ArchiMob — A Corpus of Spoken Swiss German

Tanja Samardžic
URPP Language and Space
University of Zurich

Yves Scherrer
LATL-CLCL
University of Geneva

Elvira Glaser
German Department
University of Zurich

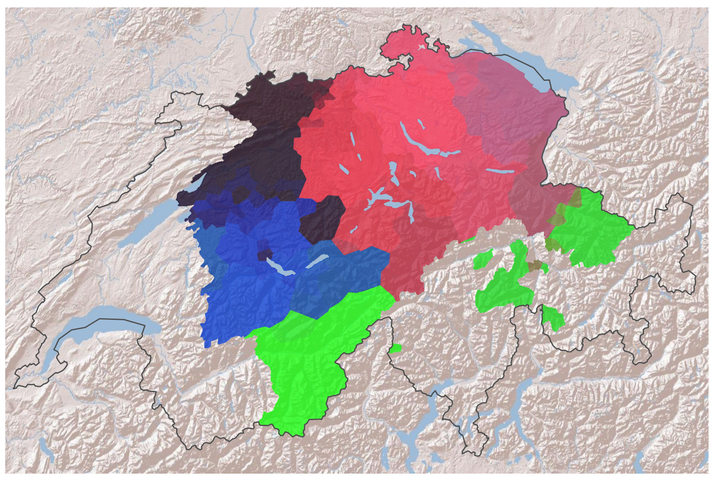
Data

www.archimob.ch

34 documents, around 500 000 tokens
*
Normalised, PoS tagged
*
Aligned with the audio source at segment level (4-8 s long)
*
Available from
http://www.spur.uzh.ch/en/departments/korpuslab.html

Transcription	Normalisation	PoS
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

Dialect Areas



Transcription and alignment

Transcription tools:
16 documents transcribed without specialised software
7 with FOLKER and 11 with EXMARaLDA

Writing conventions:
Based on Dieth guidelines, but gradually simplified
Reflect intra-speaker variation as well as dialectal variation

Text-to-sound alignment tools:
WebMAUS, EXMARaLDA

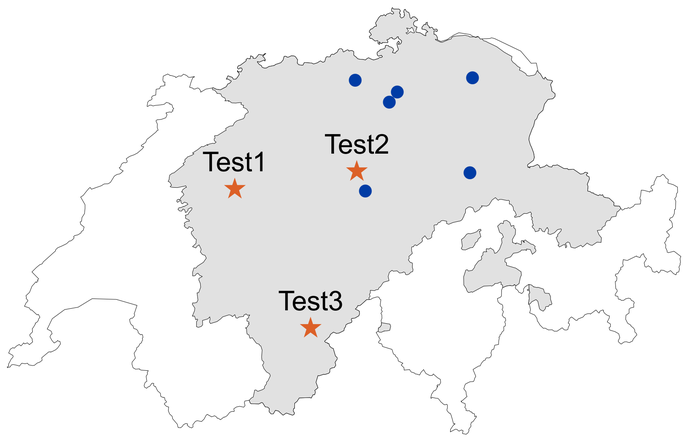
Normalisation

Goal: Establish lexical identities between the tokens that represent "the same word"

Manual normalisation: 6 documents (Initial set) + 1 (Test1) + 1 (Test2) using VARD, IGT

Automatic normalisation experiments: (Character-based) machine translation

Training data	Initial		Initial		Initial		Initial		Initial+T1+T2	
Test data	Cross-validation		Test1		Test2		Test3		Test3	
	Prop.	Acc.	Prop.	Acc.	Prop.	Acc.	Prop.	Acc.	Prop.	Acc.
Unique	46.69	98.13	33.63	96.50	43.02	98.60	41.82	95.28	43.50	95.33
Ambiguous	41.63	81.96	46.81	76.42	43.39	87.96	37.49	78.42	37.65	78.39
New	11.68	35.90	19.57	50.40	13.59	51.47	20.70	44.45	18.85	42.01
All		84.13		78.08		87.58		78.44		78.90



