University of Zurich[UZH]

Department of Computational Linguistics

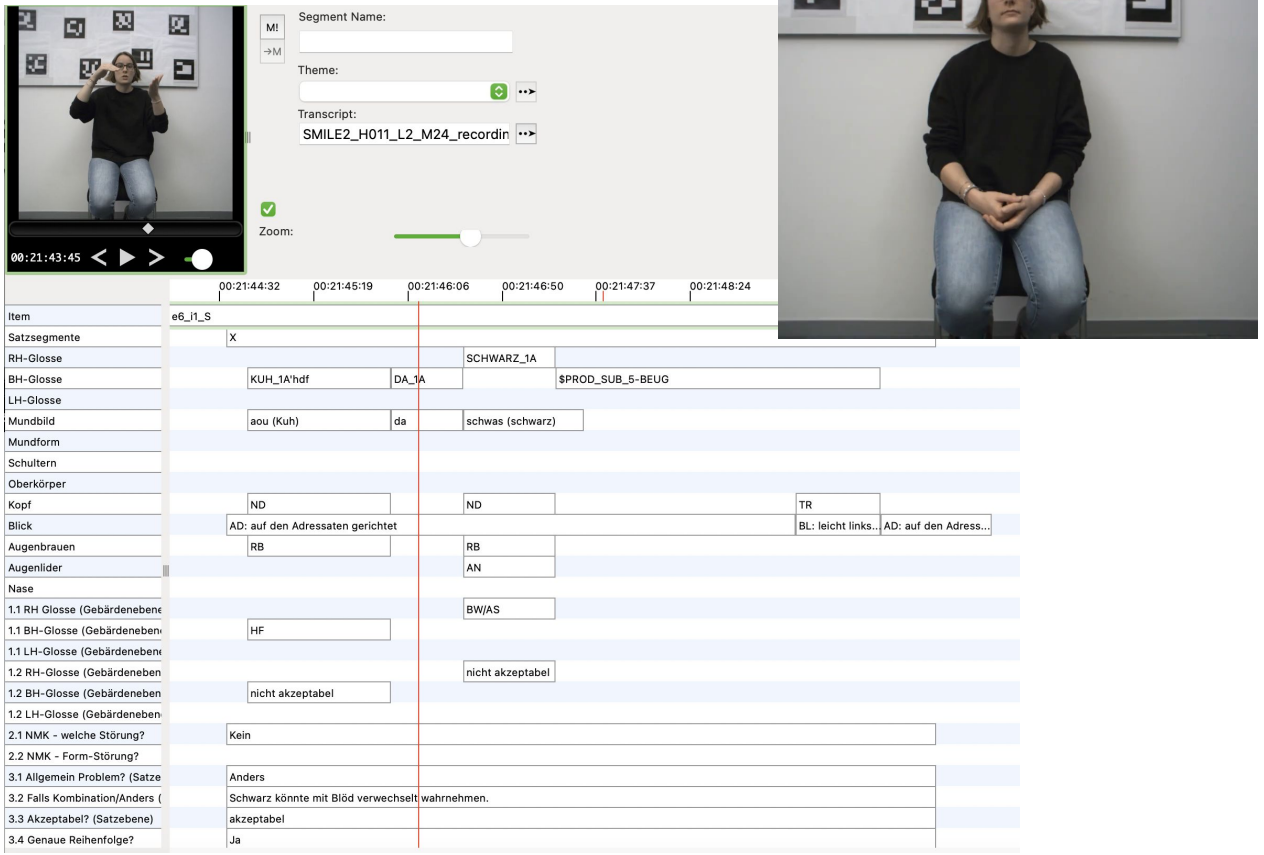
Sarah Ebling
ebling@cl.uzh.ch

Automatic sign language recognition on SMILE2 data

Supervisor: Sarah Ebling

Summary: Using human-labelled data from a recently concluded SNSF project (SMILE2), train an automatic sign language recognition model (both using existing models and developing new ones).

Requirements: Deep learning, Python, familiarity with or interest in computer vision techniques (in addition to NLP techniques), ideally successful completion of the course "Artificial Intelligence for Language Accessibility", "Visual Language Processing" or "Digital Accessibility".



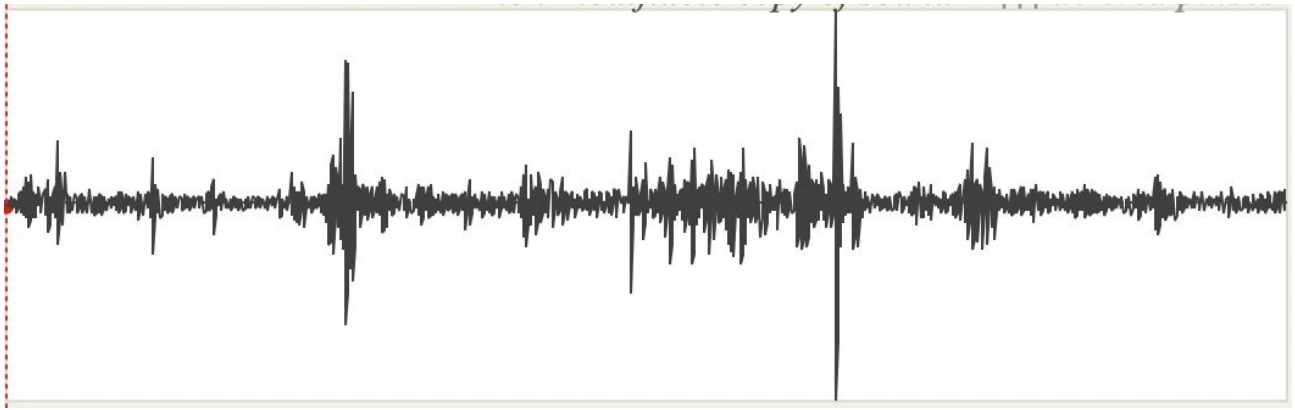
Multimodal automatic sign language recognition

Supervisor: Sarah Ebling

Summary: Conduct automatic sign language processing experiments using visual and acoustic cues (sounds of mouth actions). Link to master's thesis and ongoing PhD project by Longqian Ming

Ming, L. (2024). Investigating Breathing Patterns in Sign Language Storytelling: An Exploratory Study [Master's thesis]. University of Zurich.

Requirements: Deep learning, Python, familiarity with or interest in computer vision techniques (in addition to NLP techniques), ideally successful completion of the course "Artificial Intelligence for Language Accessibility", "Visual Language Processing" or "Digital Accessibility"



Crosslingual sign language analysis

Supervisor: Sarah Ebling

Summary: Using a multiparallel sign language corpus (SwissSLI), conduct insights into form similarities across the three sign languages involved.

Jiang, Z., Göhring, A., Moryossef, A., Sennrich, R., & Ebling, S. (2024). SwissSLi: The Multi-parallel Sign Language Corpus for Switzerland. In N. Calzolari et al. (eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (S. 15448–15456).

Carpentieri, S. (2025). Aligning a multi-parallel sign language corpus for six languages [Master's thesis]. University of Zurich.

Requirements: Deep learning, Python, familiarity with or interest in computer vision techniques (in addition to NLP techniques), ideally successful completion of the course "Artificial Intelligence for Language Accessibility", "Visual Language Processing" or "Digital Accessibility"