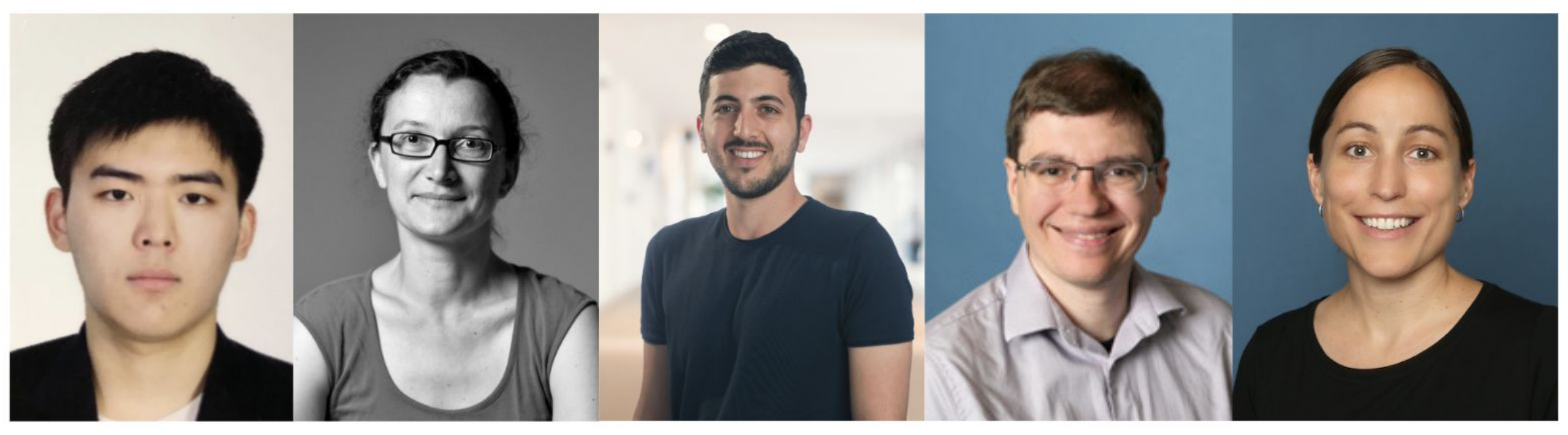
SwissSLi: the Multi-parallel Sign Language Corpus for Switzerland



LREC-COLING 2024

Zifan Jiang (jiang@cl.uzh.ch), Anne Göhring, Amit Moryossef, Rico Sennrich, Sarah Ebling

Department of Computational Linguistics, University of Zurich

Motivation

> Include sign languages in natural language processing (Yin et al., 2021), as they **are** fully natural languages.

