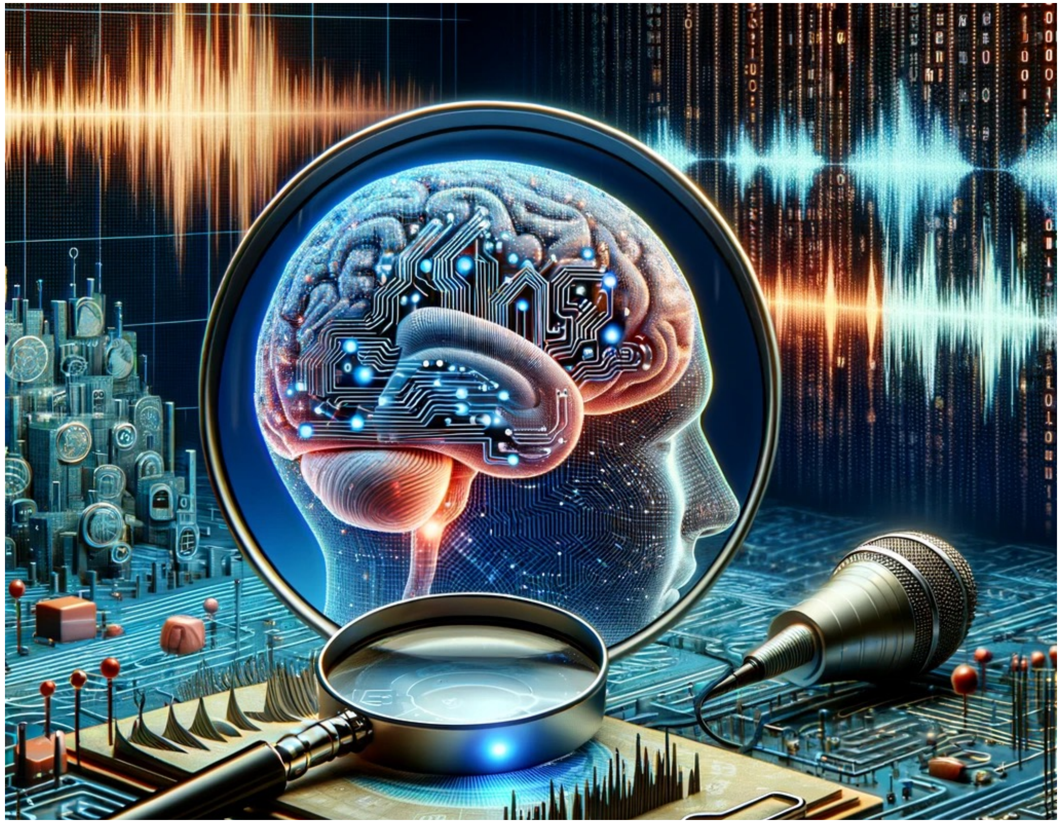
University of Zurich^UZH

Department of Computational Linguistics

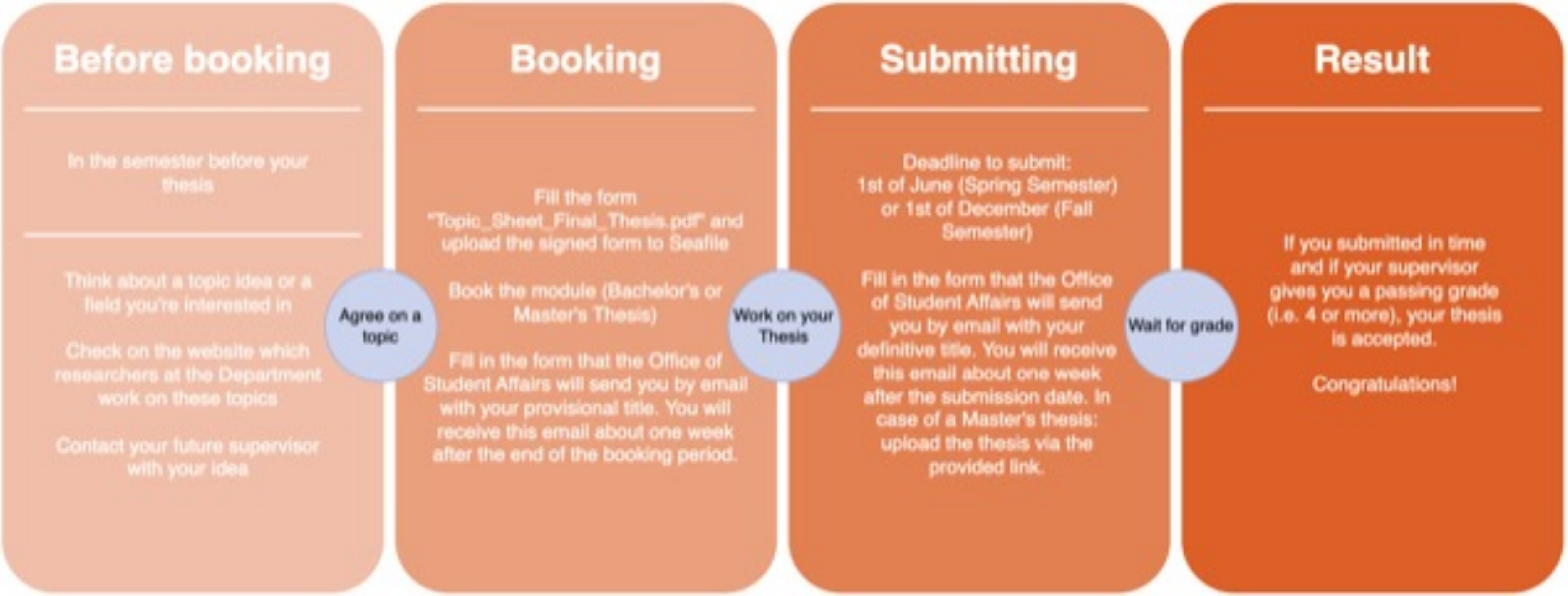
BA/MA projects at the Department of CL



Jeannette Roth

Daniel Friedrichs

Process and Organisation



Before booking

In the semester before your thesis

Think about a topic idea or a field you're interested in

Check on the website which researchers at the Department work on these topics

Contact your future supervisor with your idea

Agree on a topic

Booking

Fill the form "Topic_Sheet_Final_Thesis.pdf" and upload the signed form to Seafile

Book the module (Bachelor's or Master's Thesis)

Fill in the form that the Office of Student Affairs will send you by email with your provisional title. You will receive this email about one week after the end of the booking period.

Work on your Thesis

Submitting

Deadline to submit:
1st of June (Spring Semester) or 1st of December (Fall Semester)

Fill in the form that the Office of Student Affairs will send you by email with your definitive title. You will receive this email about one week after the submission date. In case of a Master's thesis: upload the thesis via the provided link.

Wait for grade

Result

If you submitted in time and if your supervisor gives you a passing grade (i.e. 4 or more), your thesis is accepted.

Congratulations!

