University of Zurich[UZH]

Department of Computational Linguistics

Bachelor and Master projects at the CL Department

Elisa Pellegrino and Jeannette Roth

May 7, 2026

12.00-14.00

BIN-0-K.02

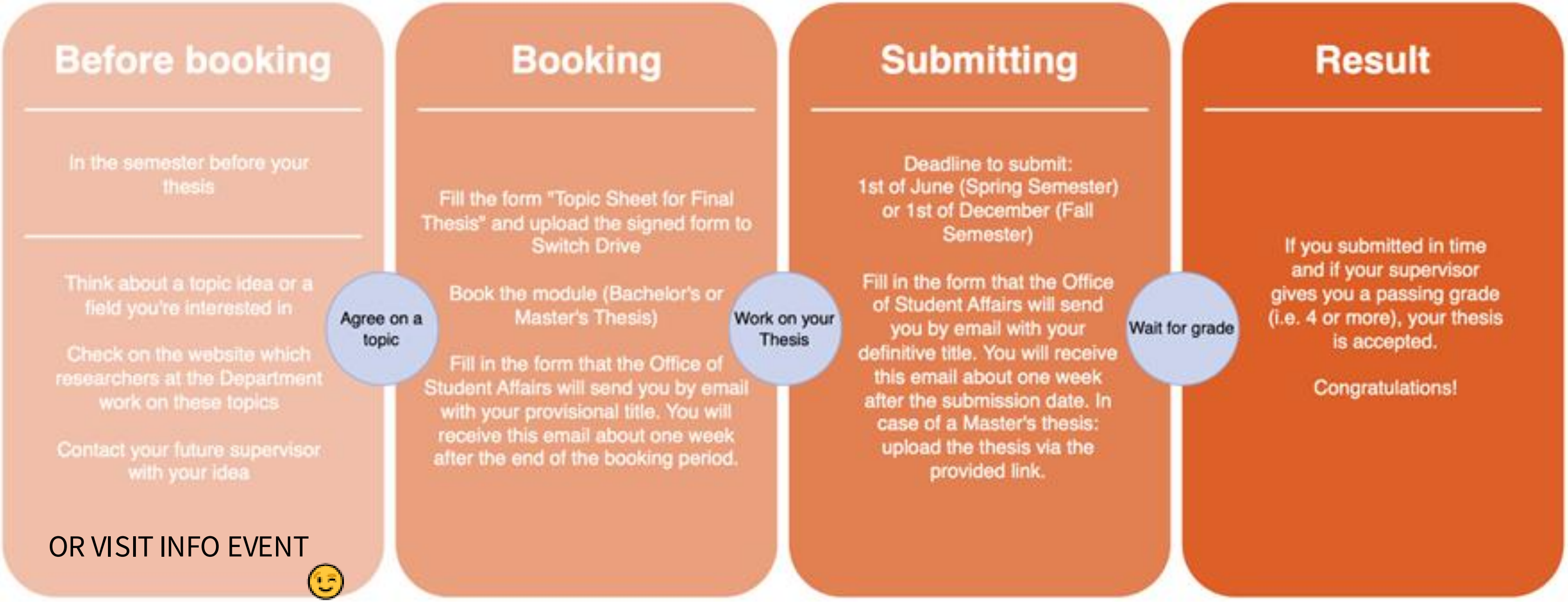
BA Thesis
(Study regulations §25 - §28)

- **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



OR VISIT INFO EVENT 😉

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

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)

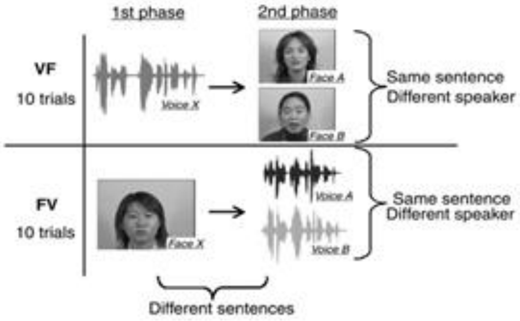
Previous MA projects (under review)

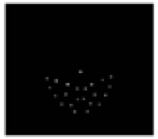
Corpus

Less expressive (ADS)

More expressive (IDS)

