



# Basic natural language processing for Swiss German texts

Tanja Samardžić





## Long-term contribution

*Noëmi Aepli*

*Fatima Stadler*

*Yves Scherrer*

*Elvira Glaser*

## Funding

Hasler Foundation grant No 16038

UZH URPP 'Language and Space'

Agreement with Spitch

## Specific tasks

*Henning Beywl*

*Christof Bless*

*Alexandra Bünzli*

*Matthias Friedli*

*Anne Göhring*

*Noemi Graf*

*Anja Hasse*

*Gordon Heath*

*Agnes Kolmer*

*Mike Lingg*

*Patrick Mächler*

*Eva Peters*

*Uliana Petrunina*

*Janine Richner-Steiner*

*Hana Ruch*

*Beni Ruef*

*Phillip Ströbel*

*Simone Ueberwasser*

*Alexandra Zoller*



**University of  
Zurich** <sup>UZH</sup>

**Language and Space Lab**

**Data**

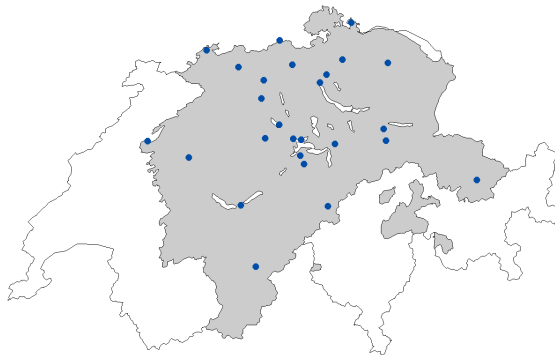


## Oral history project ArchiMob





## The ArchiMob corpus sample





## Some numbers

**44** documents selected by Janine Richner-Steiner and Matthias Friedli, supervised by Elvira Glaser

Release 1.0 (2016):

- **34** documents, around **500 000** word tokens
- **23/44** documents transcribed in the period 2004–2014
- **11/44** documents transcribed in 2015, in collaboration with Spitch

Next release (2017):

- **43** documents, around **650 000** word tokens
- **6/44** documents transcribed in 2016
- **3/44** in progress



**University of  
Zurich** <sup>UZH</sup>

**Language and Space Lab**

**Format**



## Current format

```
<u start="media_pointers#d1007-T1604" xml:id="d1007-u951" who="person_db#EJos1007">
  <w normalised="ja" tag="ADV" xml:id="d1007-u951-w1">je</w>
  <w normalised="dann" tag="ART" xml:id="d1007-u951-w2">de</w>
  <w normalised="hat" tag="VAFIN" xml:id="d1007-u951-w3">het</w>
  <w normalised="man" tag="PIS" xml:id="d1007-u951-w4">me</w>
  <w normalised="noch" tag="ADV" xml:id="d1007-u951-w5">no</w>
  <w normalised="gelugt" tag="VVPP" xml:id="d1007-u951-w6">gluegt</w>
  <w normalised="gedacht" tag="VVFIN" xml:id="d1007-u951-w7">tänkt</w>
  <w normalised="das ist" tag="KOUS+" xml:id="d1007-u951-w8">dasch</w>
  <w normalised="jetzt" tag="ADV" xml:id="d1007-u951-w9">ez</w>
  <w normalised="der" tag="ART" xml:id="d1007-u951-w10">de</w>
  <w normalised="general" tag="NN" xml:id="d1007-u951-w11">generaal</w>
  <w normalised="ja" tag="ITJ" xml:id="d1007-u951-w12">jaa</w>
  <w normalised="das" tag="PDS" xml:id="d1007-u951-w13">das</w>
  <w normalised="ist" tag="VAFIN" xml:id="d1007-u951-w14">isch</w>
  <w normalised="en" tag="PPER" xml:id="d1007-u951-w15">en</w>
  <w normalised="jetzt" tag="ADV" xml:id="d1007-u951-w16">ez</w>
</u>
```





## Content

je	ja	ITJ
de	dann	ADV
het	hat	VAFIN
me	man	PIS
no	noch	ADV
gluegt	gelugt	VVPP
tänkt	gedacht	VVPP
dasch	das_ist	PDS+
ez	jetzt	ADV
de	der	ART
genneraal	general	NN
jaa	ja	ITJ
das	das	PDS
ischsch	ist	VAFIN
en	en	PPER
ez	jetzt	ADV



**University of  
Zurich** <sup>UZH</sup>

**Language and Space Lab**

# Transcription



## Transcription

je	ja	ITJ
de	dann	ADV
het	hat	VAFIN
me	man	PIS
no	noch	ADV
gluegt	gelugt	VVPP
tänkt	gedacht	VVPP
dasch	das_ist	PDS+
ez	jetzt	ADV
de	der	ART
genneraal	general	NN
jaa	ja	ITJ
das	das	PDS
ischsch	ist	VAFIN
en	en	PPER
ez	jetzt	ADV



## Manual transcription

1. 16 documents - Nisus Writer
  - No segmentation (only turns)
  - No text to speech alignment
  - Converted into XML, added segmentation and alignment
2. 7 documents - FOLKER (Schmidt, 2012)
  - Segmented into chunks of 4-10 seconds
  - XML and alignment output
3. 11 documents - EXMARaLDA (Schmidt, 2012)
  - same as FOLKER, just more convenient



## Some details

- Based on Dieth guidelines, but gradually simplified
- Utterance as the basic unit
- Turns not explicitly annotated
- Inconsistence in writing (pronouns and clitics)
- Pauses, repetitions
- Incomprehensible speech



## Normalisation

je	ja	ITJ
de	dann	ADV
het	hat	VAFIN
me	man	PIS
no	noch	ADV
gluegt	gelugt	VVPP
tänkt	gedacht	VVPP
dasch	das_ist	PDS+
ez	jetzt	ADV
de	der	ART
genneraal	general	NN
jaa	ja	ITJ
das	das	PDS
ischsch	ist	VAFIN
en	en	PPER
ez	jetzt	ADV



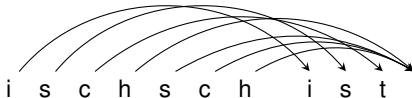
## Approach

- Manual normalisation of 6 documents, VARD2 and IGT
- Automatic normalisation
  - Character-level machine translation (CSMT) with MOSES
  - Training on the 6 manually normalised documents

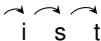


## CSMT

Translation model:  $p(\textit{normalised}|\textit{transcribed})$



Language model:  $p(\textit{normalised}_i|\textit{normalised}_{i-1})$







## Current state of the art

Yves **Scherrer** and Nikola **Ljubešić** (KONVENS 2016)

- Larger translation units (utterances instead of words)
- Language model augmented with German spoken data
- Improved tuning
- Result: 90.46 % accuracy



**University of  
Zurich** <sup>UZH</sup>

**Language and Space Lab**

# Part-of-speech tagging



## Part-of-speech

je	ja	ITJ
de	dann	ADV
het	hat	VAFIN
me	man	PIS
no	noch	ADV
gluegt	gelugt	VVPP
tänkt	gedacht	VVPP
dasch	das_ist	PDS+
ez	jetzt	ADV
de	der	ART
genneraal	general	NN
jaa	ja	ITJ
das	das	PDS
ischsch	ist	VAFIN
en	en	PPER
ez	jetzt	ADV



## Tagger development

STTS+ tag set

	Train	Test	% Acc.	% OOV
Starting	TüBa-D/S	Normalised	70.31	24.21
	NOAH	Original	60.56	30.72
Removed punctuation	TüBa-D/S	Normalised	70.68	24.21
	NOAH	Original	73.09	30.72
Adapted	NOAH + ArchiMob	Original	<b>90.09</b>	–



## Current activities

Tagger adaptation:

- Active learning: gradually add ArchiMob data in the train set
- CRF tagger



**University of  
Zurich** <sup>UZH</sup>

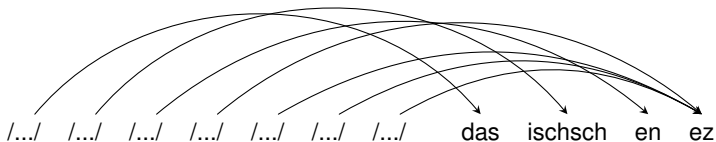
**Language and Space Lab**

**Speech-to-text**

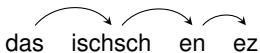


## Speech-to-text

Acoustic model:  $p(\textit{transcribed}|\textit{sound})$



Language model:  $p(\textit{transcribed}_i|\textit{transcribed}_{i-1})$





## Approach

- Improving Spitch prototype with new language models
- Our own speech-to-text development with Kaldi
- Manual transcription





**University of  
Zurich** <sup>UZH</sup>

**Language and Space Lab**

**Next steps**



## Next steps

- Continue transcription, PoS tagging, normalisation
- Neural transducers (deep learning) for normalisation
- Subword language models for speech-to-text
- New data



**University of  
Zurich** <sup>UZH</sup>

**Language and Space Lab**

**Your feedback!**