Part-of-speech tagging

Initial annotation:
BTagger trained on TüBa-D/S and NOAH
Applied to original or normalised forms

Train	Test	% Acc.	% OOV
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

Tagset:
STTS+

Tagger adaptation:
Gradually add ArchiMob documents in the train set (4 x 1 document)

Round	1	2	3	4
Accuracy	79.99	84.08	88.55	90.09

Conclusion

- First corpus of spoken Swiss German developed using up-to-date technology
- Suitable for studying syntactic regional variation and for training new natural language processing tools

Recent experiments (Scherrer & Ljubešić, KONVENS 2016)

		Baselines and ceilings				Isolated words		Segments		Constr.
	Prop.	Ident.	Baseline	Combi	Ceiling	1 LM	2 LM	1 LM	2 LM	2 LM
Unique	46.63	22.84	98.79	98.79	98.98	98.30	98.22	97.25	97.64	98.69
Ambig.	42.12	23.52	84.06	84.20	84.64	83.45	82.52	86.27	87.54	87.92
New	11.25	9.46	9.46	35.33	99.57	53.15	53.91	52.50	63.59	65.87
All		21.62	82.54	85.51	93.00	86.96	86.62	87.59	89.56	90.46

- Isolated words

  CSMT for all words, 7-gram LM, no reordering, tuned with WER (=CER)
- Segments

  Segment $\approx$ utterance, between 4 and 8 seconds long, 8 words/avg

  CSMT for all words, 10-gram LM, no reordering, tuned with WER (=CER)
- Constrained

  Constrain to baseline translation for Unique words (Moses XML)
- 2 LM

  Training data + German OpenSubtitles2016

Corpus-based dialectology

Example:

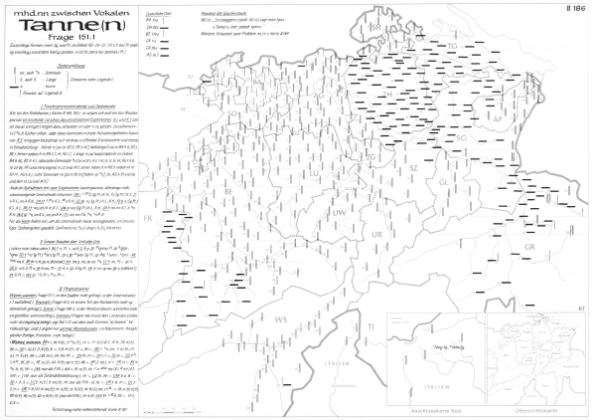
- Which dialect phonemes are normalized as *nn*?
- Train distinct CSMT models, look for $p(* \mid nn)$ in phrase tables
- Document 1:

```
n n ||| n n ||| 0.464052 0.212927 0.934211 0.698887 ||| 0-0 1-1 ||| 153 76 71 |||
n ||| n n ||| 0.535948 0.46144 0.0363636 0.698887 ||| 0-0 0-1 ||| 153 2255 82 |||
```

- Document 2:

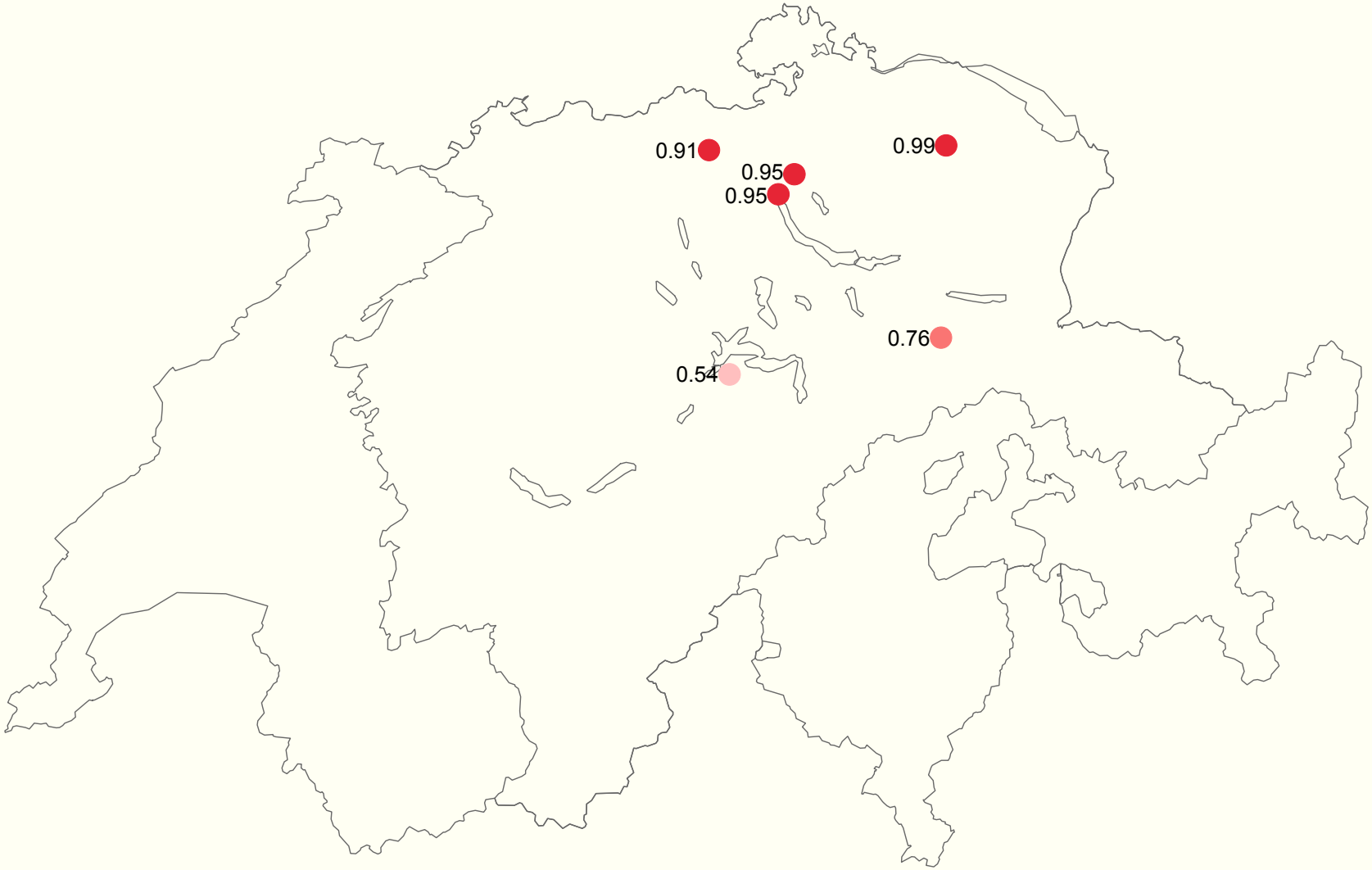
```
m ||| n n ||| 0.00456108 0.002902 0.00116594 0.000606745 ||| 0-0 ||| 511 1999 3 |||
n e ||| n n ||| 0.00118076 0.104172 0.00251403 0.133153 ||| 0-0 1-1 ||| 511 240 1 |||
n n ||| n n ||| 0.037182 0.303291 0.76 0.681866 ||| 0-0 1-1 ||| 511 25 19 ||| |||
n ||| n n ||| 0.95499 0.550719 0.129066 0.681866 ||| 0-0 0-1 ||| 511 3781 488 |||
```

- Comparison with existing atlas data:
  Sprachatlas der deutschen Schweiz
  (SDS, Hotzenköcherle et al., 1962-1997)