Perceptual Testing





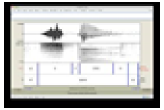
Current available projects

(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 ADS

Skills/Software

Acoustic analysis



Perception tests

Prolific

Data analysis

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

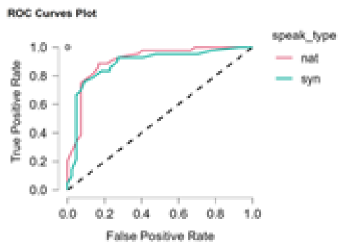
SPEECH MATERIAL

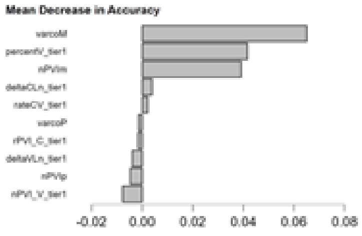
- **685 natural utterances**
  - 137 utterances * 5 speakers (1 source, 4 targets)
    - **44** statements (SVO; 5 words): **LONG UTTERANCE**
    - **10** y/n questions (based on 5-word statement)
    - **83** statements (SV; 2 words): **SHORT UTTERANCE**
- **548 voice converted utterances**
  - 137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

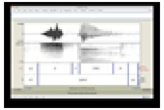
- The rise of multimodal large language models can transform the way we interact with technology
    - From task-oriented interactions to more natural and fluid conversations with users
        - The ontological boundary between AI-powered voice assistants and humans as conversational partners is challenged

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

Annette Rios
Jannis Vamvas

Text Technology Group

University of Zurich[UZH]

Department of Computational Linguistics

rios@cl.uzh.ch

Automatic subtitling for Romansh TV

- Build a dataset consisting of Swiss TV shows and existing German subtitles
- Combine ASR and MT at inference time, or train or adapt a multimodal model to generate Romansh subtitles
- Evaluate performance across language varieties

Supervisors:

Dr. Annette Rios (main)

Dr. Jannis Vamvas

You will deepen the following skills:

- Model training with PyTorch
- Hugging Face Transformers

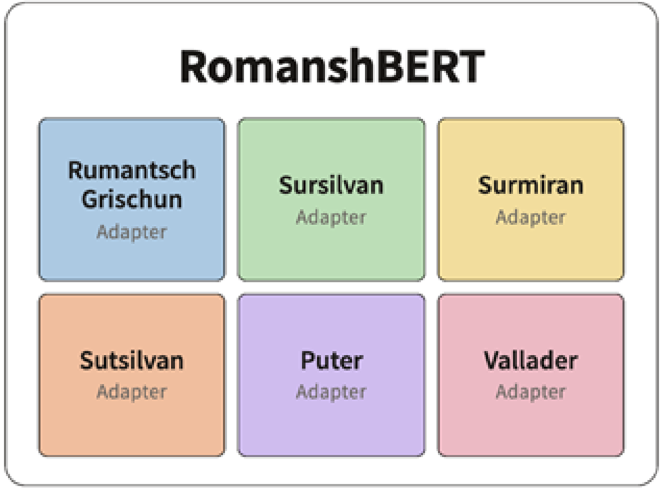
rios@cl.uzh.ch

Text Encoder Model for Romansh Idioms

— Train a BERT-like model on 130M tokens of Romansh text.

— Evaluate your model on tasks such as text classification or named entity recognition.

— Present the results at the SwissText 2027 conference.



Supervisors:

Dr. Annette Rios (main)

Dr. Jannis Vamvas

You will deepen the following skills:

— Model training with PyTorch

— Hugging Face Transformers

University of Zurich[UZH]

Department of Computational Linguistics

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

Magnitude for Analysis of Language Models

Magnitude:

- A geometric tool that measures “diversity” of a point cloud
- Has been used in CL, e.g., to “fingerprint” language models via analysis of their embeddings

Idea/task:

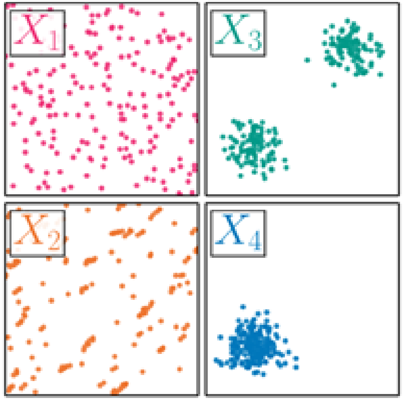
- Develop necessary understanding of magnitude as a tool, and investigate the value of magnitude for CL by applying it to problems such as, e.g.,
  - translation (direction) detection;
  - uniformity/alignment tradeoff in contrastive learning;
  - explainability of language models, such as, e.g., effects of training and/or fine tuning on magnitude of latent spaces;
  - …

Prerequisites:

- Python, NLP; “Mathematical Foundations of Computational Linguistics 2” course

Intended for:

- MSc students



Point clouds with varying magnitudes

Michelle Wastl
michelle.wastl@uzh.ch

michelle.wastl@uzh.ch

Sentence Ordering with Text Generation Models

- Automatic sentence ordering has previously relied on explicit discourse modeling or task-specific supervision.
- *Can modern text generation models do this without any task-specific training?*

- Prerequisites: Successful completion of the modules *Text Generation with Language Models* and/or *Large Language Models*

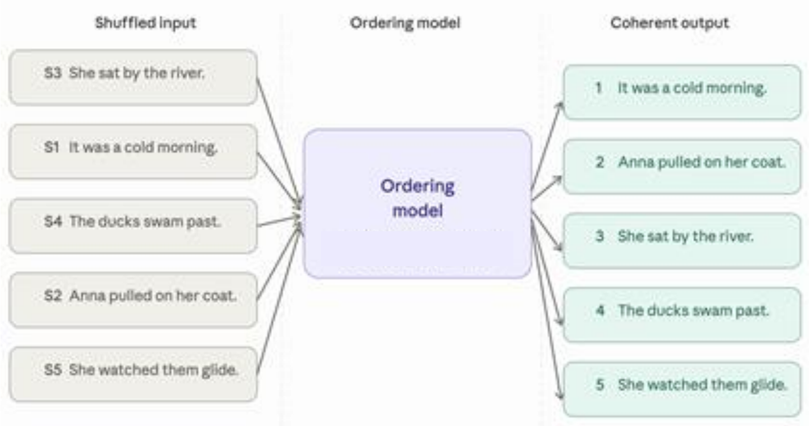


Image generated with Claude.

University of Zurich[UZH]

Department of Computational Linguistics

Rico Sennrich
rico.sennrich@uzh.ch

University of Zurich[UZH]

Department of Computational Linguistics

Automating Formal Review of Grant Applications

Grant applications are long, multilingual, multimodal (including transcripts, passports, etc.)

Collaboration with Research and Grants Office of UZH:

Can current LLMs help the formal review of grant applications?

- are all necessary documents included?
- does candidate fulfill the eligibility criteria of the grant?

Task brings interesting technical challenges:

- processing of long, multilingual and multimodal docs
- long checklist of criteria
- model restriction due to data confidentiality

Opportunity for use-inspired research with real grant applications and pathway to impact.

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

University of Zurich[UZH]
Department of Computational Linguistics

Cui Ding, Lena Jäger
{cui.ding, lenaann.jaege}@uzh.ch

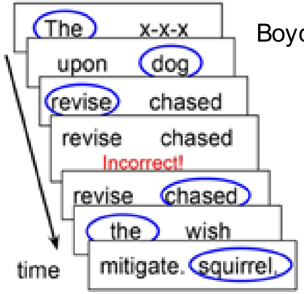
Test working-memory effects in syntactic ambiguity resolution with Mandarin relative clauses

Background:

- Previous work shows that lower-WM readers often prefer high attachment in ambiguous RCs (a finding that is counterintuitive), but English materials introduce a positional confound.
  - **The maid** (N1) of **the princess** (N2) <u>who scratched herself in public</u> (relative clause) was terribly embarrassed.
  - Who scratched herself in public? the maid → high attachment; the princess → low attachment
- In English, the high-attachment noun (**The maid**) is usually the first noun in the sentence. So lower-working-memory readers may choose the high-attachment interpretation not because they prefer high attachment, but because of primacy effect.
- Mandarin Chinese can help test this problem because relative clauses in Mandarin usually come before the noun. This makes it possible to design sentences where structural position and word order are less confounded.
  - **昨天在宴会上喝醉酒的**经理的秘书最近刚在市区买了新房子。

Your Tasks:

- Create stimulus that is similar to the above example sentences + comprehension questions
- Implement a SPR or MoTR or Maze experiment (code exists; need small adaptation).
- Implement working memory test (code exists; need adaptation).
- Collect Chinese participants' reading data + comprehension question + working memory test data
- Analyze the data to study whether high-/low-working memory people prefer to attach high or low when reading Mandarin relative clauses.