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 Qualifications:** Master's degree or higher
- **Booking:** Via Student Portal in the standard booking period

MA Thesis (Study regulations § 33 - § 35)

- **Credits:** 30 ECTS, compulsory, graded
- **Duration**: 2 semesters
- **Submission Deadlines:** June 1 (spring semester) / December 1 (fall semester)
- **Thesis:** Individual, no co-authorship
- **Supervisor Qualifications:** Must hold a PhD
- **Booking:** Via Student Portal in the standard booking period

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

University of Zurich^UZH

Department of Computational Linguistics

Supervisors Presenting Today

Daniel Friedrichs
Steven Moran
Eleanor Chodroff
Sandra Schwab
Jean-Philippe Goldman
Jan Brasser
Lena Jäger
Elisa Pellegrino
Jannis Vamvas
Sarah Ebling
Amit Moryossef
Rico Sennrich
Deborah Jakobi
Andrianos Michail
Simon Clematide
Gerold Scheider
Janis Goldzycher
Nora Hollenstein (slides attached only)

Daniel Friedrichs & Steven Moran

Some ideas:

Kinetic Task Analysis: Explore differences in diadochokinetic tasks, such as Alternating Motion Rates (AMR) and Sequential Motion Rates (SMR). Use data from Electromagnetic Articulography (EMA), Ultrasound Tongue Imaging (UTI), and Electroencephalography (EEG) to understand better why SMR demonstrates quicker productions.

Articulatory Synthesis and Biomechanical Modeling: Enhance 3D models of articulatory movement by integrating combined EMA and UTI data. Develop comprehensive simulations (e.g., a dynamic tongue model) using the modeling toolkit/platform ArtiSynth.

Biological and Environmental Effects on Language: Investigate how anatomical variation and climatic conditions influence human sound systems. Assess the extent to which these factors contribute to the diversity of language.

Contact:
✉ daniel.friedrichs@uzh.ch , bambooforest@gmail.com

Eleanor Chodroff

Phonetic typology with massively multilingual speech corpora

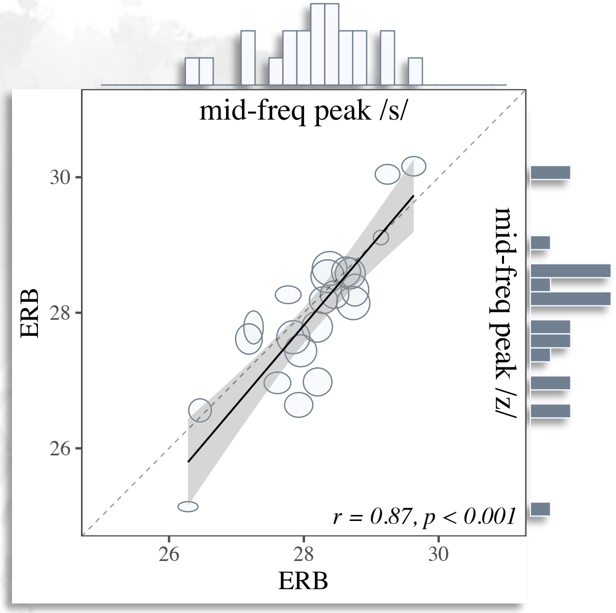
Prof. Dr. Eleanor Chodroff

Prerequisites:

- Praat scripting
- Statistical analysis with R or Python

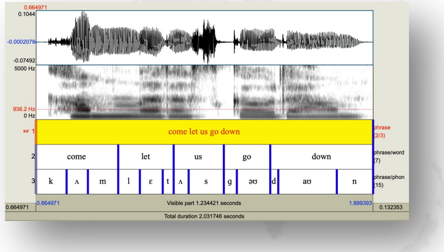
Goals:

- Investigate previously proposed phonetic "universals" at the descriptive and analytic levels across a large number of diverse languages
- Develop and refine crosslinguistic speech resources



Important notes:

- This is a broad proposal → Considerable room for variability in the exact project topic (intrinsic f0, consonant f0, uniformity, suprasegmental aspects of speech)



Variability and systematicity in L1 and L2 speech patterns

Prof. Dr. Eleanor Chodroff

Prerequisites:

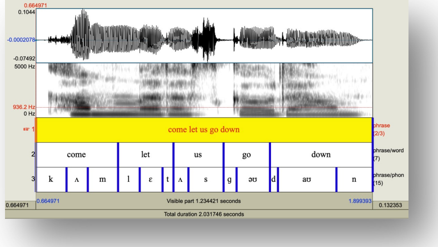
- Praat, scripting is a bonus
- Basic statistical analysis with R or Python

Goals:

- Investigate variability and systematicity in phonetic realisation across a speaker's L1 and L2 speech productions
- Methods: use the ALLSSTAR Corpus (Archive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings) or an alternative collection of bilingual speech for acoustic-phonetic analysis

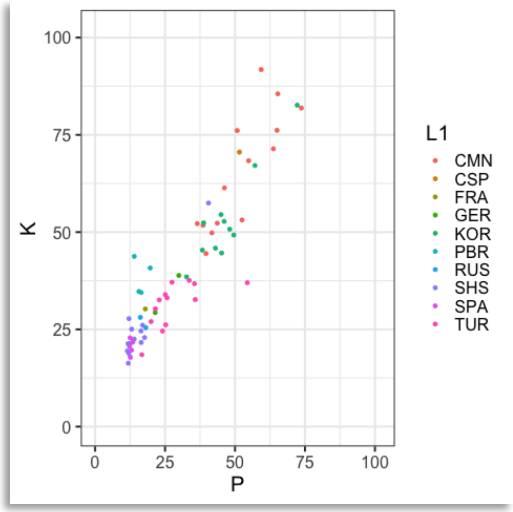
Important notes:

- This is a broad proposal → Considerable room for variability in the exact project topic (stop voice onset time, sibilant spectral properties, vowel formants, rhythm, etc.)



Visible part 1.234421 seconds

Total duration 2.031746 seconds



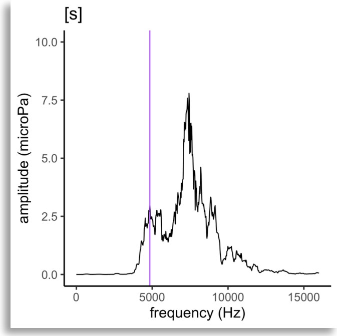
Shift from L1 /p/–/k/ VOT to L2 English /p/–/k/ VOT (ms)

Acoustic–articulatory relationships in sibilant fricatives using EMA

Prof. Dr. Eleanor Chodroff

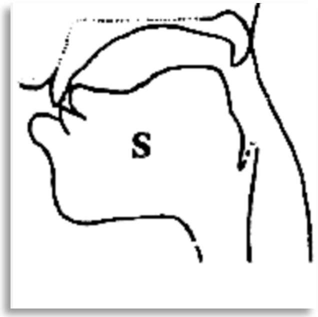
Prerequisites:

- Praat, scripting is a bonus
- Basic statistical analysis with R, Python or Matlab



Background:

- Mid-frequency peak: a major peak frequency typically located between 2000 and 7000 Hz that supposedly reflects the front cavity resonance — that is the space between the tongue constriction and the front teeth
- Limited validation of the relationship



Goals:

- Use electromagnetic articulography (EMA) and acoustic analysis to investigate the relationship between tongue constriction location and the mid-frequency peak in sibilant fricatives (/s z ʃ ʒ/)

Perceptual generalization in adapting to novel speakers

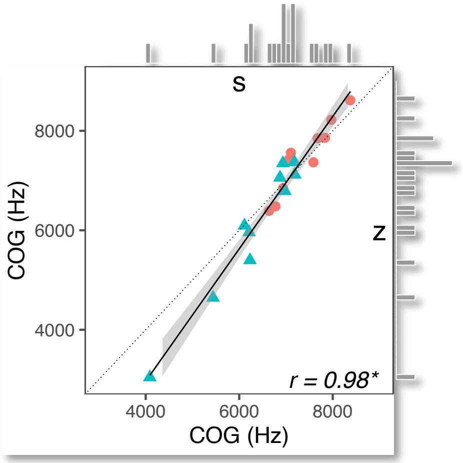
Prof. Dr. Eleanor Chodroff

Prerequisites:

- Phonetics / speech science coursework

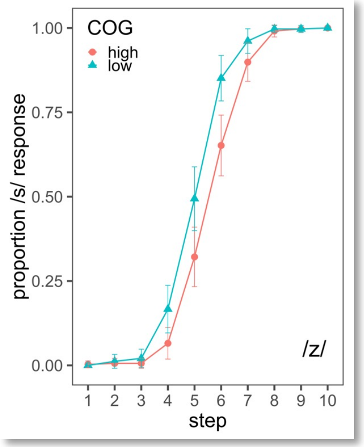
Background:

- Listeners may generalise speaker-specific patterns from the speech sounds they have heard to speech sounds that have not yet occurred in the exchange
- Limited or incomplete evidence for when this might occur



Goals:

- Conduct a series of speech perception experiments to investigate whether listeners generalize aspects of one speech sound (e.g. /s/) to another speech sound (e.g., /z/)
- Determine whether listeners generalize using knowledge of the phonetic relationship or a more general auditory mechanism such as spectral contrast



Important notes:

- This is a broad proposal → Considerable room for variability in the exact project topic (stop voice onset time, sibilant spectral properties, vowel formants, etc.)

Sandra Schwab & Jean-Philippe Goldman (UniGE)

Automatic prominence detection in L2

Framework

- Computer-assisted pronunciation teaching (**CAPT**)
- Focus on L2 **stress contrasts**: e.g. import vs. import
- **Automatic prominence detection in speech signal**

Two goals

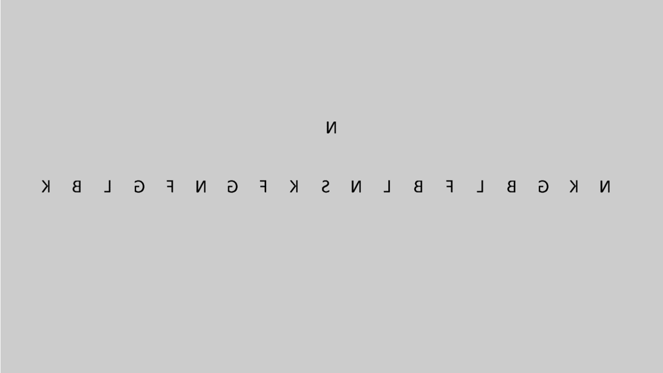
- **Train** system to develop **L1 German/Italian stress detector**
- **Assess** and **adapt the L1 system to L2** German/Italian to be implemented in **Miaparle** (miaparle.unige.ch)

Requirement: Strong background in **Machine Learning**

Co-supervision: Sandra Schwab (UZH) & Jean-Philippe Goldman (UniGe)

Lena Jäger & Jan Brasser

Symbol Complexity in Visual Search and Reading



- As part of the project “Lesen im Blick” (Eye-Tracking-based dyslexia detection), we investigate the difference in visual search behavior between letter and non-letter stimuli
- Research Question:
  - Are there differences in complexity between the symbols used within each experimental condition?

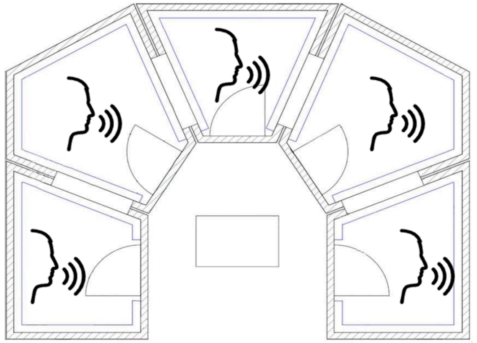
- For your BA/MA thesis you will:
  - Conduct an eye-tracking experiment containing the visual search task with adult participants
  - Analyze the complexity of the different symbols in each of the three conditions based on the eye-tracking data

Elisa Pellegrino

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

Elisa Pellegrino, Volker Dellwo in collaboration with A. Hervais Adelman and E. Varano

Data collection



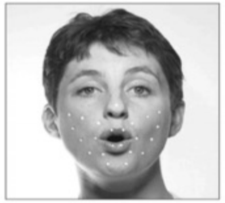
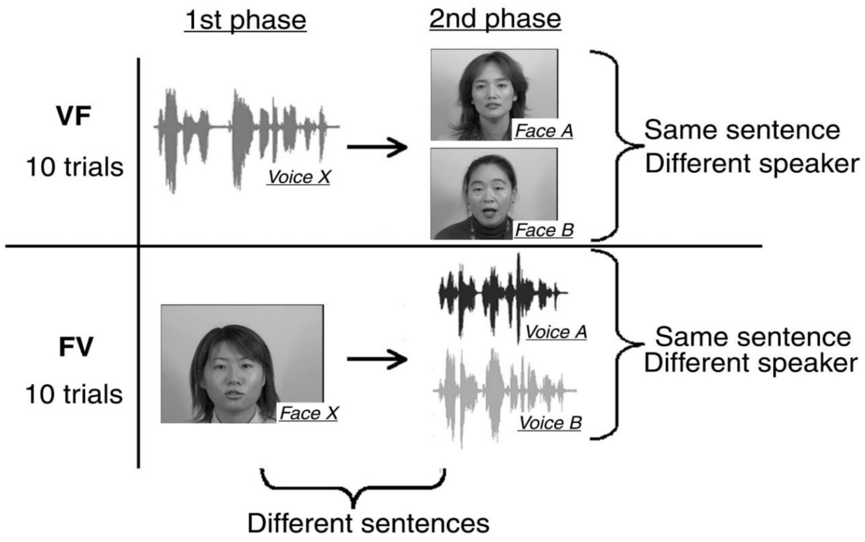


Fig. 2 Arrangement of reflexive dots

Fig. 3 Point-light video

Fig. 4 Fully illuminated displays

Cross-modal identity matching task



1st phase

2nd phase

VF
10 trials

Voice X → Face A, Face B

Same sentence
Different speaker

FV
10 trials

Face X → Voice A, Voice B

Same sentence
Different speaker

Different sentences

STWF-21-19

The Role of Speech Acoustics to Detect Deep Fake Voices

Elisa Pellegrino

Voice conversion dataset

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
- 45 y.o.

