How We Crowdsourced a Large Swiss German Speech-to-Text Dataset: **Die Schweizer Dialektsammlung**

Manuela Hürlimann, Michel Plüss, Marc Cuny, Alla Stöckli,
Anna Ulasik, Manfred Vogel, Mark Cieliebak

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

n|w Fachhochschule
Nordwestschweiz

SPINNINGBYTES
Text Analytics & Speech
Processing

Outline

- Motivation + Goals
- Data Collection
  - Web App
  - PR Campaign
- Data collected so far
- Lessons learnt
- Questions + Discussion

Motivation + Goals

Speech-to-Text

Die Schweizer Dialekt sammlung

Motivation

Swiss German is lacking a large-scale public dataset for Speech-to-Text!

- Walliserdeutsch, 7h
- ArchiMob (UZH), 70h
- Swiss Parliaments corpus (FHNW), 300h

- *State-of-the-art systems need **thousands of hours**!*

What?

- **Crowdsource** audio recordings from volunteers with an easy-to-use web application

- **Goal**: collect **2000h** of audio

  ➡ reach at least **24’000 people** in all German-speaking cantons

Data collection

Data Collection: Web App

www.dialektsammlung.ch

Input Data

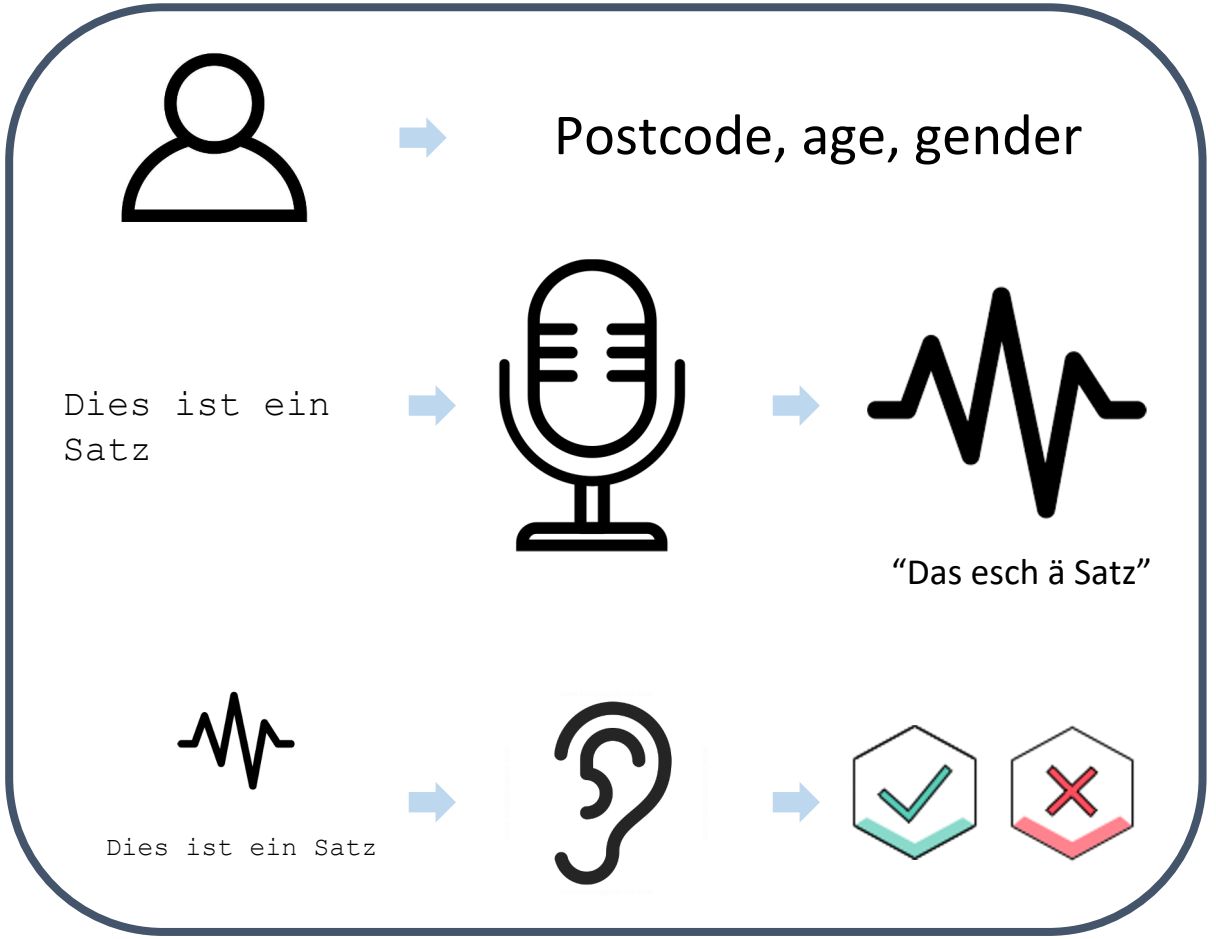
CommonVoice DE

Subtitles

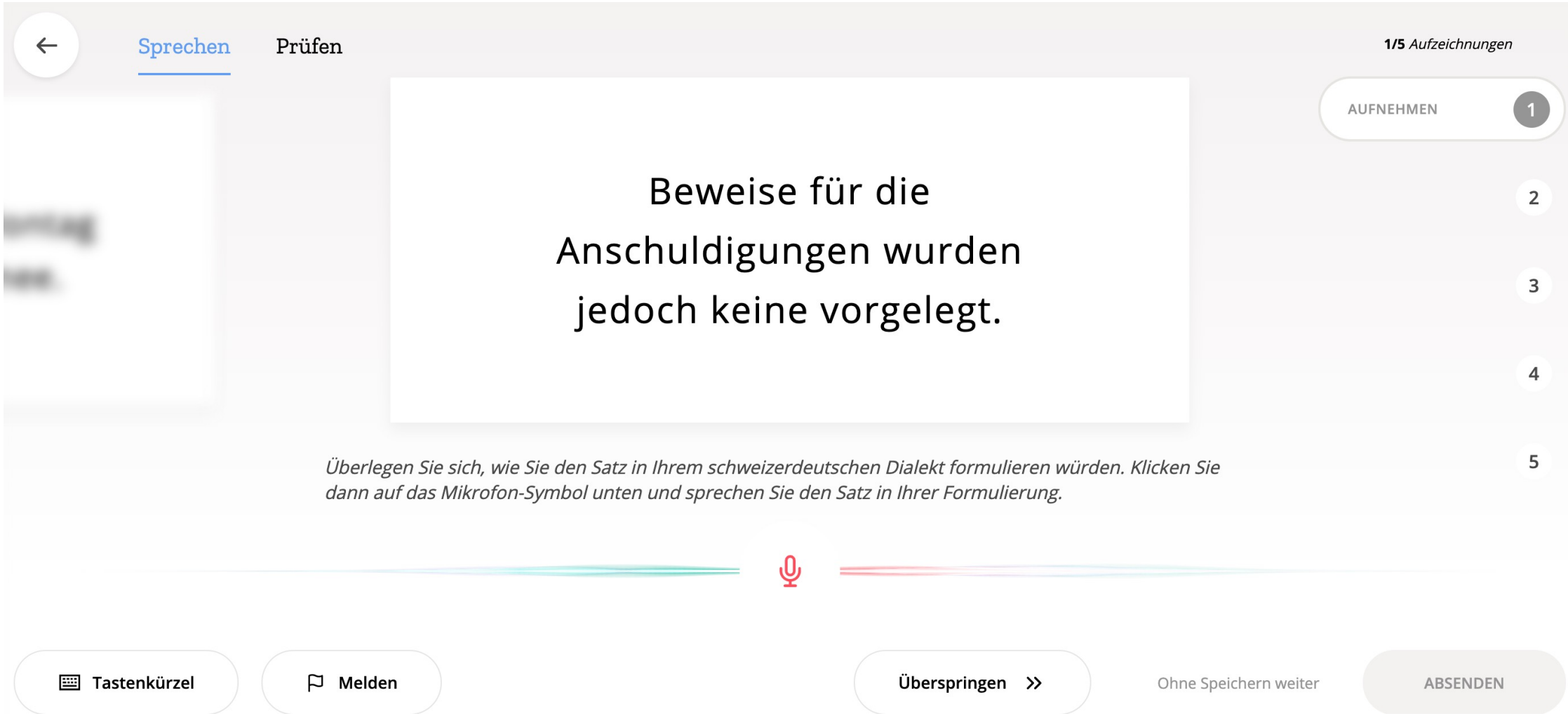
Sentence selection

- Easy to translate into dialect
  (max. 12 words)
- Diverse vocabulary
  (not too specialised)

Web application

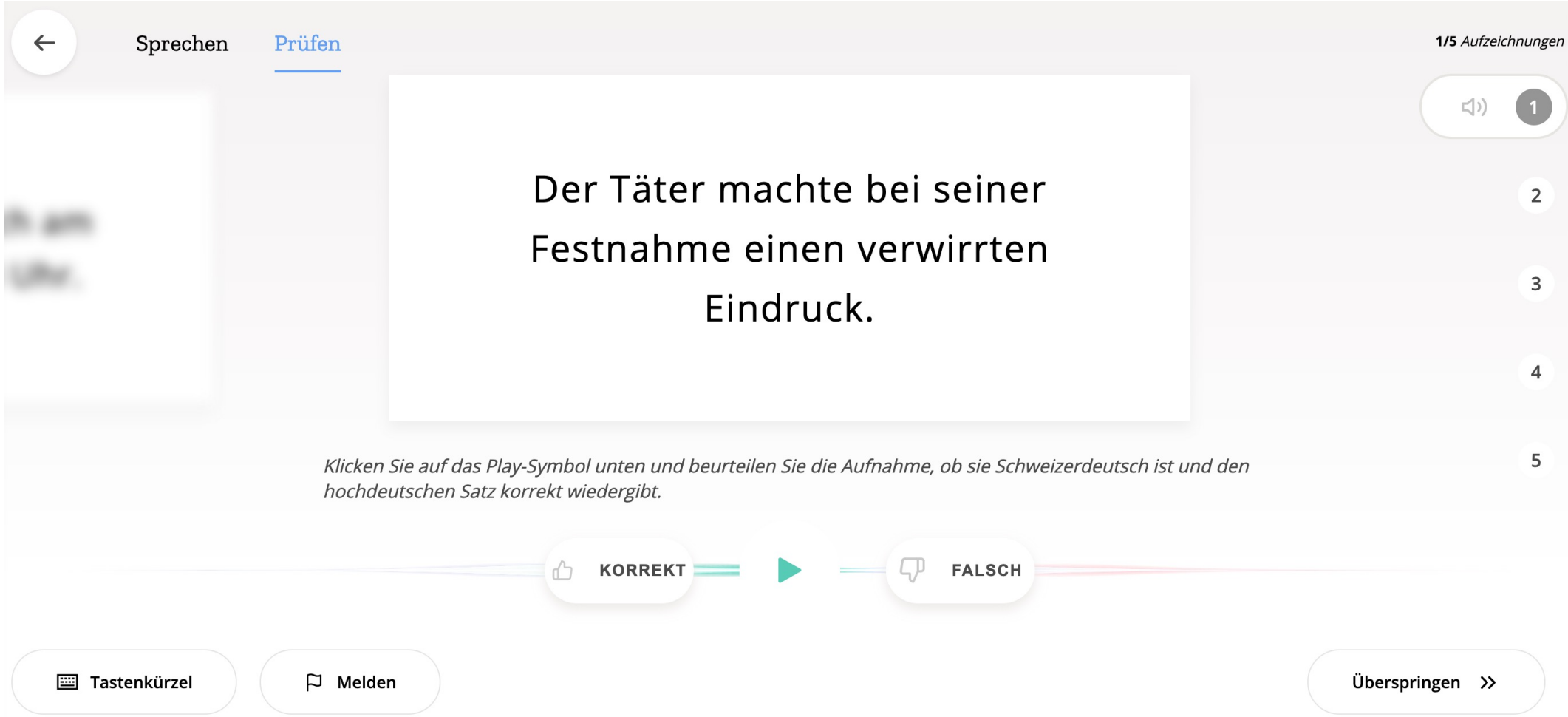


Web App: Recording



Die Schweizer Dialekt sammlung

Web App: Validation



Die Schweizer Dialekt sammlung

PR Campaign

Computer mit Thurgauer Wörter füttern

03.06.2021 07:00

ZHAW und FHNW digitalisieren Schweizer Dialekte

Zwei Hochschulen wollen die Dialekte der Schweiz digitalisieren, um künstlicher Intelligenz Schweizerdeutsch beizubringen. Die Bevölkerung kann sich am Projekt beteiligen.

blue News UEFA Champions League

News 24h-Ticker Schweiz Regional International Vermischtes

Siri, Alexa und Co. sollen endlich Schweizerdeutsch lernen

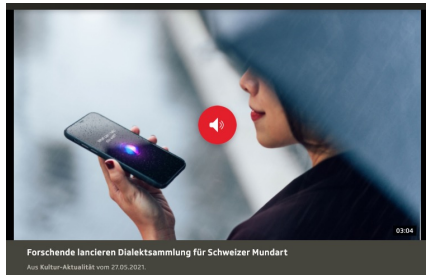
neo1 MEIN RADIO

News Programm Gewinnen Werbung Unternehmen Coronavirus Info

News » News

Können Siri & Alexa bald Dialekt? 26.05.2021 10:27

Kultur Gesellschaft



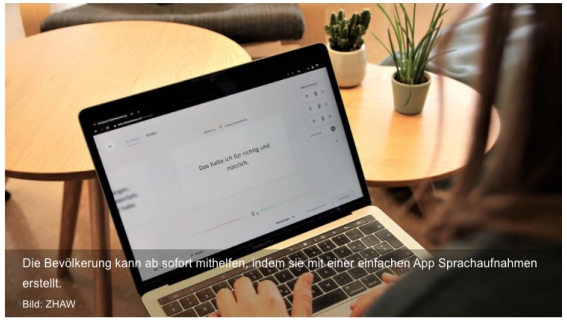
Simon Enzler - offiziöll
3 June at 16:45 ·

Übersetzig: Wir brauchen nun wirklich jeden Dialekt in dieser neuen Schweizer Dialektsammlung. Ja gut, vor allem natürlich den Appenzeller Dialekt. Ganz ehrlich: Die wenigsten können ihn (sprechen) und noch weniger verstehen ihn.
😉

Spende auch du deinen Dialekt: www.dialektsammlung.ch
ZHAW School of Engineering

ZHAW und FHNW digitalisieren Schweizer Dialekte

25. MAI 2021



Suchen südostschweiz Login

Im Fokus News ohne Corona Meistgeklickt Regional Nachrichten > Wirtschaft Sp

Forschende lancieren Dialektsammlung für Schweizer Mundart

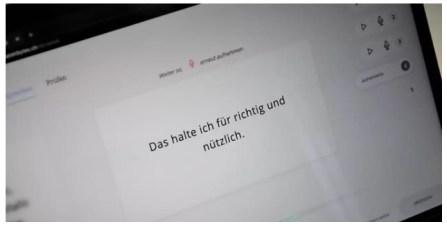
Ein Projekt von Schweizer Hochschulen sammelt und digitalisiert Deutschschweizer Dialekte. Das Ziel ist, eine Grundlage zu schaffen, um etwa Chatbots und Sprachassistenzen Mundart beizubringen.

Nau.ch Suchwort oder Ortschaft eingeben

Home > News > Schweiz

Forschende lancieren Dialektsammlung für Schweizer Mundart

Ein Projekt von Schweizer Hochschulen sammelt und digitalisiert Deutschschweizer Dialekte. Das Ziel ist ein Programm für die Übersetzung von Mundart.



PCtipp

Forschende sammeln Schweizer Dialekte fürs Computerprogramm-Training

Die beiden Hochschulen rufen zu einer Datensammlung von Schweizer Dialekten in der Deutschschweiz auf. Mit den digitalisierten Dialekten sollen Computerprogramme trainiert werden.

In the summer:
“Schoggitaler-Battle” +“Clash of the Cantons”

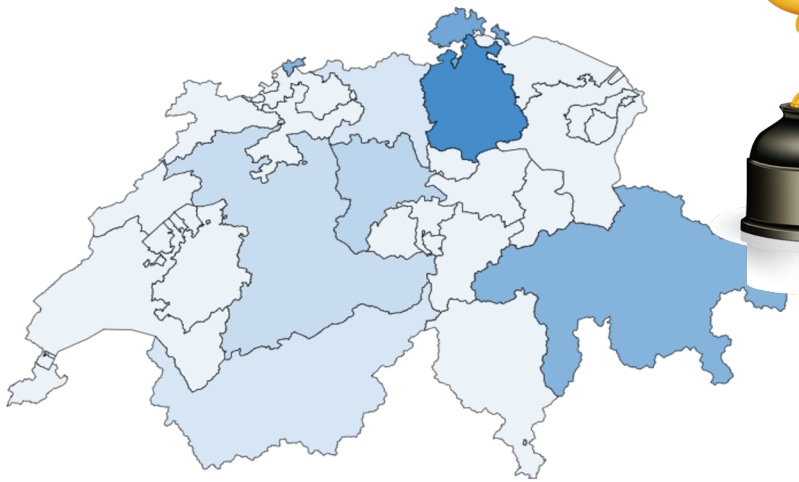
Rangliste

SCHOGGITALER

Ihr Rang: 1 von 22

01	Julia (Sie)	103
02	Hans R.	40
03	Sven	28
04	Martina	23
05	Ueli	18

Die Grafik zeigt die Kantone mit den meisten Punkten pro Einwohner. Hilf Deinem
indem du fleissig aufnimmst, Übersetzungen generierst und Aufnahmen prüfst.



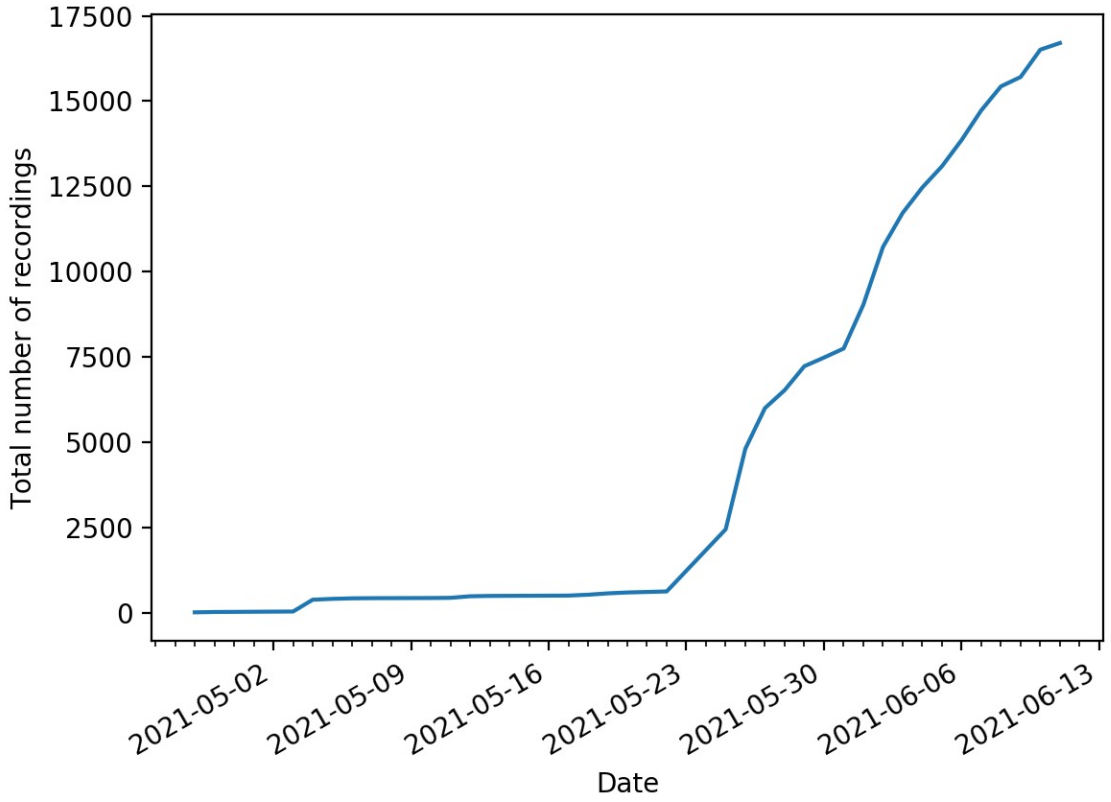
am pf der Kantone

		Punkte	Aufnahmen
01	Jura	606	5
02	Basel-Stadt	41	10
03	Aargau	37	30
04	Appenzell Innerrhoden	0	0
05	Appenzell Ausserrhoden	0	0
06	Bern	0	0
07	Basel-Landschaft	0	0
08	Freiburg	0	0

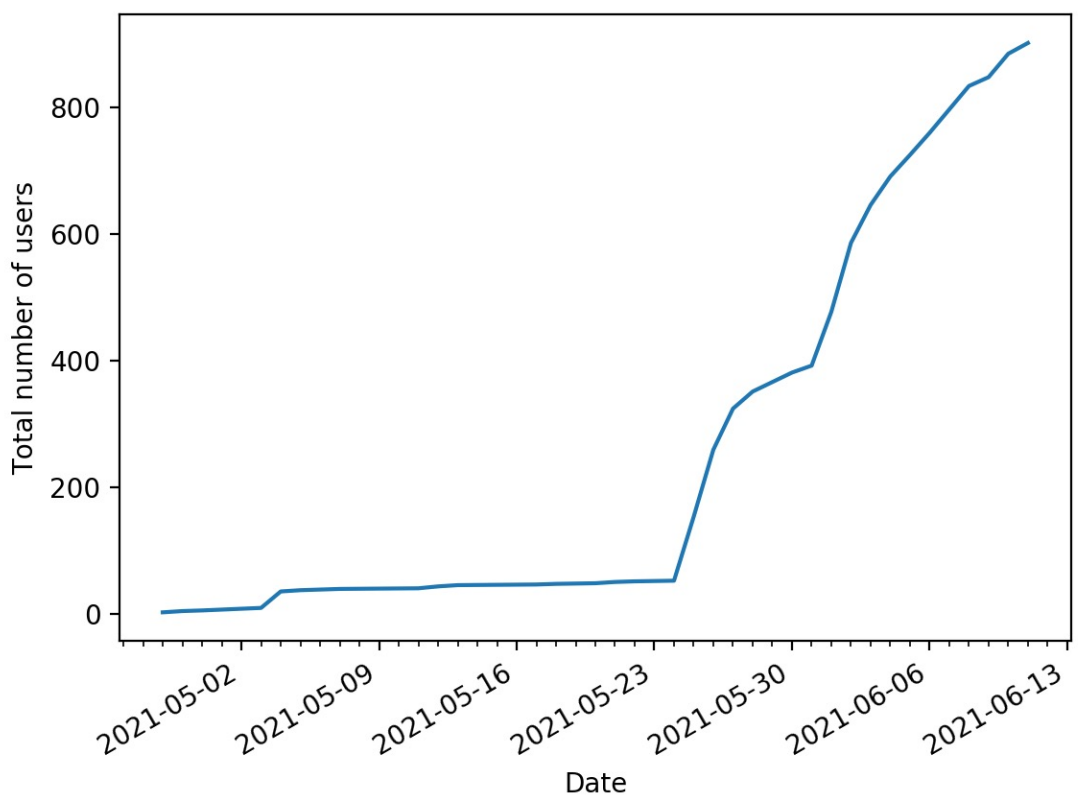
Data collected so far

Data collected so far

17’318 Recordings



929 Users



Contribute now!

www.dialektsammlung.ch

Everyone can participate! It'll help us a lot if you take **5 minutes** to create recordings ☺

Die Schweizer Dialekt sammlung

Lessons learnt

Lessons Learnt

- No-one wanted to support the data collection, but now that we are doing it there is **huge interest in the data**
- Building, testing and running a **web application** for thousands of users takes a LOT of time
- It is better to use **several different ways to reach our audience**, rather than rely on one channel
- **“Teaching Siri Swiss German”** is a simple and effective pitch
- Many people **care about dialects**, but this does not automatically translate to a large number of recordings

Questions?