Boyce et al., 2020

What Does the Maze Task Measure?

Background:

- The Maze task is widely used to collect reading data, but it may capture more than reading difficulty. Because participants must continuously choose the correct word while rejecting distractors. Participants' performance in Maze may also reflect their working memory, cognitive control abilities.

Research Question:

- **How do distractor properties and individual cognitive abilities jointly shape performance in the Maze task?**

Your Tasks:

- Prepare stimuli + comprehension questions or use existing ones.
- Design experiment, e.g., distractor difficulty levels and types; maybe add a SPR/MoTR experiment to make comparison
- Run Maze distractor generation software (code exists; need adaptation depending on the target languages)
- Implement a Maze experiment (code exists).
- Implement working memory and/or cognitive control and/or reading fluency test(code exists; need small adaptation).
- Collect data
- Analyze data. We are particularly interested in how Maze difficulty/type interact with WMC or cognitive control ability.
  - How is the interaction different from the interaction with reading fluency?
  - *How is the interaction different from the one in SPR/MoTR experiment?

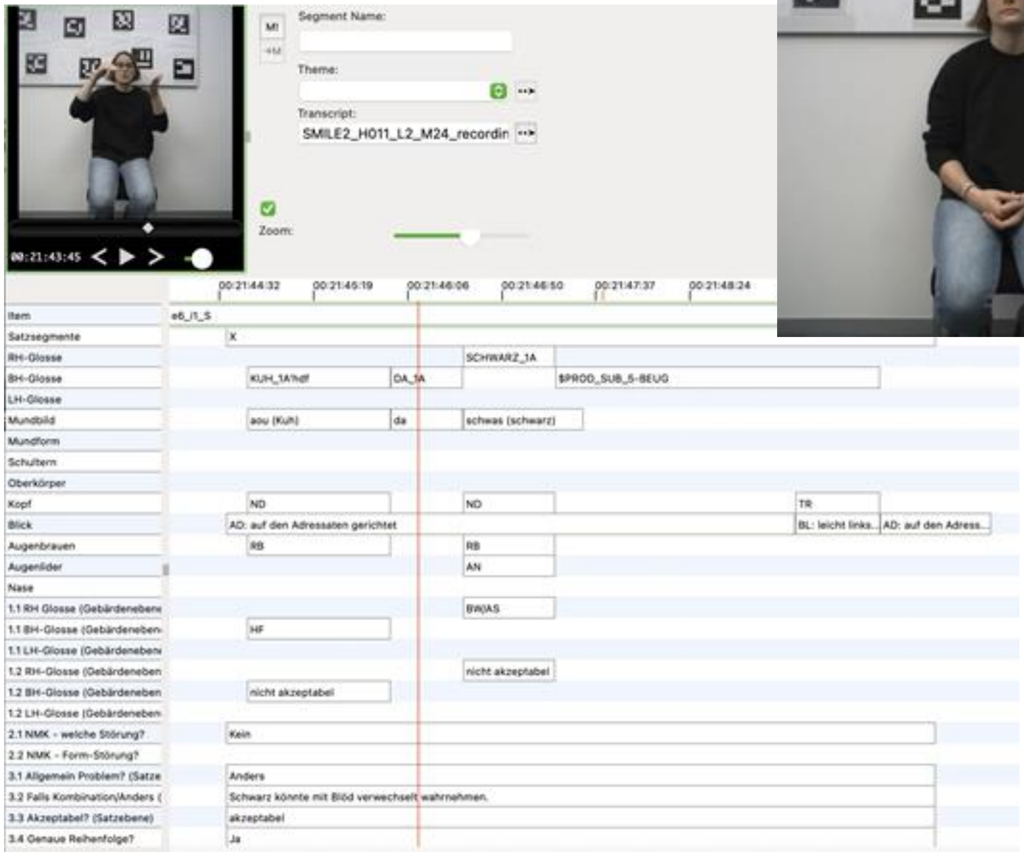
Sarah Ebling
ebling@cl.uzh.ch

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, specifically considering non-manual aspects of signing

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".