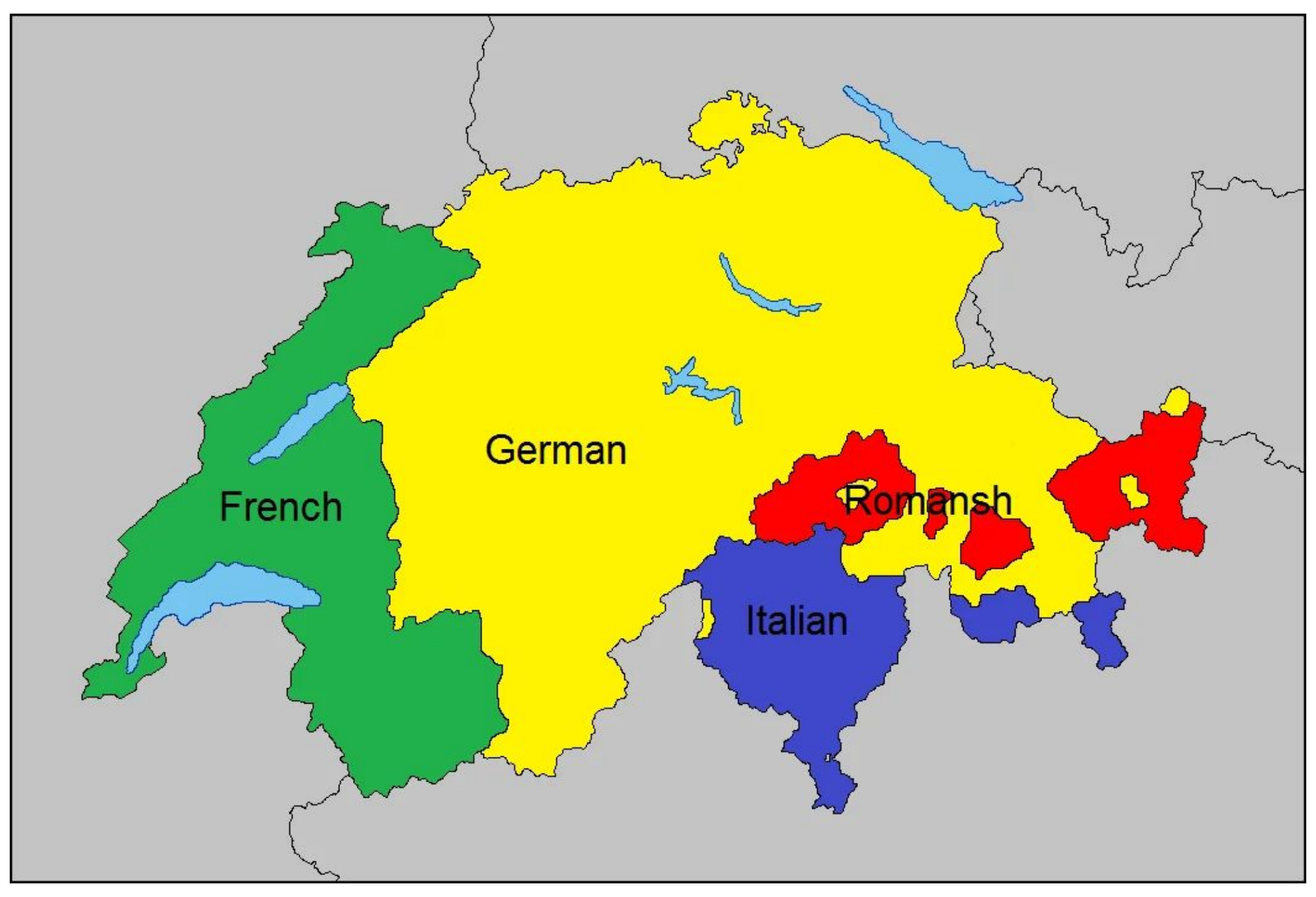
> Limited data is a main challenge, as shown by *WMT-SLT* (Müller et al., 2022, 2023).

> The multilingual landscape in Switzerland offers the possibility for a **multi-parallel** sign language corpus.