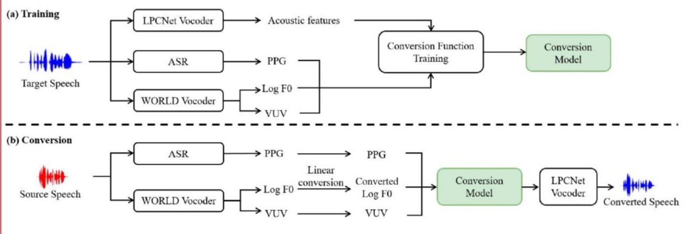
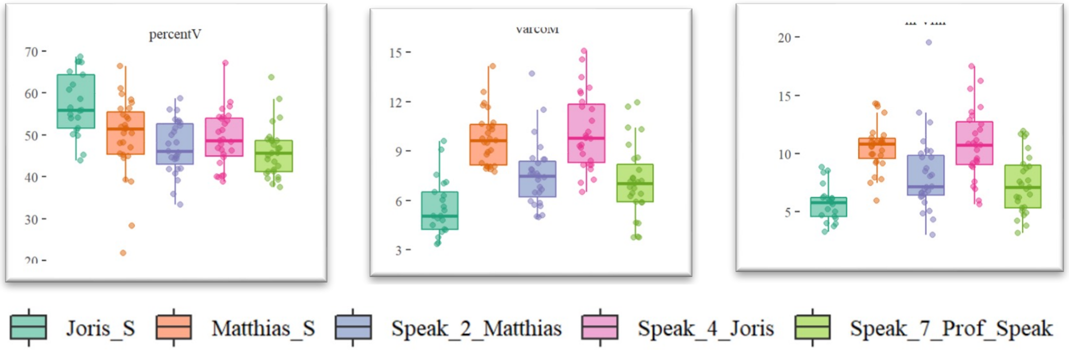


Figure 1: *The framework of the baseline system.*

→ Voice Converted Utterances have the **timbre/segmental properties** of the **TARGET SPEAKER** but the **prosodic features** of the **SOURCE SPEAKER**

Preliminary results



Expected outcomes

Contribution of acoustic features to the distinction between natural and synthesized counterpart

A. Leemann et al. / Forensic Sc

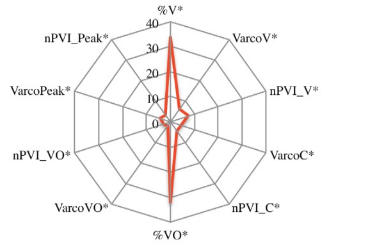


Fig. 1. Radar chart illustrating the individual contribution of the ten predictors i the nominal logistic regression model for *speaker* based on read speech.

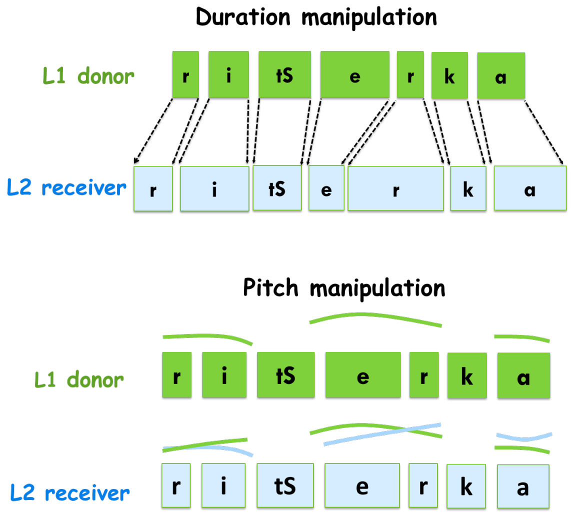
The Effect of Prosody Training on Second Language Acquisition

Elisa Pellegrino

Prosodic Training via Self-imitation

Foreign accent predicted by prosodic and segmental deviations from L1 speakers

In classroom setting little to no attention to the L2 prosody acquisition



Self-imitation prosodic training improves pragmatic competence

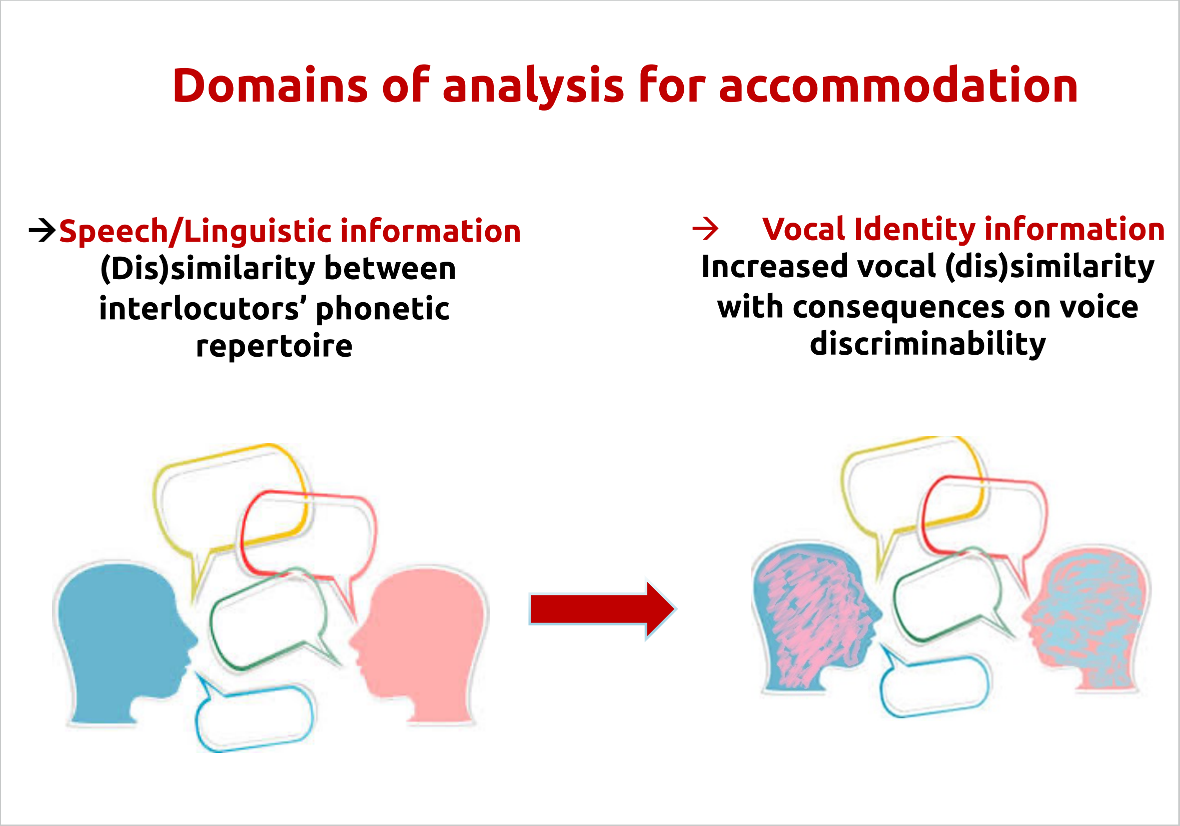
	Pre-training (A)	Post-training (B)	Difference (B – A)
Request	52,52%	75,21%	22,69
Command	39,92%	57,98%	18,06
Grant	8,40%	47,06%	38,66

Possible investigations:
In what way does prosodic training affect the performance?

Only prosody? Also pronunciation of individual sounds'? Does prosodic training improve accentdness?

The time course of vocal accommodation in speech communication and Its Effect on Speaker Recognizability

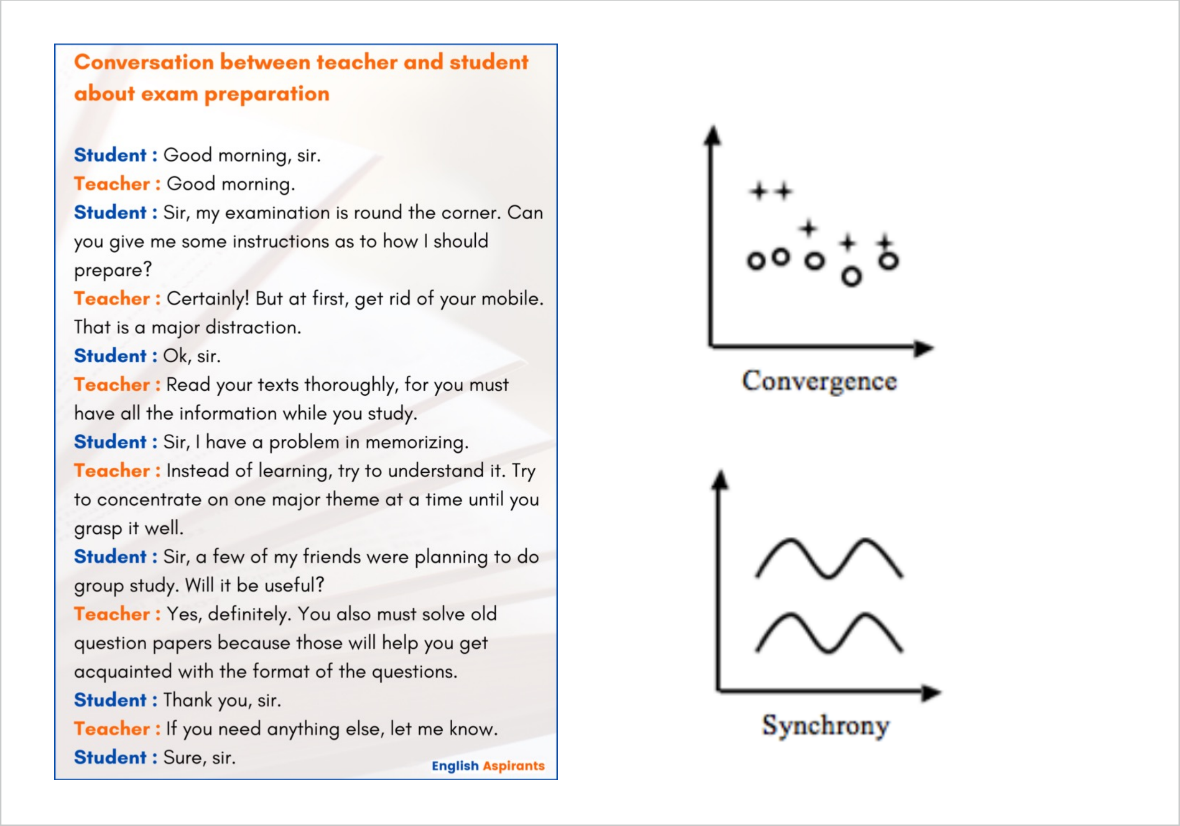
Elisa Pellegrino



Domains of analysis for accommodation

→**Speech/Linguistic information**
(Dis)similarity between interlocutors' phonetic repertoire

→ **Vocal Identity information**
Increased vocal (dis)similarity with consequences on voice discriminability



Conversation between teacher and student about exam preparation

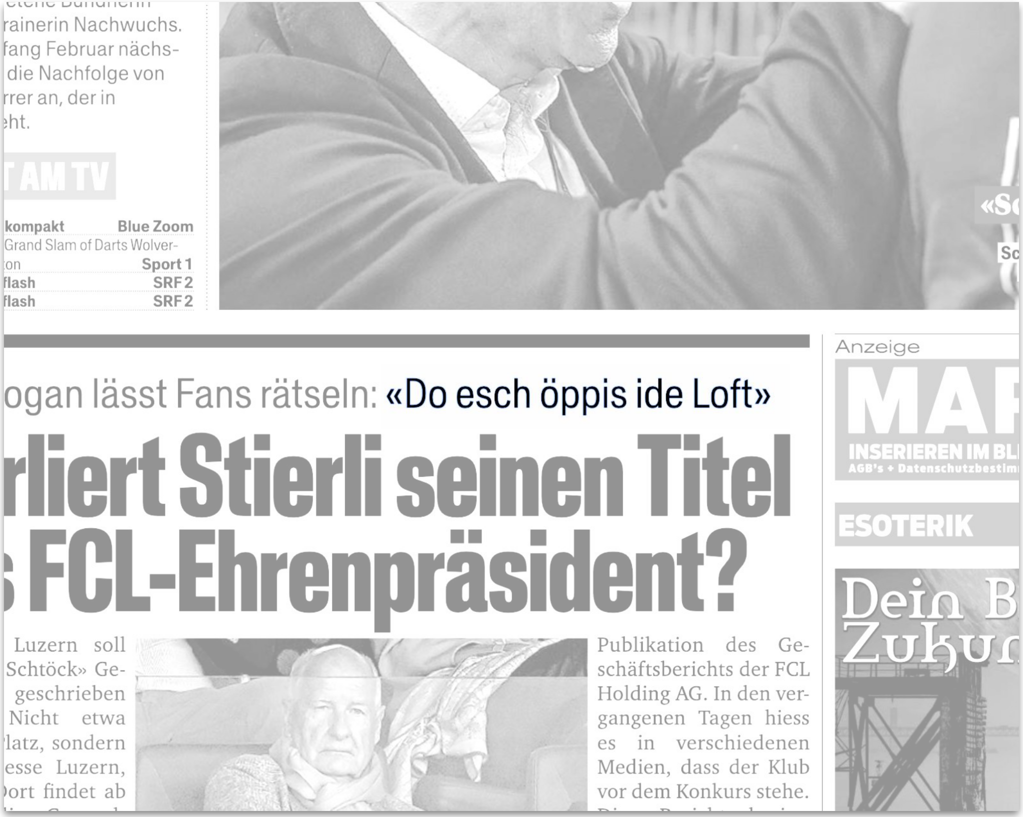
Student : Good morning, sir.
Teacher : Good morning.
Student : Sir, my examination is round the corner. Can you give me some instructions as to how I should prepare?
Teacher : Certainly! But at first, get rid of your mobile. That is a major distraction.
Student : Ok, sir.
Teacher : Read your texts thoroughly, for you must have all the information while you study.
Student : Sir, I have a problem in memorizing.
Teacher : Instead of learning, try to understand it. Try to concentrate on one major theme at a time until you grasp it well.
Student : Sir, a few of my friends were planning to do group study. Will it be useful?
Teacher : Yes, definitely. You also must solve old question papers because those will help you get acquainted with the format of the questions.
Student : Thank you, sir.
Teacher : If you need anything else, let me know.
Student : Sure, sir.

English Aspirants

Jannis Vamvas

Mining a newspaper archive for Swiss German

Jannis Vamvas
vamvas@cl.uzh.ch



rainerin Nachwuchs.
fang Februar nächs-
die Nachfolge von
rrer an, der in
eht.

T AM TV

kompakt	Blue Zoom
Grand Slam of Darts Wolver- ton	Sport 1
flash	SRF 2
flash	SRF 2

ogan lässt Fans rätseln: «Do esch öppis ide Loft»

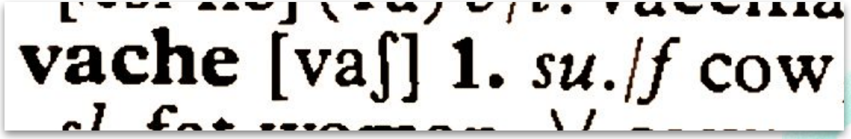
rliert Stierli seinen Titel s FCL-Ehrenpräsident?

Luzern soll Schtöck» Gegeschrieben Nicht etwa latz, sondern esse Luzern, Dort findet ab

Publikation des Geschäftsberichts der FCL Holding AG. In den vergangenen Tagen hiess es in verschiedenen Medien, dass der Klub vor dem Konkurs stehe.

Generation of L2 mnemonics

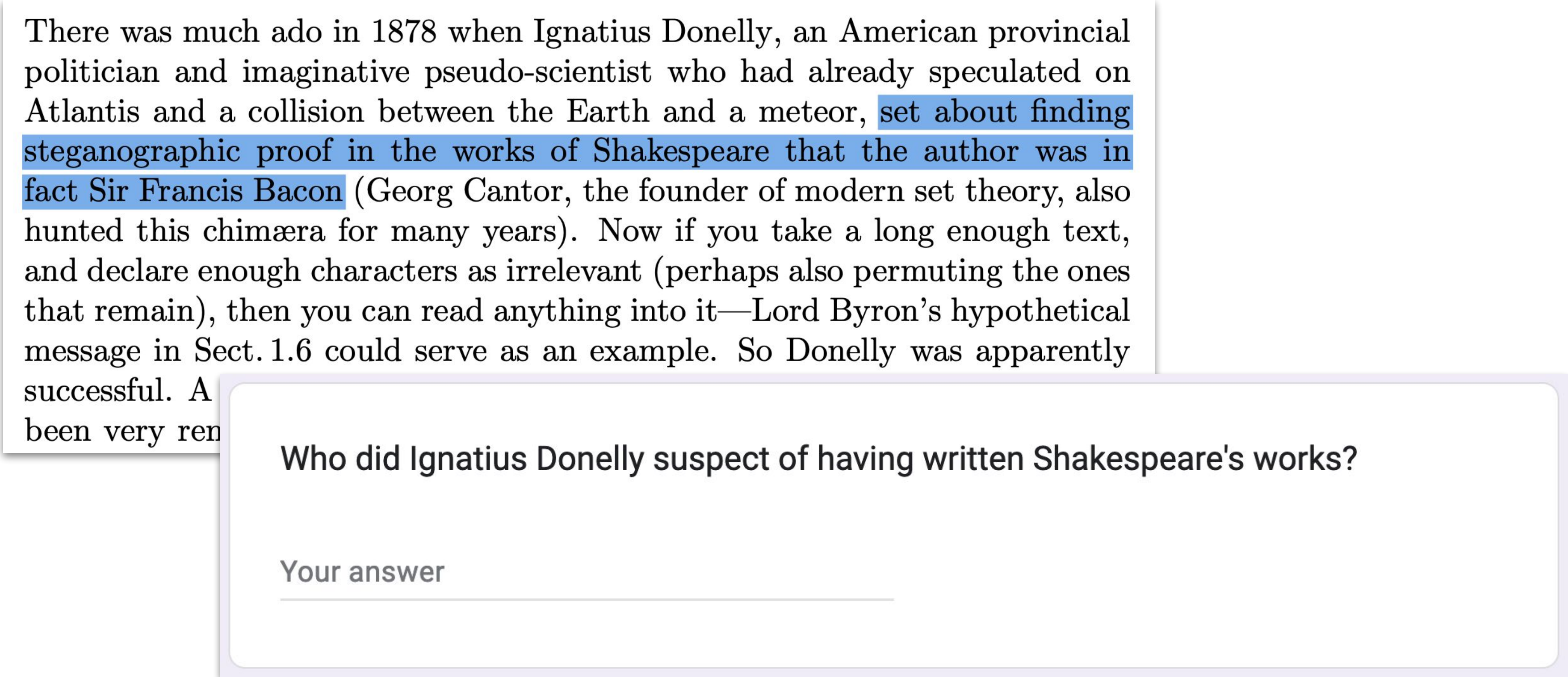
Jannis Vamvas
vamvas@cl.uzh.ch



vache [vaʃ] **1.** *su./f* cow

Book-to-quiz conversion

Jannis Vamvas
vamvas@cl.uzh.ch



There was much ado in 1878 when Ignatius Donelly, an American provincial politician and imaginative pseudo-scientist who had already speculated on Atlantis and a collision between the Earth and a meteor, set about finding steganographic proof in the works of Shakespeare that the author was in fact Sir Francis Bacon (Georg Cantor, the founder of modern set theory, also hunted this chimæra for many years). Now if you take a long enough text, and declare enough characters as irrelevant (perhaps also permuting the ones that remain), then you can read anything into it—Lord Byron's hypothetical message in Sect. 1.6 could serve as an example. So Donelly was apparently successful. A

been very ren

Who did Ignatius Donelly suspect of having written Shakespeare's works?

Your answer

Sarah Ebling & Amit Moryossef

Exploring Production Difficulty of Manual Parameters of Swiss German Sign Language Signs

Sign

Handshape
Hand orientation
Location
Movement

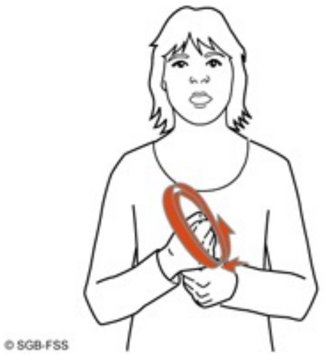
Production difficulty

Which parameters and combinations of parameters contribute to sign production difficulty?

Connections to CEFR?

Additional elements:

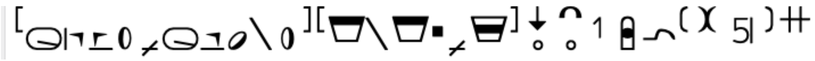
Semantics, movement trajectories, etc.



https://www.sgb-fss.ch/signsuisse/lexikon/114371/arbeiten

ARBEITEN

Hamburg Notation System:

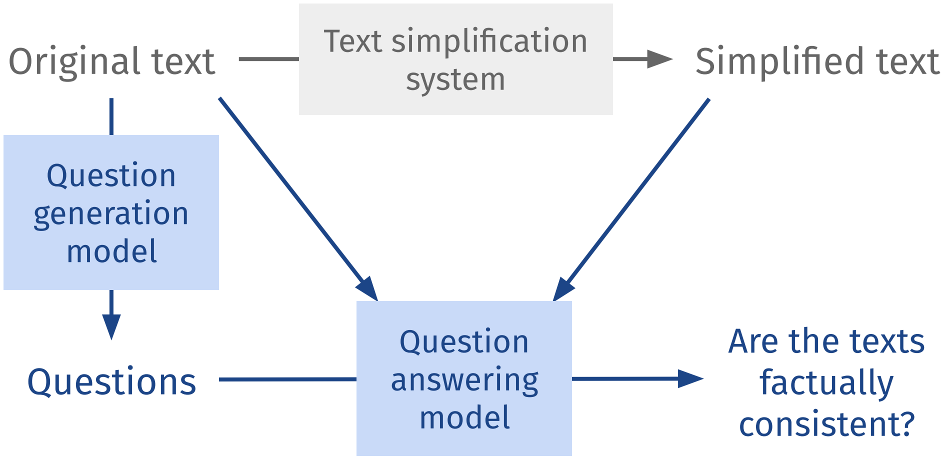


Evaluation of factual consistency in simplified texts

How can we evaluate whether an automatically simplified text contains all (and only) the information that was present in the original text?

Goals:

- **Develop an automatic evaluation pipeline** based on question generation and question answering models.
- Test the pipeline on automatically and manually simplified texts.
- Compare the pipeline to human judgments of factual consistency.



Original text
Text simplification system
Simplified text
Question generation model
Questions
Question answering model
Are the texts factually consistent?

Contact: Sarah Ebling (ebling@cl.uzh.ch)

Quality Estimation for Text Simplification / Simplicity Scoring

Background:

- part of the “Inclusive Information and Communication Technologies (IICT)" project:
- subproject 1: automatic text simplification (ATS)
- goal: develop system that can reliably score simplified texts
- first version: German

Outline:

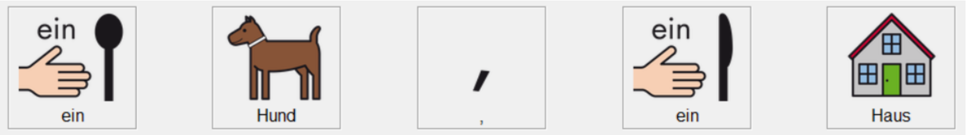
- you will get a set of guidelines on how to write ‘easy language’
- these guidelines have ‘rules’, you decide then:
    - what can be implemented as a formal rule, e.g. with spacy
    - what is better represented as a score, e.g. from a LLM
- later, you will also receive a set of translations with human scores
- you can then ‘tune’ your scoring system with those human scores as a reference

Requirements:

- good python skills, familiarity with spacy + huggingface is a plus
- good German skills
- familiarity with Unix/shell scripts is a plus

Investigating the potential of text-to-image models for pictogram generation

- BA thesis, jointly supervised by Amit Moryossef and Sarah Ebling
- Goal: generate images in the style of METACOM pictograms (https://www.metacom-symbole.de/) in an open-vocabulary setting



Rico Sennrich

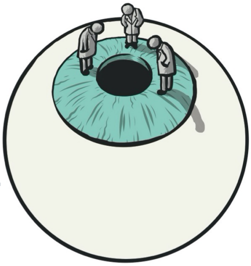
Machine Translation and Multilinguality (Rico Sennrich)

sample topics:

- **create NLP dataset for low-resource language** (MT, summarization, sentiment, …)
  get the most out of limited training data with cross-lingual transfer learning
- **reproduce recent publications and gain novel insights into them (understand limitations, generalization to new settings, perform error analysis)**
- **investigate low-resource machine translation with Large Language Models**
  first results show poor quality. Explore main errors and investigate how to improve it.

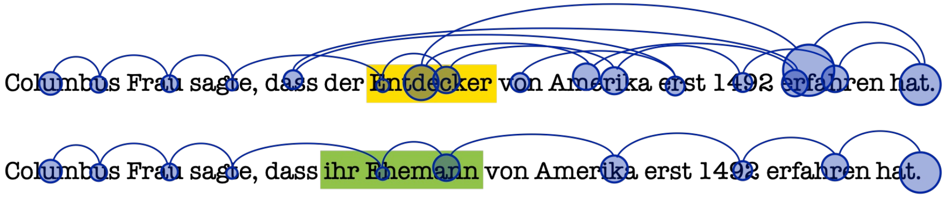
requirements: you have taken class "Machine Translation" or "Advanced Techniques of MT"

Lena Jäger & Deborah Jakobi



MA topic: Studying Psycholinguistic Theories on Language Processing using a Natural Reading Eye-Tracking Corpus

- Eye-tracking data collection in our lab & preprocessing
  - Participants read different texts and we track the eye-movements
  - You're not the only one collecting this data! Researchers across Europe are collecting data using the same texts in different languages
- Analysing the data
  - Eye movements can give insight into cognitive processes
  - Study psycholinguistic research questions
  - E.g are there text types that are more difficult to process?



Columbus Frau sagte, dass der Entdecker von Amerika erst 1492 erfahren hat.

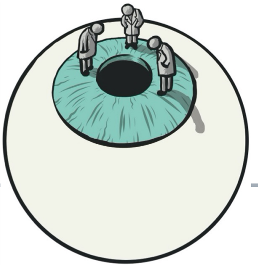
Columbus Frau sagte, dass ihr Ehemann von Amerika erst 1492 erfahren hat.

Prerequisites
- Experience with statistical analysis
- R (or Python)