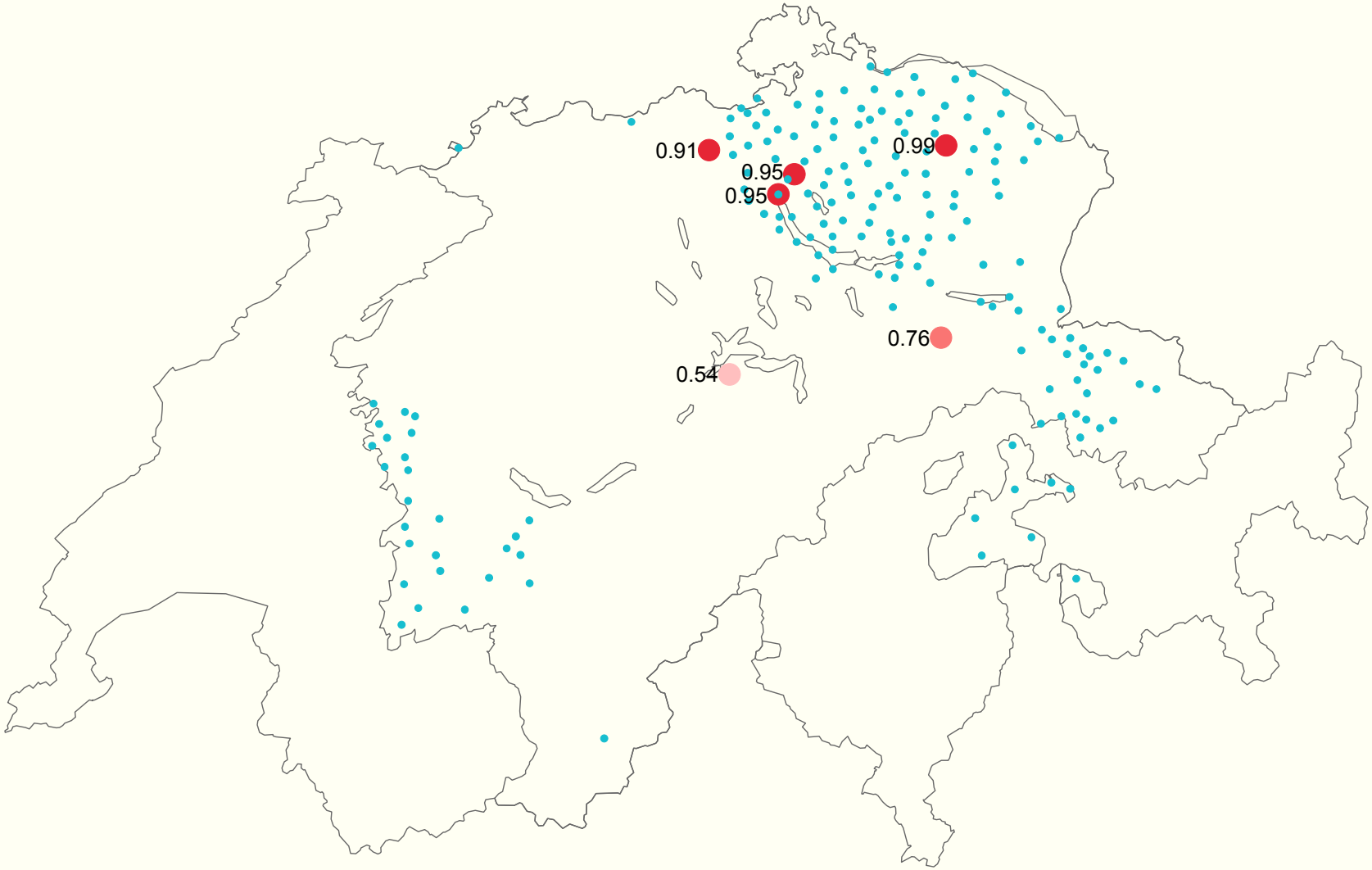
p(n|nn)

Example: Tane|Tanne ‘fir tree’



