

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



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



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



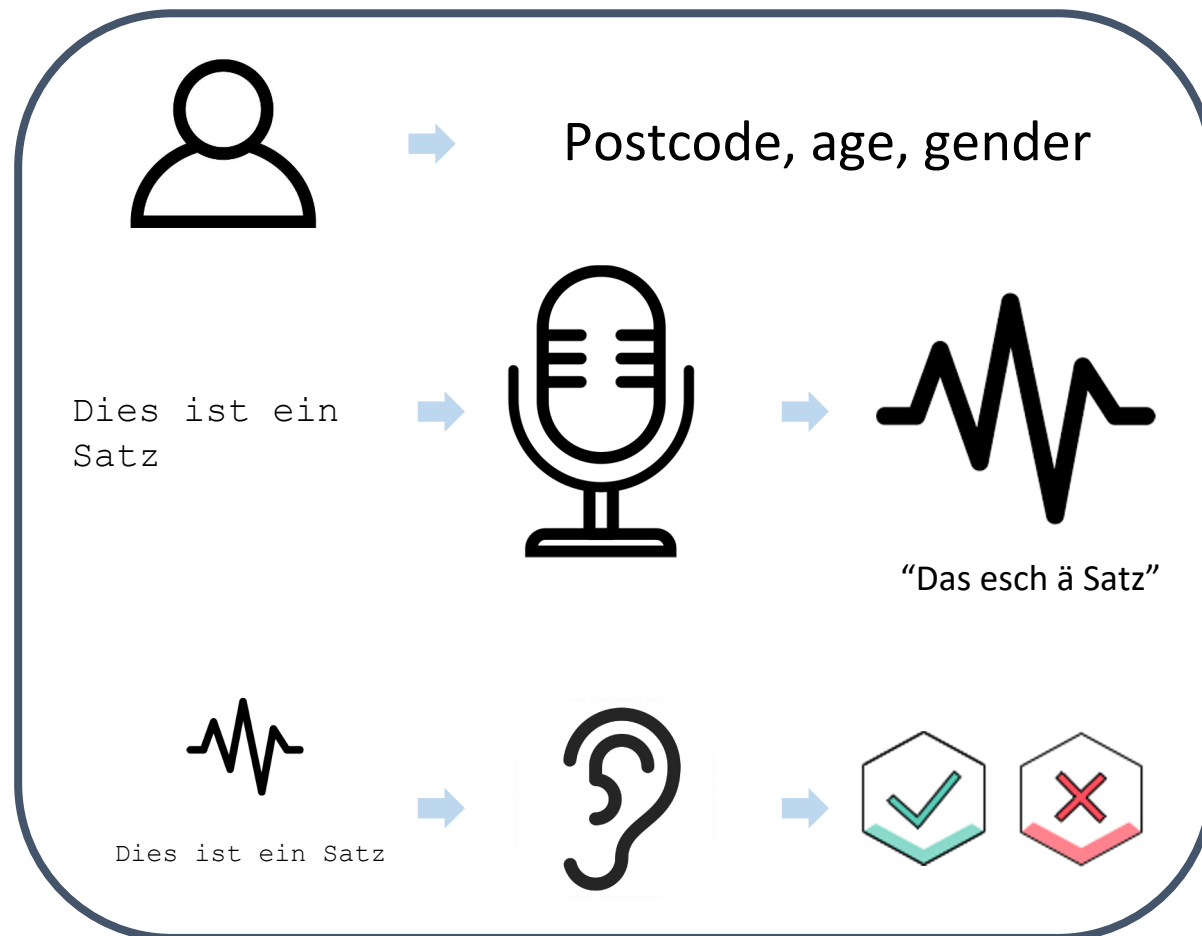
CommonVoice DE

Subtitles

Sentence selection

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

Web application



Web App: Recording

The screenshot shows a web application interface for recording a sentence. 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 contains a large white box with the text: "Beweise für die Anschuldigungen wurden jedoch keine vorgelegt." To the right of this box is a vertical list of five numbered circles (1-5), with the first one containing the word "AUFNEHMEN". Below the main text box, there is a smaller text box with the instruction: "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 center, there is a red microphone icon. The bottom navigation bar contains several buttons: "Tastenkürzel" (with a keyboard icon), "Melden" (with a flag icon), "Überspringen >>" (with a double arrow icon), "Ohne Speichern weiter", and "ABSENDEN".

Web App: Validation

← Sprechen Prüfen 1/5 Aufzeichnungen

Der Täter machte bei seiner Festnahme einen verwirrten Eindruck.

1
2
3
4
5

Klicken Sie auf das Play-Symbol unten und beurteilen Sie die Aufnahme, ob sie Schweizerdeutsch ist und den hochdeutschen Satz korrekt wiedergibt.