Supervision
- Digital Linguistics lab
- Prof. Lena Jäger & Deborah Jakobi

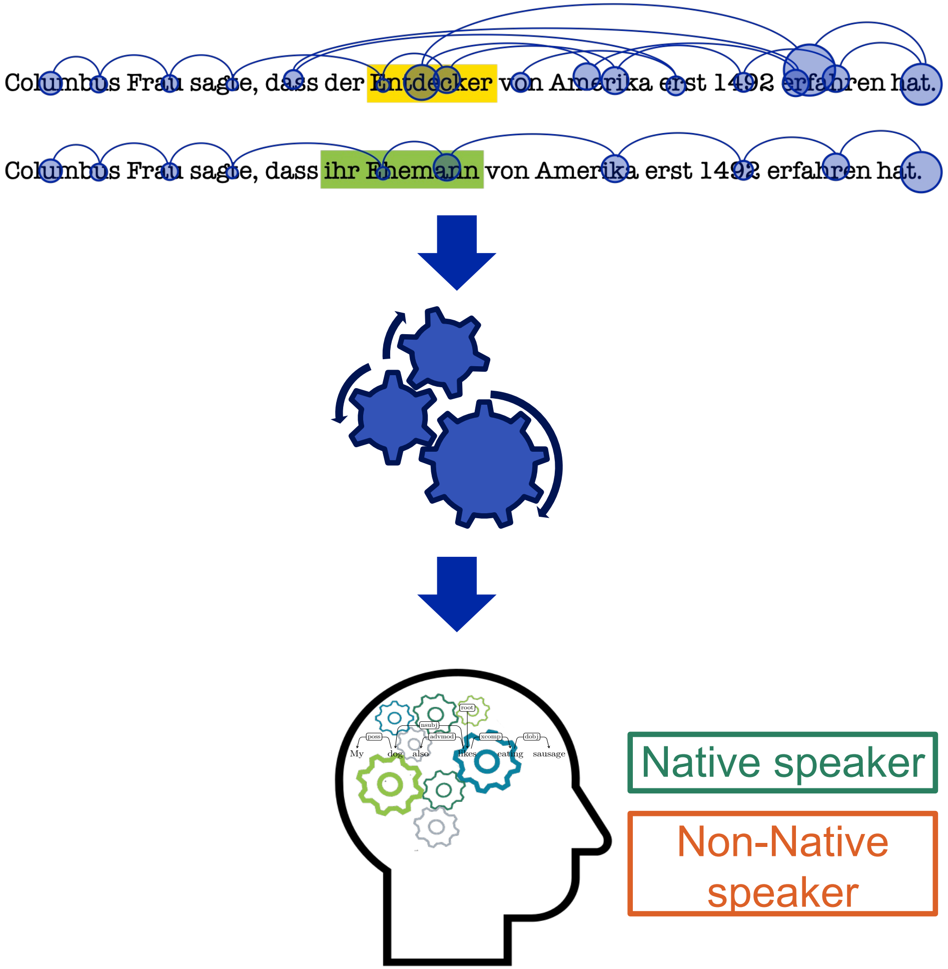
University of Zurich^UZH

Department of Computational Linguistics



MA topic: Leveraging Eye-Tracking Data for Assessing Different Reader Characteristics

- Predict reader characteristics based on eye-tracking data
  - You will be given an eye-tracking corpus that you will need to preprocess in order to use it as input for an ML model
- Design different ML models that can be used to predict specific reader characteristics (=reader inference)
  - Compare different approaches
  - Evaluate / tune / adapt / … your models
  - E.g. try to predict whether a reader is a native speaker of the given language or not



Columbus Frau sagte, dass der Entdecker von Amerika erst 1492 erfahren hat.

Columbus Frau sagte, dass ihr Ehemann von Amerika erst 1492 erfahren hat.

Native speaker

Non-Native speaker

Prerequisites
- Solid background in machine learning
- Python

Supervision
- Digital Linguistics lab
- Prof. Lena Jäger & Deborah Jakobi

Andrianos Michail & Simon Clematide

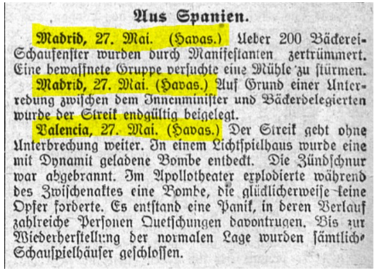
Crosslingual Semantic Search

We expect the student to focus on exploration of this topic through the following methods:

- Study literature and identify existing N-Way Parallel Datasets and how they were sourced.
- Study literature on Bitext Mining to understand the current State Of The Art and its limitations.
- Explore and benchmark different methods identified on existing datasets.
- Apply methods in a limited collection of documents from the Impresso project

Outcome: Robust Cross Lingual Semantic search on Historical newspaper data.

Keywords: Multilingual Models, Bitext Mining, Information Retrieval, Historical news, Deep Learning



Aus Spanien.

Madrid, 27. Mai. (Havas.) Ueber 200 Bäckerei-Schaufenſter wurden durch Manifeſtanten zertrümmert. Eine bewaffnete Gruppe verſuchte eine Mühle zu ſtürmen.

Madrid, 27. Mai. (Havas.) Auf Grund einer Unterredung zwiſchen dem Innenminiſter und Bäckerdelegierten wurde der Streik endgültig beigelegt.

Valencia, 27. Mai. (Havas.) Der Streik geht ohne Unterbrechung weiter. In einem Lichtſpielhaus wurde eine mit Dynamit geladene Bombe entdeckt. Die Zündſchnur war abgebrannt. Im Apollotheater explodierte während des Zwiſchenaktes eine Bombe, die glücklicherweiſe keine Opfer forderte. Es entſtand eine Panik, in deren Verlauf zahlreiche Perſonen Quetſchungen davontrugen. Bis zur Wiederherſtellung der normalen Lage wurden ſämtliche Schauſpielhäuſer geſchloſſen.

impresso
Media
Monitoring
of the Past

Suitable for **MA** Thesis

Andrianos Michail, Simon Clematide

*Note: Come to us if you have another idea connected to keywords

Gerold Schneider & Janis Goldzycher

Categorizing Religious Online Hate Speech

- Context: Hate speech, often in relation to religion, is an omnipresent problem on the internet.
- Research question:
  - How can we create categories of religious hate in a data-driven manner?
  - These categories could be semantic (who is hated how?) or syntactic/stylistic (how is the hate expressed?).
- Methods and Skills:
  - training and using transformer-based supervised text classifiers for detecting hate speech
  - using unsupervised methods, such as clustering, for recognizing and classifying types of hate speech

Contact: `gschneid@cl.uzh.ch`and `goldzycher@cl.uzh.ch`

Thank you!

Nora Hollenstein & Lena Jäger

Generating character avatars from books

Can you create animated avatars from character descriptions in books?

Details:
The goal is to generate video files showing the characters of fiction books based on the text extracted from EPUB files from ebooks without copyright.

The challenge of this project is to generate avatars that are consistent with the descriptions in the book *and* consistent across the storyline, and to make the generation of the animated images fast and efficient.

Supervision: Nora Hollenstein

Contact: nora.hollenstein@uzh.ch

Developing an eye-tracking based text readability score

Say goodbye to Flesch and other traditional readability indices

Say hello to a new readability score based on synthesized eye-tracking data

Details: This project builds on the premise that most readability scores are outdated as they do not work on modern text genres and are often language-specific. Eye-tracking data provides information about the linguistic and cognitive processes occurring during text comprehension and can therefore be used as a proxy to determine the readability level of a text. Moreover, we now have models that can accurately predict eye-tracking features for reading, making the need for real-time data obsolete. The goal of this project is to develop a multilingual text readability score based on synthesized eye-tracking data. This score will then be compared to traditional scores and can be evaluated against datasets of standardized language assessment tests or text simplification.

Supervision: Nora Hollenstein & Lena Jäger

Contact: nora.hollenstein@uzh.ch

Webcam eye-tracking for machine learning applications

Collect webcam eye-tracking data in an online experiment to improve fixation detection and gaze prediction algorithms

Details: Webcam-based eye-tracking technology has improved in recent years but has not yet been thoroughly tested for reading experiments. The aim of this project is to extend the WebQAmGaze dataset by collecting a control set in the lab, representing the upper bound of achievable data quality from webcams. The data is then used to improve fixation detection, fixation correction and/or gaze prediction algorithms to ensure webcam gaze data can be used in machine learning applications.

Supervision: Nora Hollenstein & Lena Jäger

Contact: nora.hollenstein@uzh.ch