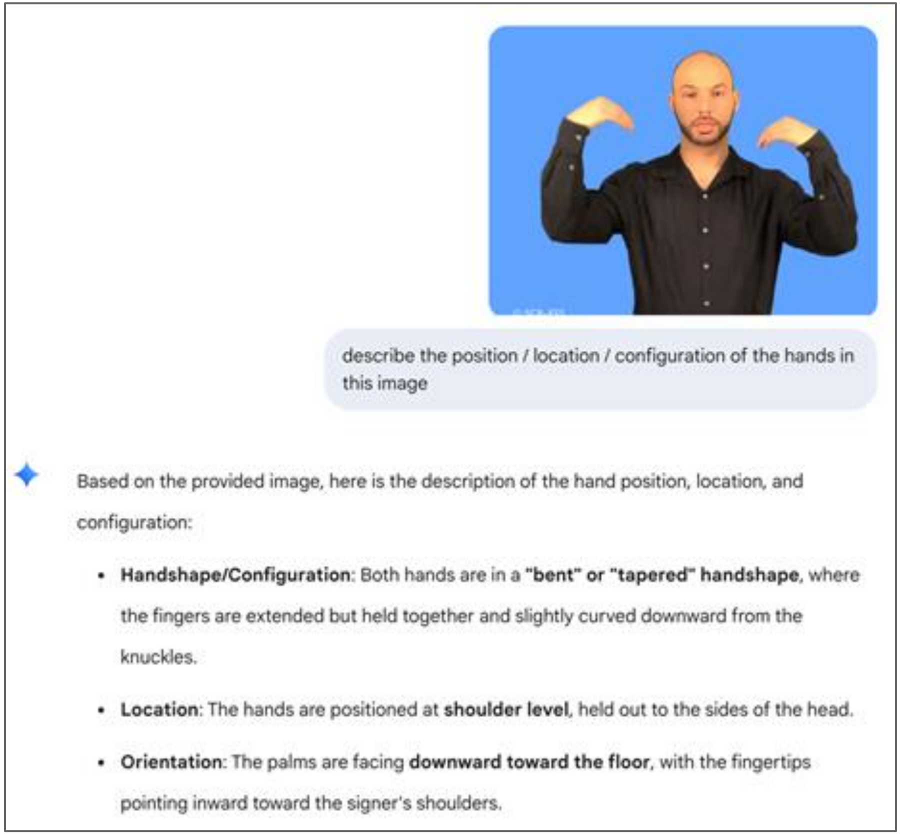


Describing sign language movement with a multimodal language model

Supervisor: Sarah Ebling

Summary: Use the newest and most powerful multimodal LLMs (text plus images or video) to describe the motion in a sign language video. Evaluate whether this could be helpful for translation systems.

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"



University of Zurich[UZH]

Department of Computational Linguistics

Mathias Müller, Zifan Jiang

{mmueller,jiang}@cl.uzh.ch

Automatic segmentation of sign languages

Supervisors: Mathias Müller, Zifan Jiang

Summary: A segmentation model takes a sign language video as input and splits it up into meaningful parts. Help us analyze our existing segmentation models on more benchmarks and then train better ones

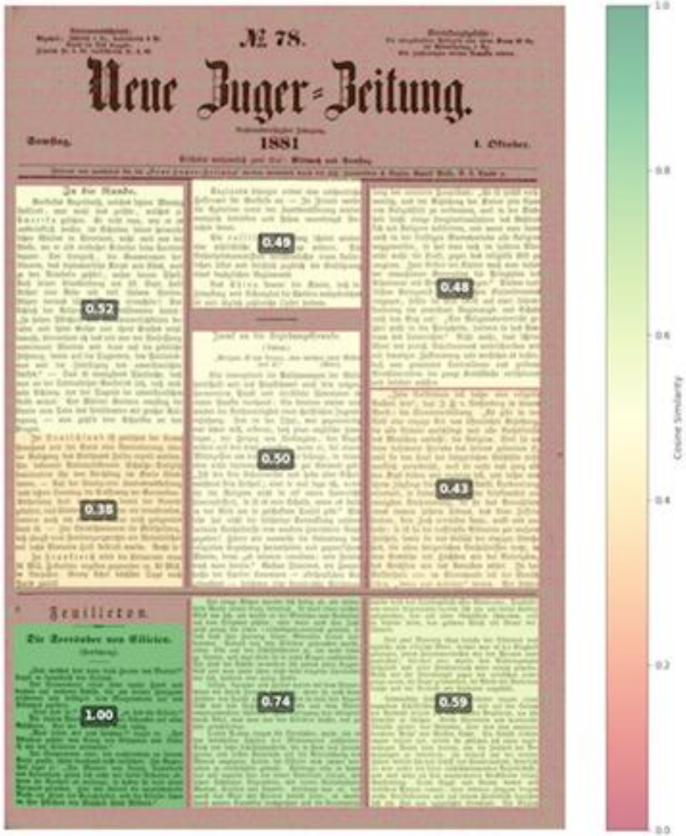
Requirements: Deep learning, Python, familiarity with or interest in computer vision techniques (in addition to NLP techniques), ideally successful completion of the course "Artificial Intelligence for Language Accessibility", "Visual Language Processing" or "Digital Accessibility"

University of Zurich[UZH]

Department of Computational Linguistics

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

Impresso II UZH CL Student Projects

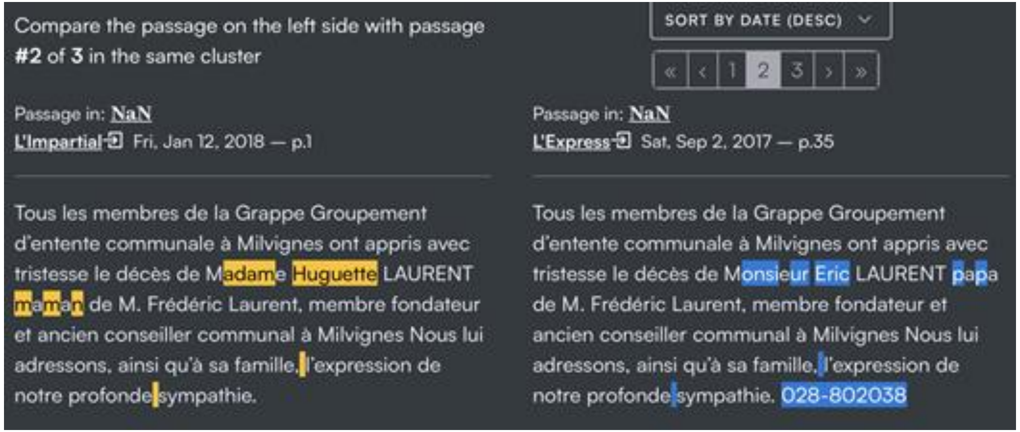


Andrianos Michail

Juri Opitz

Simon Clematide

impresso
Media
Monitoring
of the Past



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**

Machine Unlearning Embedding Model Biases

Research problem:

- **Embedding models** have implicitly learned **undesirable biases** through their training processes. An examples of interest of us include **over-reliance on surface-level (lexical) similarity**.

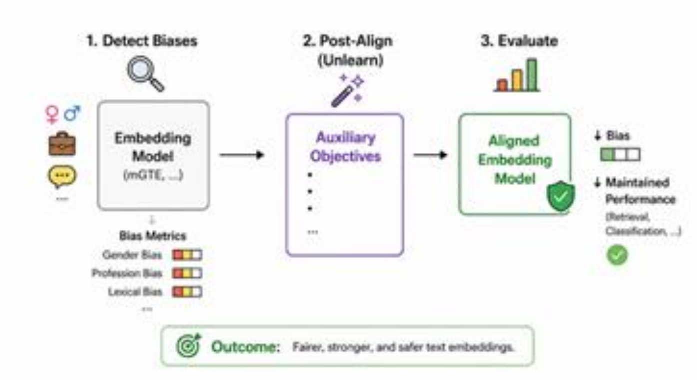
RQ:

1. Which of these undesirable biases are present in current embedding models, and how can they be measured automatically?

2. Can embedding models (e.g., mGTE) be **post-aligned via auxiliary objectives** to **unlearn** these properties while preserving downstream performance?

Expected outcome:

- Text embedding models with reduced biases
- Quantitatively evaluated



Suitable for an **MA**(CL/IFI) Thesis
Contact: **Andrianos Michail**

Efficient Person-Place Relation Detection in News Articles

Research problem:

- Through our **HIPE 2026** evaluation campaign we have seen that Large Language Models are reasonably capable in qualifying the relations between Entity and Locations that co-occur in a newspaper.

RQ: 1. Can we also find efficient methods of performing this **Person Place** Relation predictio using smaller models, perhaps through knowledge distillation?


JOURNAL DE GENEVE
et Gazette de Lausanne
LE DERNIER NUMERO
Au revoir
Une journée pour en finir

2. **Expected outcome:**

- Quantitative evaluation of efficient to implement at scale Entity and Locations relatio predictions.
- A cookbook on how to efficiently perform the task at scale within Impresso II.

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

Contact: **Simon Clematide**

Distillation of Keyphrase Abstraction Capabilities on Efficient to Compute Models

Research problem:

- **Keyphrase Abstraction** (finding fitting keywords to make longer texts searchable) works reasonably well when using large enough (e.g DeepSeek 670B) models. Can we reach similar performance using a recently released efficient to compute model?

RQ:

1. Which recent (smaller) models perform similarly well in Keyphrase Abstraction on **historical newspapers?**

2. Can we make our Keyphrase Abstraction **inference at scale (million of newspapers) more efficient**? Is there some configuration with the latest models that allows us to reduce computation overload without perf. loss?

Expected outcome:

- Quantitative evaluation of recent Large Language Models in the task.
- A cookbook on how to efficiently perform the task at scale within Imp.II

Suitable for a **BA/MA** (CL/IFI) Thesis
Contact: **Simon Clematide**

Other Topics

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
- Topic Modelling
- Relation Mining
- Post-OCR Correction of historical texts.

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

Sina Ahmadi
(sina.ahmadi@uzh.ch)

Anything ***low-resourced***, come talk to me!

University of Zurich[UZH]

Department of Computational Linguistics

Loanword-Resilient Machine Translation (master's thesis)

- Bilingual speakers rely on code-switching and loanwords constantly
- Additional challenges for MT, especially for low-resource communities where lexical pressure is highest
- We already know that:
  - MT systems underperform on "pure" language
  - LLMs prefer loanwords
  - Pretrained models are loanword-blind
- **Questions for you:**
  - **How to mitigate this bias and build MT systems that are resilient to, and controllable over, loanword variation?**
  - **Compare and develop lightweight adaptation strategies**
  - **Analyze whether effectiveness varies with typological distance, borrowing density, and resource level**

More on this:

- ConLoan: A Contrastive Multilingual Dataset for Evaluating Loanwords (ACL 2025)
- Are Language Models Borrowing-Blind? A Multilingual Evaluation of Loanword Identification across 10 Languages (LREC 2026)

Evaluation Metrics and Benchmarks for Dialectal Variation in MT

- Do our metrics actually measure what we think they measure?
  - Surface-level metrics (e.g., BLEU) rely on exact or near-exact string matching
  - Trained metrics (e.g., COMET) inherit biases from their training data
- Both approaches penalize valid dialectal forms and break down when variants are equally correct
- **Goal: develop variety-aware metrics and multilingual benchmarks for dialectal MT**

Variety	Source Sentence	MT Output (NLLB)
Standard	*Wir leben im Zeitalter der Technik.*	We live in the age of technology.
Bern	*mir läbä im zitauter vor technik.*	I'm a little overwhelmed by the technique.
Graubünden	*miar leben im ziitalter dr technik.*	I think we're living in the age of technology.
St. Gallen	*mir lebed im ziitalter de technik.*	I lived in the age of technology.
Wallis	*mir läbu im zitalter der technik.*	I was in the age of technology.
Zürich	*mir läbu im zitalter der technik.*	I was in the age of technology.