👍 KORREKT ▶️ ❌ FALSCH

Tastenkürzel Melden Überspringen >>

PR Campaign

Computer mit Thurgauer Wörter füttern

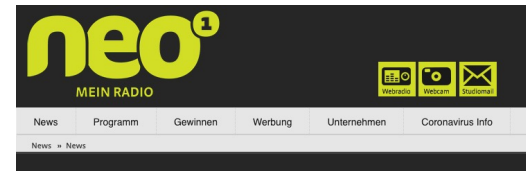
03.06.2021 07:00

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.



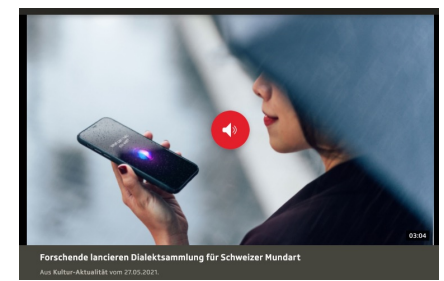
Siri, Alexa und Co. sollen endlich Schweizerdeutsch lernen



Können Siri & Alexa bald Dialekt?

26.05.2021 10:27

Kultur Gesellschaft



Forschende lancieren Dialektsammlung für Schweizer Mundart
Aus Kultur-Gesellschaft vom 27.05.2021

Simon Enzler - offiziell
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.

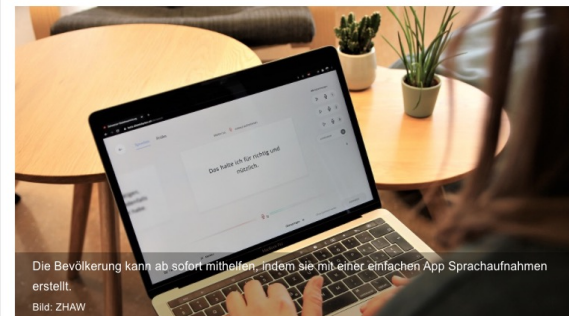


Spende auch du deinen Dialekt: www.dialektsammlung.ch
ZHAW School of Engineering



ZHAW und FHNW digitalisieren Schweizer Dialekte

26. MAI 2021



Die Bevölkerung kann ab sofort mithelfen, indem sie mit einer einfachen App Sprachaufnahmen erstellt.
Bild: ZHAW

Suchen



Login

Im Fokus News ohne Corona Meistgeklickt Regional Nachrichten > Wirtschaft Sp

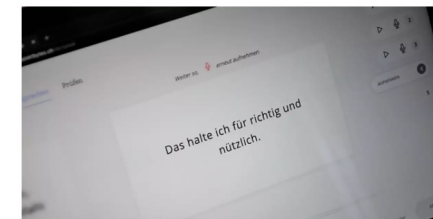
Forschende lancieren Dialektsammlung für Schweizer Mundart

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



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.



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”



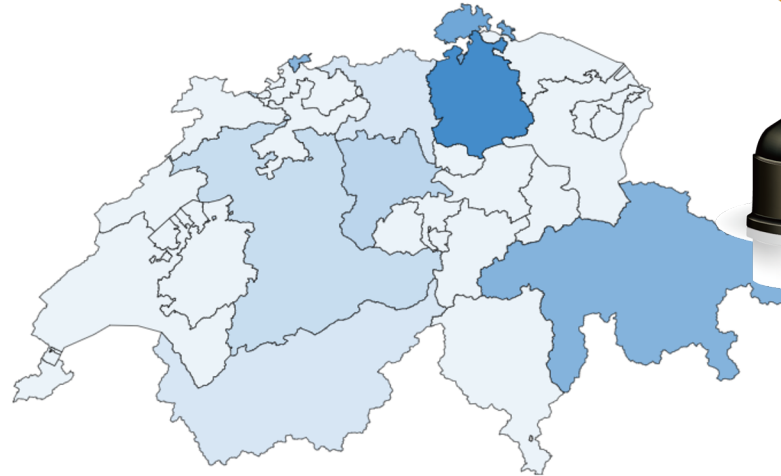
Rangliste

SCHOGGITALER

Ihr Rang: 1 von 22

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

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



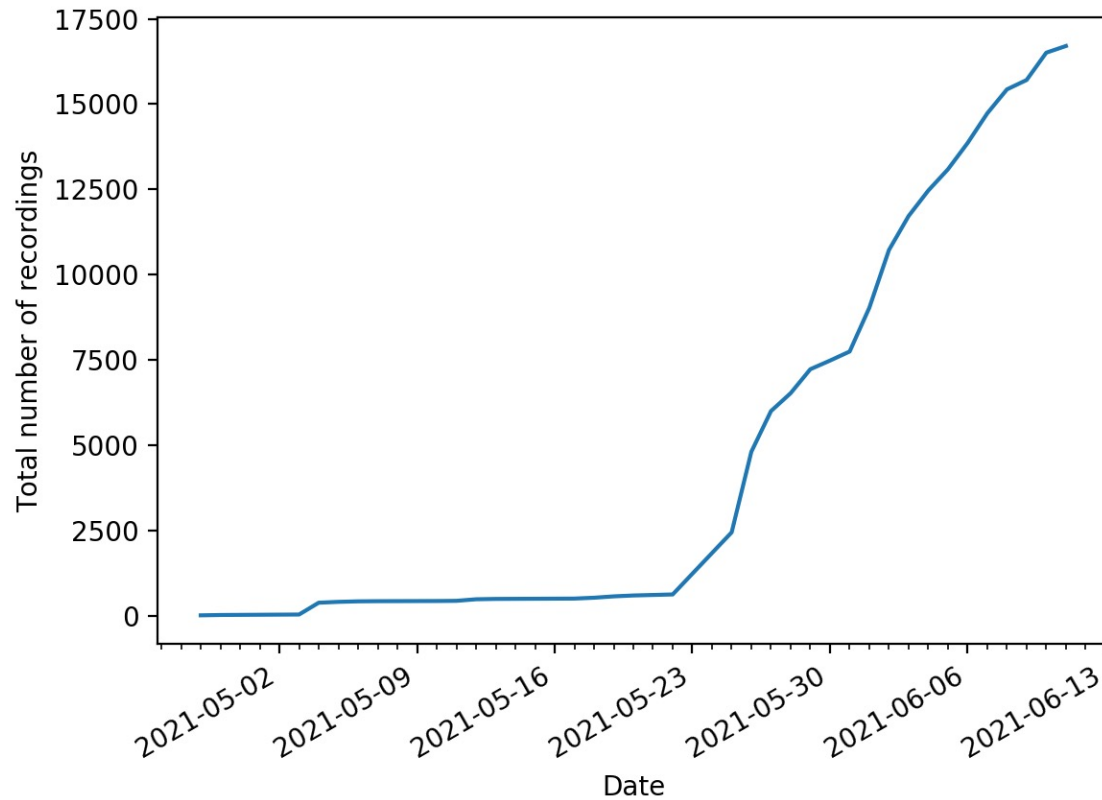
Kampf der Kantone

| | | 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 |

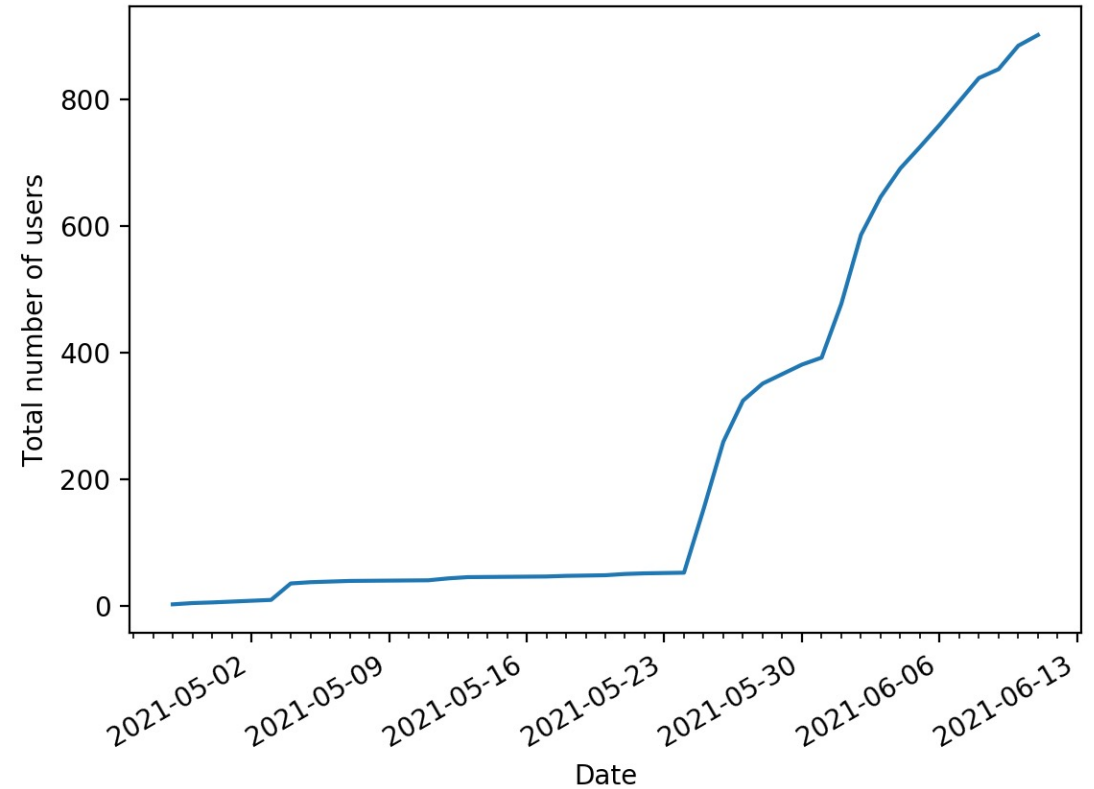
Data collected so far

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 😊

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?