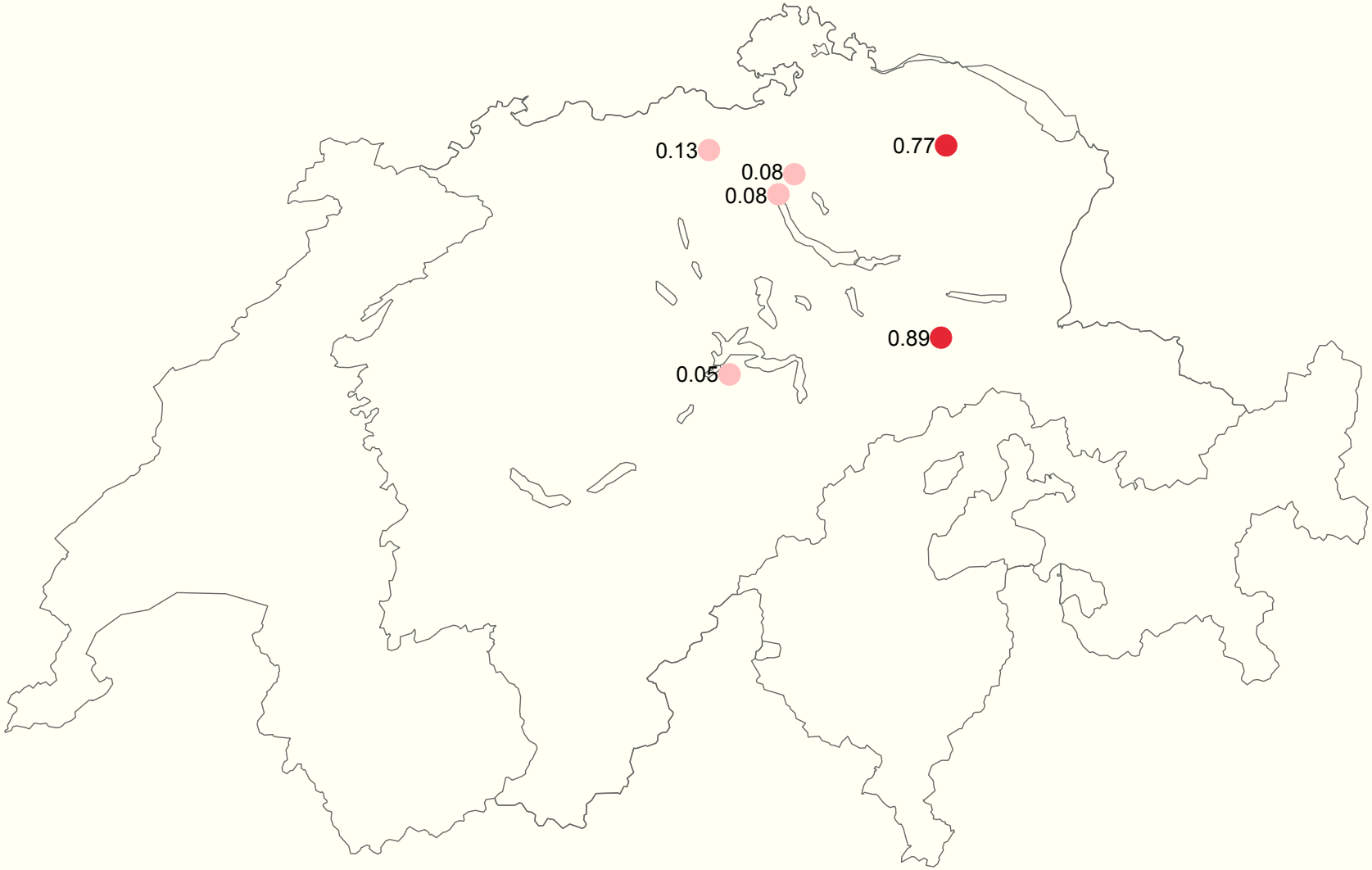
p(n|nn)

Comparison with SDS map 2/179 “Tanne”



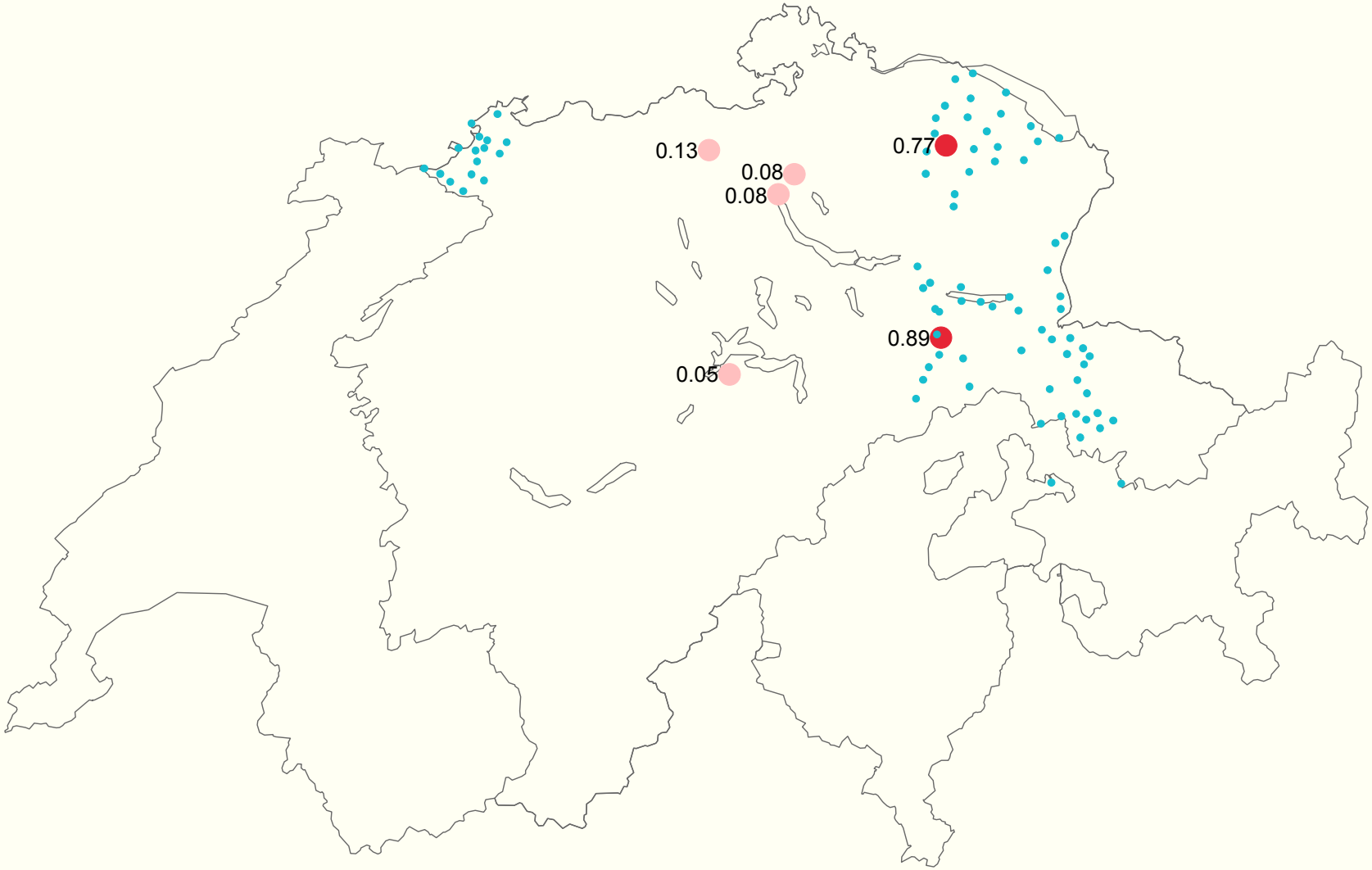
p(gg|ck)

Example: drugge|drücken ‘to push’



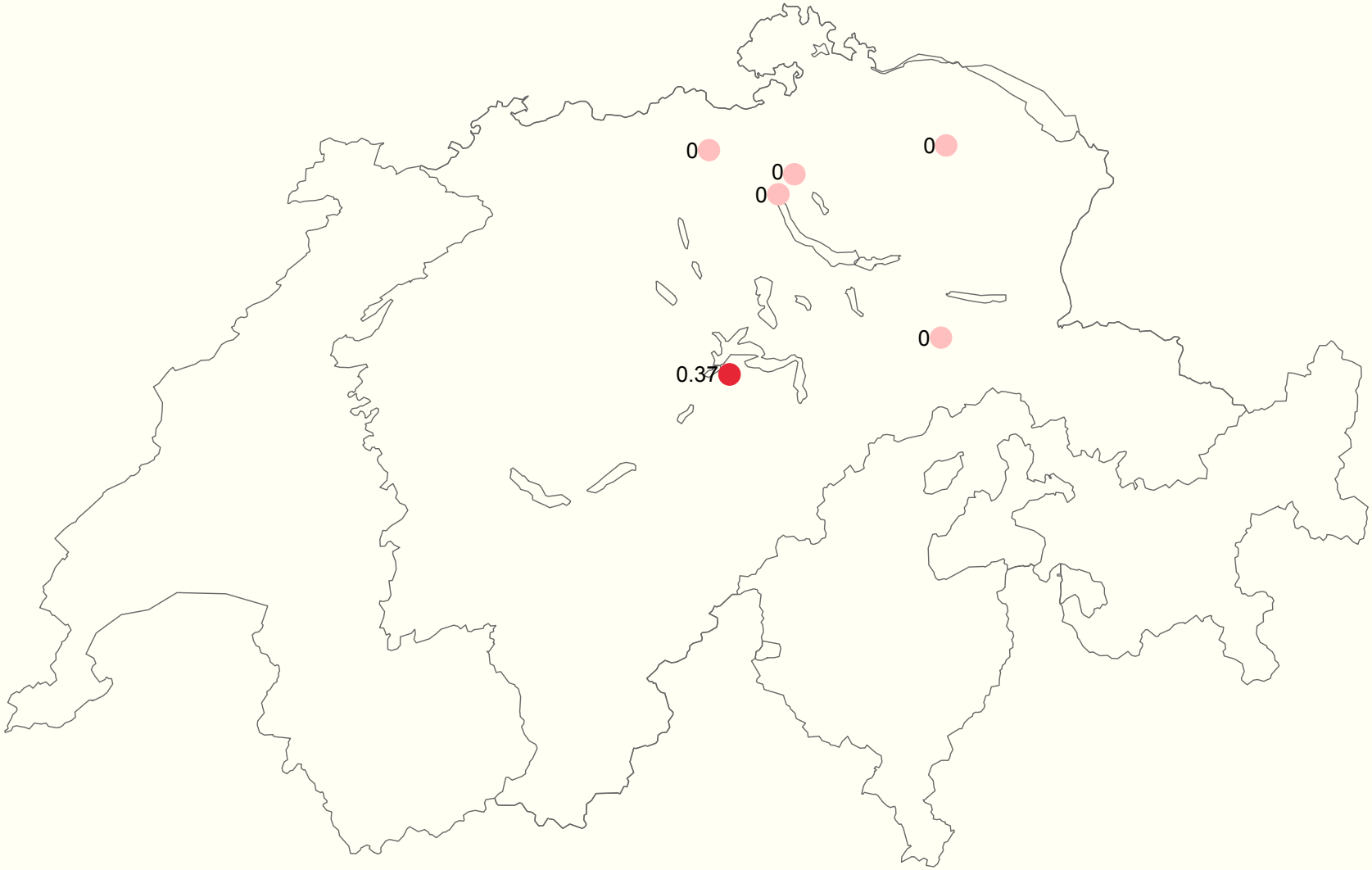
p(gg|ck)

Comparison with SDS map 2/095 “drücken”



p(ui|au)

Example: Muis|Maus ‘mouse’



p(ui|au)

Comparison with SDS map 1/106 “Maus”