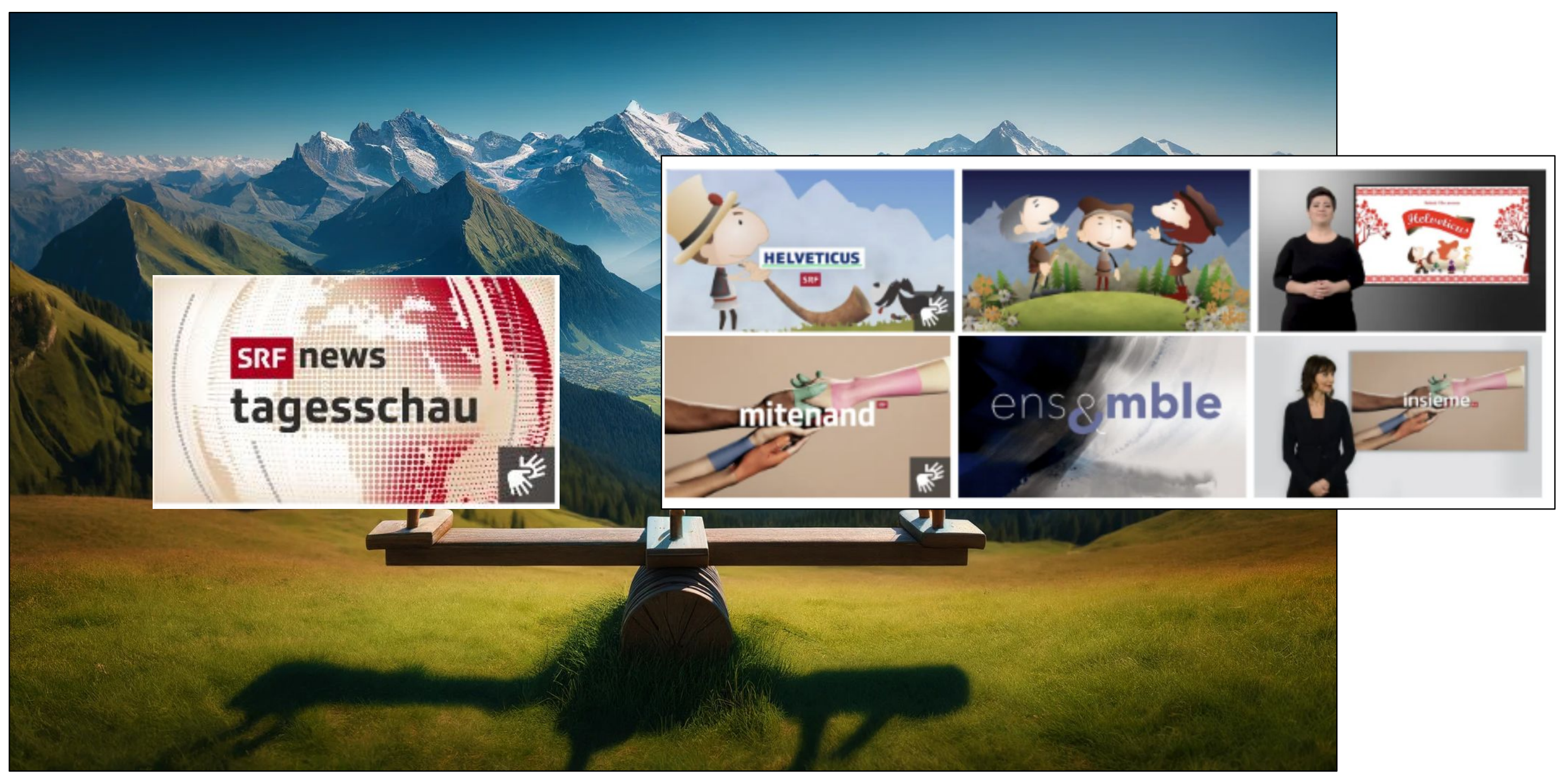
> We focus on TV programs **translated by deaf signers offline**, in total ~30 hours in 3 sign and 3 spoken languages.

Background & Discussion

1. Languages in Switzerland



2. Taxonomy of Swiss TV Programs



3. Signing Mode and Data Quality

Sign language as source

TV interpretation

Offline translation

DGS-KORPUS

BOBSL DATASET

How2sign

4. Data Quality, Quantity, and Licensing

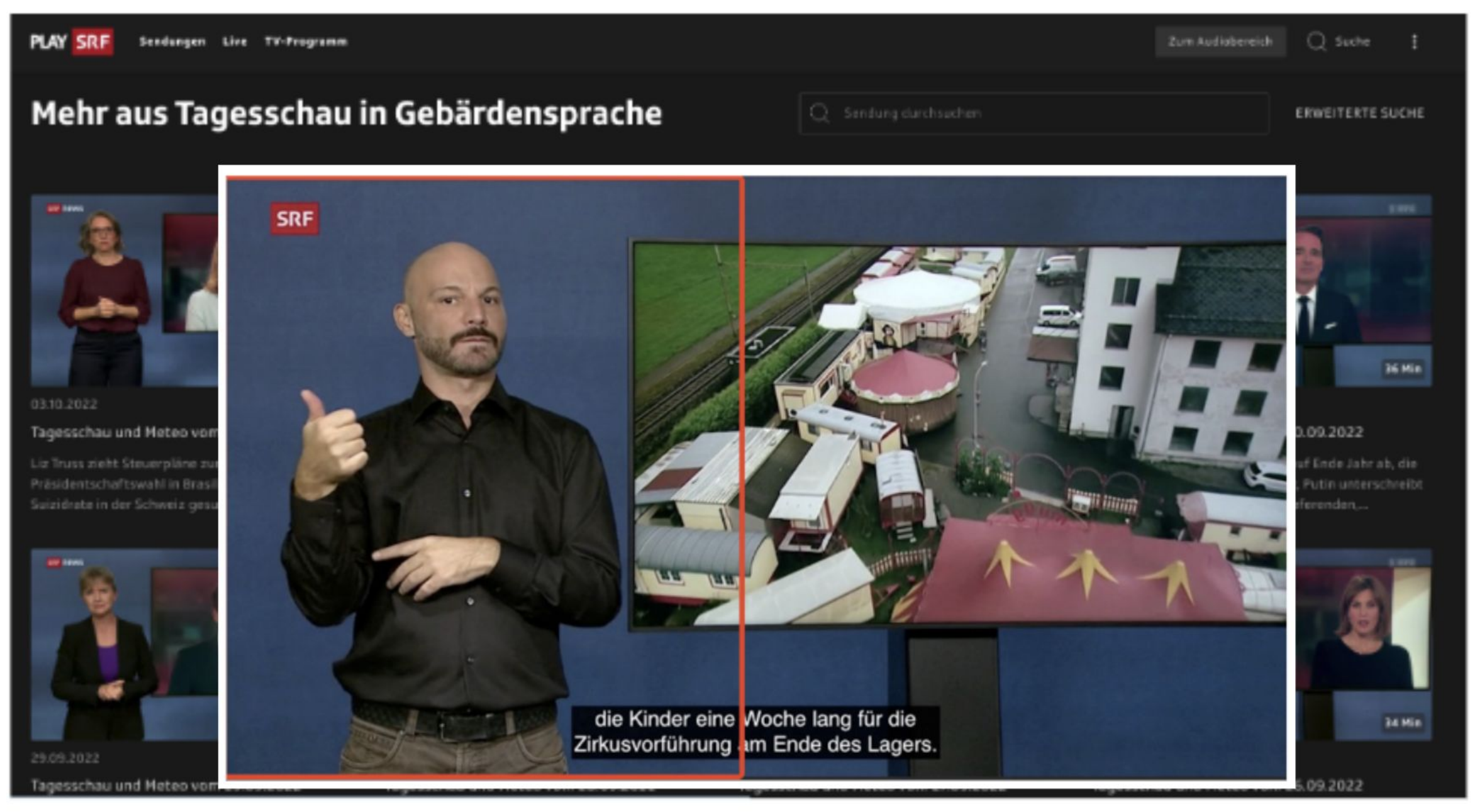
Deaf community

Policy maker

Computer scientists

Data & Processing

1. Raw Data from Program Websites


PLAY SRF
Sendungen
Live
TV-Programm
Mehr aus Tagesschau in Gebärdensprache
SRF
die Kinder eine Woche lang für die Zirkusvorführung am Ende des Lagers.

2. Video/Pose Processing

- Pose estimation

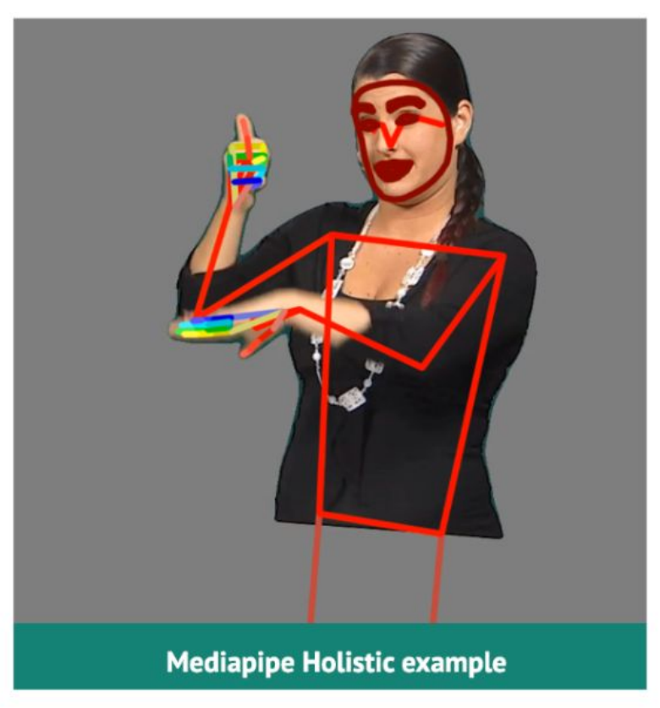
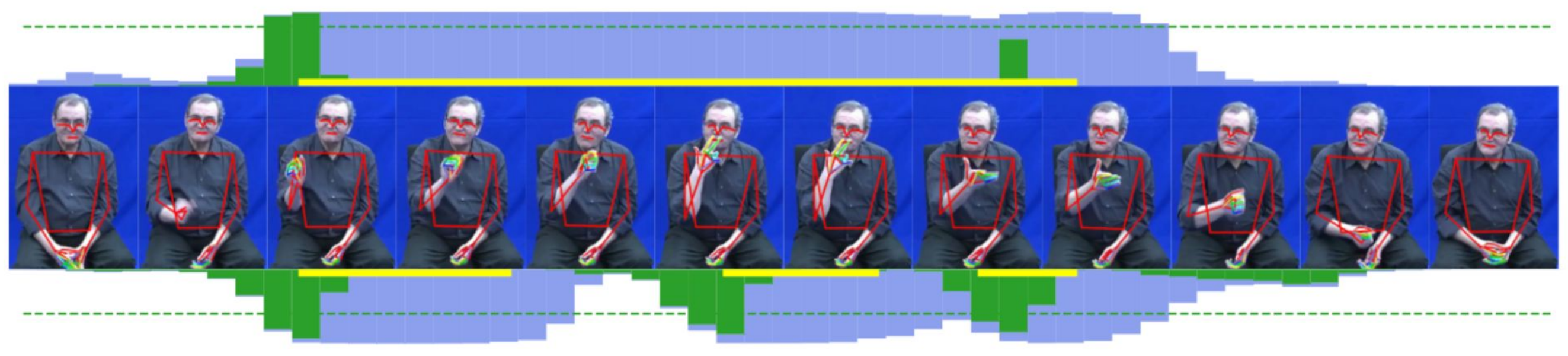


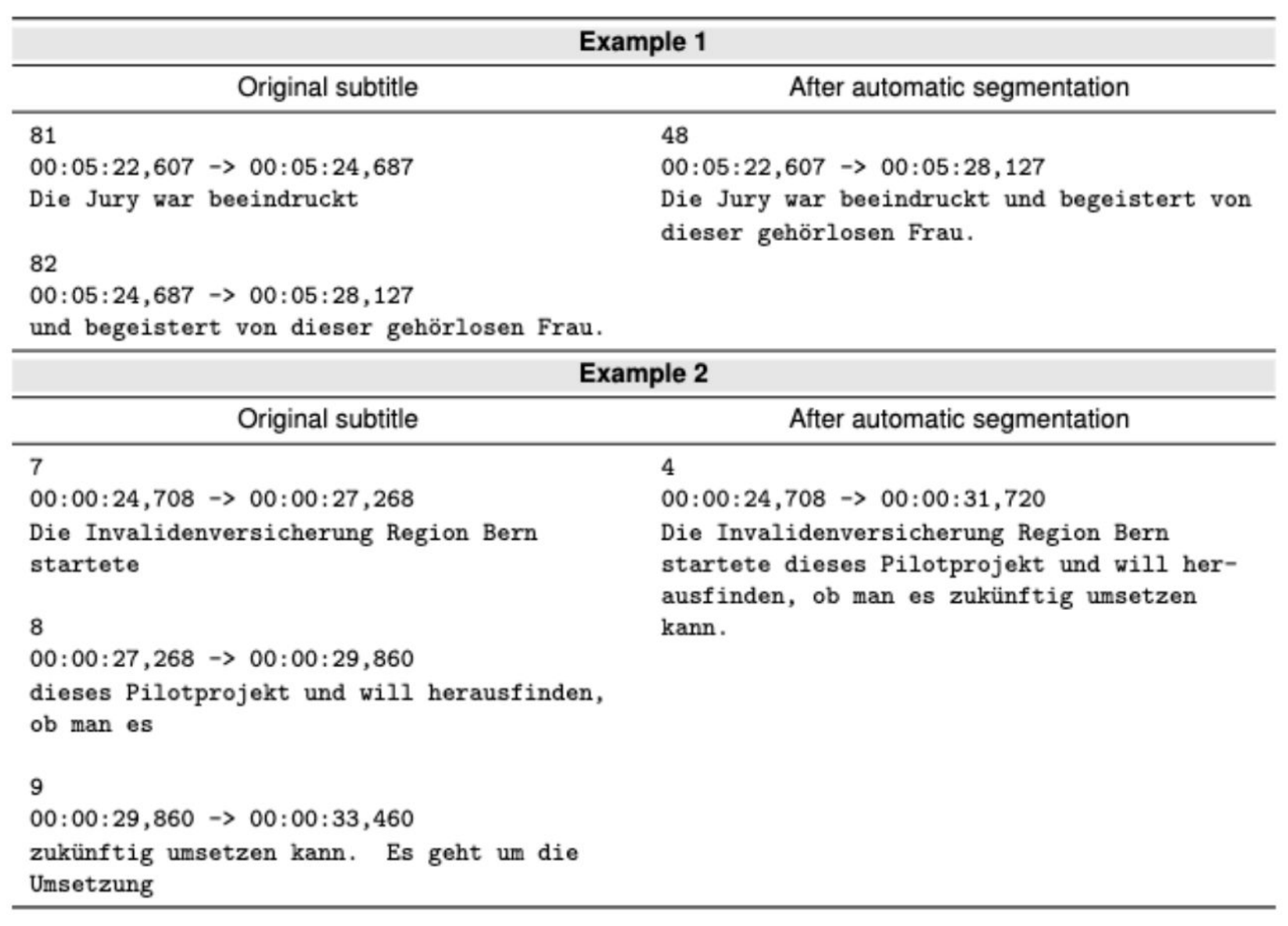
Figure taken from https://www.wmt-slt.com/data

- Video segmentation

Figure taken from Moryossef et al. (2023)



3. Subtitle Processing



Original subtitle	After automatic segmentation
Example 1	
81 00:05:22,607 -> 00:05:24,687 Die Jury war beeindruckt 82 00:05:24,687 -> 00:05:28,127 und begeistert von dieser gehörlosen Frau.	48 00:05:22,607 -> 00:05:28,127 Die Jury war beeindruckt und begeistert von dieser gehörlosen Frau.
Example 2	
7 00:00:24,708 -> 00:00:27,268 Die Invalidenversicherung Region Bern startete 8 00:00:27,268 -> 00:00:29,860 dieses Pilotprojekt und will herausfinden, ob man es 9 00:00:29,860 -> 00:00:33,460 zukünftig umsetzen kann. Es geht um die Umsetzung	4 00:00:24,708 -> 00:00:31,720 Die Invalidenversicherung Region Bern startete dieses Pilotprojekt und will herausfinden, ob man es zukünftig umsetzen kann.

Figure taken from the WMT-SLT findings paper

Conclusion & Outlook

> The corpus is publicly available on the **SWISSUbase** data platform for research purposes under *CC BY-NC-SA 4.0*.

> Shared with the **informed consent** of the signers involved and a data-sharing agreement with the *Swiss Broadcasting Corporation*.

> A valuable asset for researchers in computer vision, natural language processing, and sign language linguistics.

SWISSUbase

Flagship
supported by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Innosuisse – Swiss Innovation Agency

University of Zurich UZH