

# SDS-200: A Swiss German Speech to Standard German Text Corpus

Michel Plüss, University of Applied Sciences and Arts Northwestern  
Switzerland, Windisch

LREC 2022

Full author list:

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, Manfred Vogel

# Swiss German

- Family of German dialects
- 5 million speakers
- Differs from Standard German in phonetics, vocabulary, morphology, syntax
- Mostly spoken
- No standardized writing system

# Challenges Swiss German STT

- Many written variants for the same word
- Huge vocabulary size
- Desired language of output: Standard German → Translation
- Diversity of dialects, especially vocabulary and phonetics
- Limited amount of publicly available training data:
  - SwissDial [1], 26 h
  - Radio Rottu [2], 2 h
  - Swiss Parliaments Corpus [3], 293 h (noisy)

# Data collection

- Crowdsourcing
- Web recording tool <https://dialektsammlung.ch/>
- Based on Common Voice [4]
- Recording:
  - Standard German sentence
  - How would I say this in Swiss German?
  - Record in Swiss German
- Validation:
  - Is the recording in Swiss German?
  - Is it a faithful Swiss German translation of the Standard German sentence?

# Data collection - recording

The screenshot shows a mobile application interface for a language learning exercise. At the top left, there is a back arrow and two tabs: 'Sprechen' (highlighted) and 'Prüfen'. In the top right corner, it says '1/5 Aufzeichnungen'. The main content area is a large white box containing the text: 'Die übrigen Beach Boys gingen auf Tournee.' To the right of this box is a vertical list of five numbered circles (1-5), with the first one highlighted. Below the main text box, there is a line of instructional text in German: 'Überlegen Sie sich, wie Sie den Satz in Ihrem schweizerdeutschen Dialekt formulieren würden. Klicken Sie dann auf das Mikrofon-Symbol unten und sprechen Sie den Satz in Ihrer Formulierung.' At the bottom of the screen, there is a microphone icon with a red recording indicator. Below the microphone are several buttons: 'Tastenkürzel' (with a keyboard icon), 'Melden' (with a flag icon), 'Überspringen >>', 'Ohne Speichern weiter', and 'ABSENDEN'.

# Data collection - validation

The screenshot shows a mobile application interface for audio validation. At the top left, there is a back arrow and two tabs: 'Sprechen' and 'Prüfen', with 'Prüfen' being the active tab. In the top right corner, it says '1/5 Aufzeichnungen'. Below this is a vertical list of five numbered circles (1-5), with the first circle containing a speaker icon. The main content area is a large white box with the text 'Einen Ausblick wage ich nicht.' Below this box is a small instruction in German: 'Klicken Sie auf das Play-Symbol unten und beurteilen Sie die Aufnahme, ob sie Schweizerdeutsch ist und den hochdeutschen Satz korrekt wiedergibt.' At the bottom of the screen, there are two buttons: 'KORREKT' with a thumbs-up icon and 'FALSCH' with a thumbs-down icon, separated by a play button icon. Further down, there are three more buttons: 'Tastenkürzel' with a keyboard icon, 'Melden' with a flag icon, and 'Überspringen >>'.

# Data collection - metadata

- Dialect:
  - Zip code of origin of the dialect
  - E.g. where the participant went to school
  - Allows to group dialects in dialect regions, cantons, municipalities
- Age
- Gender

# Data collection - sentences

- Source:
  - 80 % from Swiss newspapers
  - 20 % from the German Common Voice [5]
- 5-12 tokens



# Data collection process

- Interviews and reports in Swiss television, radio, newspapers
- Support videos by 4 well-known Swiss comedians, shared on their social media accounts
- Leaderboard contest:
  - Individual contest
  - Score based on number of recordings, validations, quality of recordings
  - Attractive Switzerland-themed prizes for the top 10 participants
- Clash of Cantons:
  - Canton contest
  - Spark competition between cantons
  - Score based on number of recordings, weighted by average quality, normalized by population

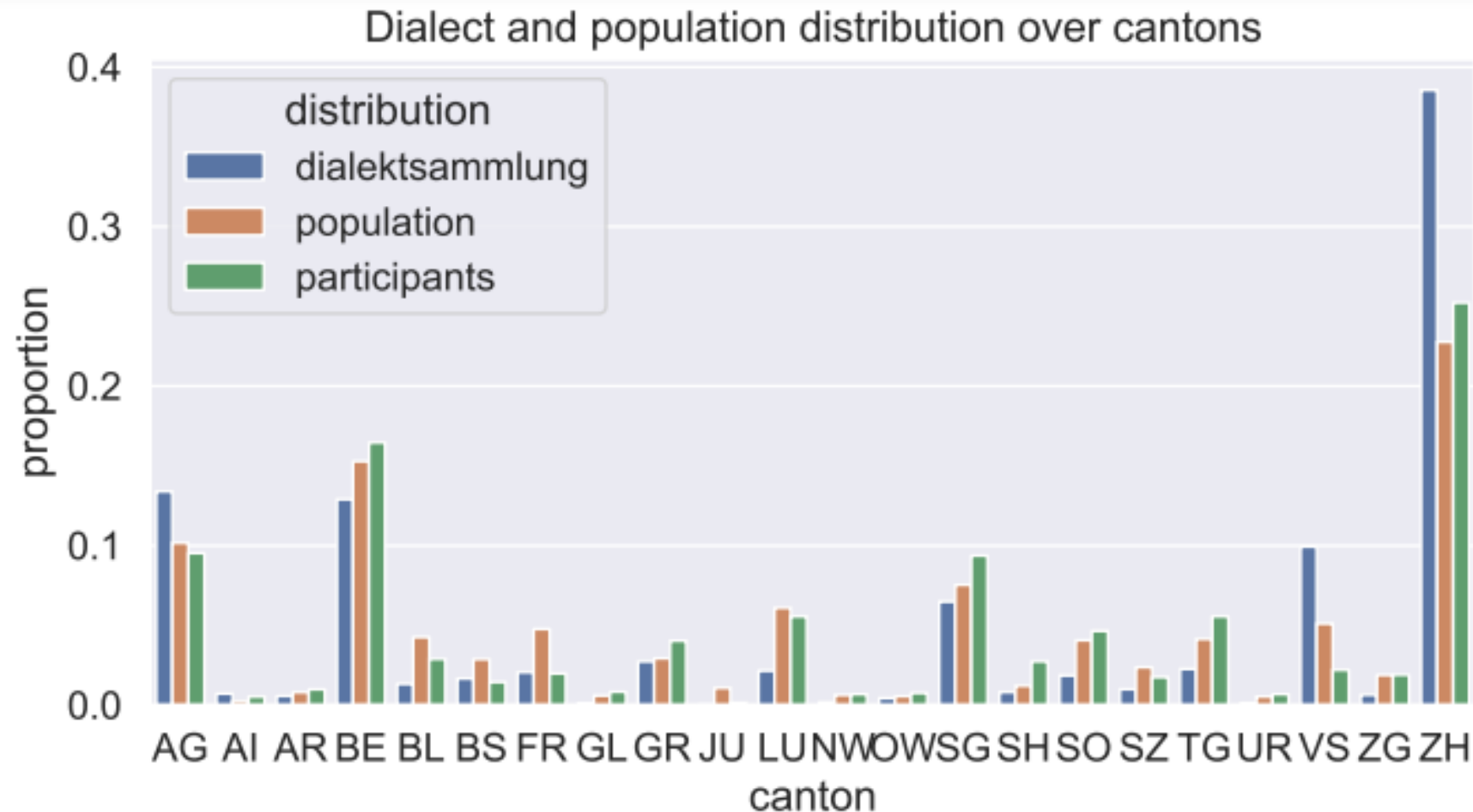
# Corpus

Split	Hours
train (raw)	188.9
train (filtered)	178.3
validation	5.2
test	5.4

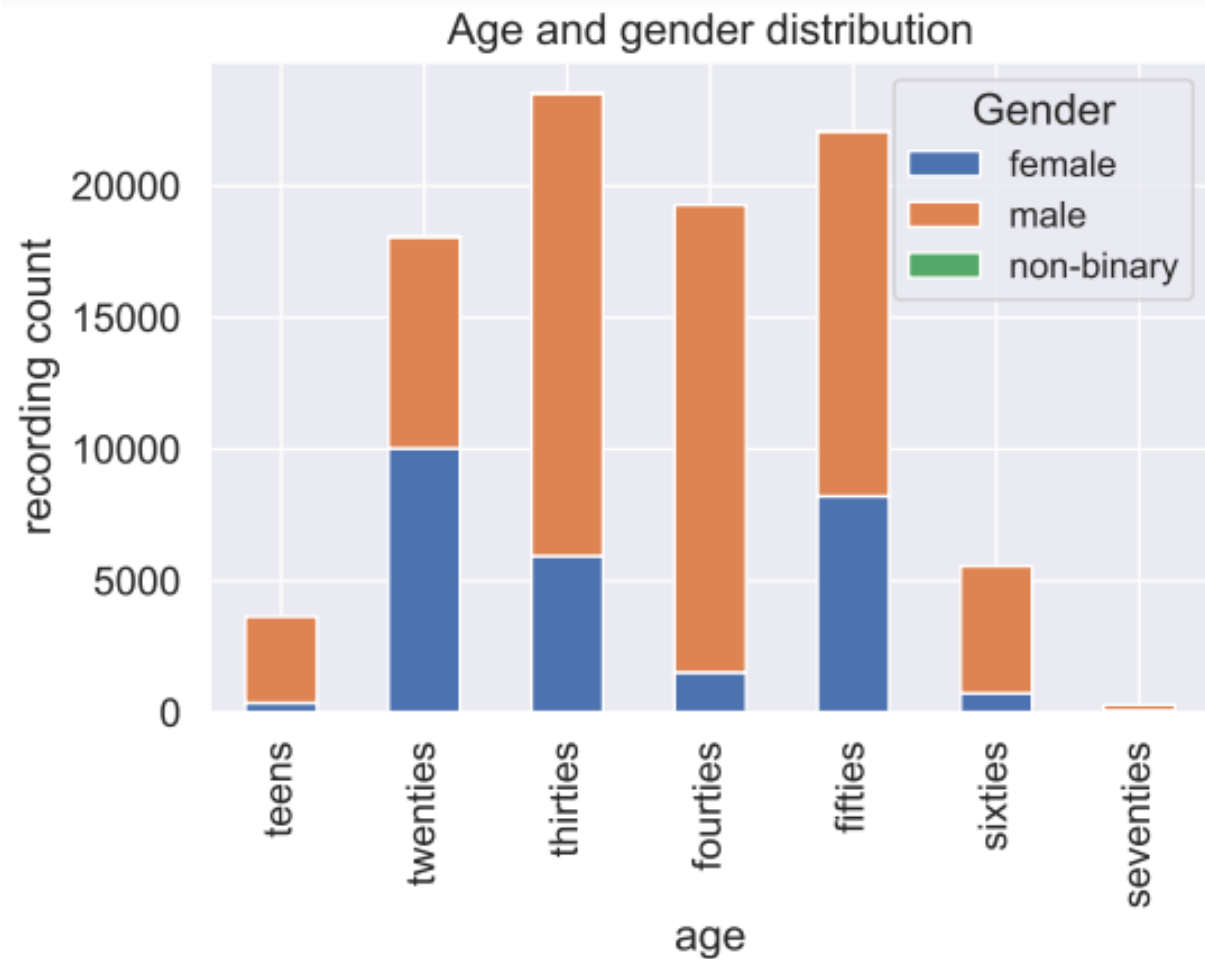
# Corpus

- train (filtered), validation, test:
  - Filtered using validations
  - 142'545 recordings
  - 3'816 speakers
  - Recording duration: 4.8 s +/- 1.3 s

# Corpus - dialects



# Corpus – age and gender



- Speakers:
  - 8 % male
  - 6 % female
  - 86 % unknown
- Recordings:
  - 46 % male
  - 19 % female
  - 35 % unknown

# Models

- All models are trained in an end-to-end fashion (Swiss German speech to Standard German text)
- *SDS-200 train (filtered)* is the only labeled corpus used for these models
- WER and BLEU reported on *SDS-200 test*

# Models

Model	Params	Train method	WER	BLEU
Transformer	72M	from scratch	30.3	53.1
XLS-R 1B	965M	finetuning	21.6	64.0
XLS-R 1B + LM	965M	finetuning + LM	<b>17.9</b>	<b>70.3</b>

- XLS-R [6]:
  - wav2vec 2.0 [7] models
  - Self-supervised pre-training of speech representations
  - 436'000 h unlabeled speech data in 128 languages (no Swiss German)
- LM:
  - KenLM [8]
  - 68M Standard German sentences

# References

- [1] Dogan-Schönberger, P., Mäder, J., and Hofmann, T. (2021). SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German.
- [2] Garner, P. N., Imseng, D., and Meyer, T. (2014). Automatic Speech Recognition and Translation of a Swiss German Dialect: Walliserdeutsch. In Proceedings of Interspeech 2014.
- [3] Plüss, M., Neukom, L., Scheller, C., and Vogel, M. (2021). Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. In Proceedings of the Swiss Text Analytics Conference 2021.
- [4] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common Voice: A Massively-Multilingual Speech Corpus. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4211–4215.
- [5] <https://github.com/common-voice/common-voice/tree/main/server/data/de>
- [6] Babu, A., Wang, C., Tjandra, A., Lakhota, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. arXiv, abs/2111.09296.
- [7] Baevski, A., Zhou, Y., Mohamed A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems, 2020.
- [8] Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation.