ArchiMob: A multidialectal corpus of Swiss German oral history interviews

Yves Scherrer[1] & Tanja Samardžić[2]

1. Department of Digital Humanities, University of Helsinki, Finland
2. Language and Space Lab, University of Zurich, Switzerland

ArchiMob – an oral history project and a corpus

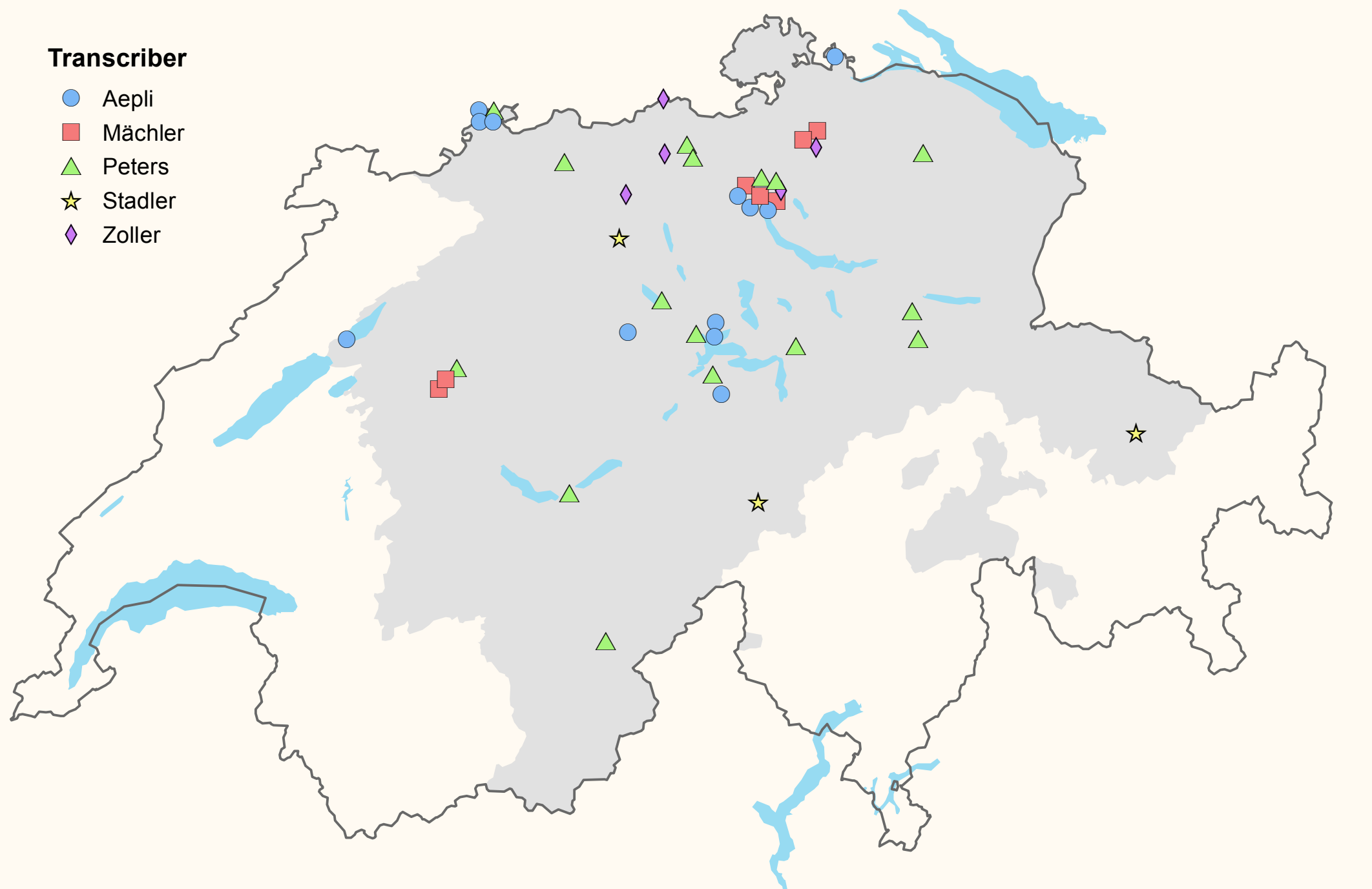
L'Histoire c'est moi
555 Versionen der Schweizer Geschichte
555 versions de l'histoire suisse
555 versioni della storia svizzera
1939–1945
Deutsch
Français
Italiano

The Archimob project:

- An oral history project set up by Swiss historians and cineasts around the year 2000
- 555 video-recorded interviews with contemporary witnesses of the WW II period in Switzerland (***Archi**ves de la **mob**ilisation*)
- Witnesses of various social and linguistic backgrounds

The ArchiMob corpus:

- 43 interviews conducted in Swiss German dialect selected for corpus
  - Good audio quality
  - Wide dialectal coverage
  - Transcribed and annotated
- Interview lengths: 1-2 hours
- Corpus size: 650 000 tokens



Annotation and access

Three annotation layers:

1. Transcription and alignment with audio source:

- Manual, using transcription tools such as EXMARaLDA
- 5 transcribers

2. Normalization:

- Manual normalization of 6 interviews
- Automatic normalization of the remaining interviews with character-level statistical machine translation
- Estimated accuracy: 90%

3. Part-of-speech tagging:

- Automatic tagging with a tagger trained on a previously annotated Swiss German corpus
- Bootstrapping by manual correction and retraining
- Estimated accuracy: around 90%

Example:

Transcr.	Norm.	POS
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	ihn	PPER
ez	jetzt	ADV

Access:

1. XML transcriptions:

- Free download

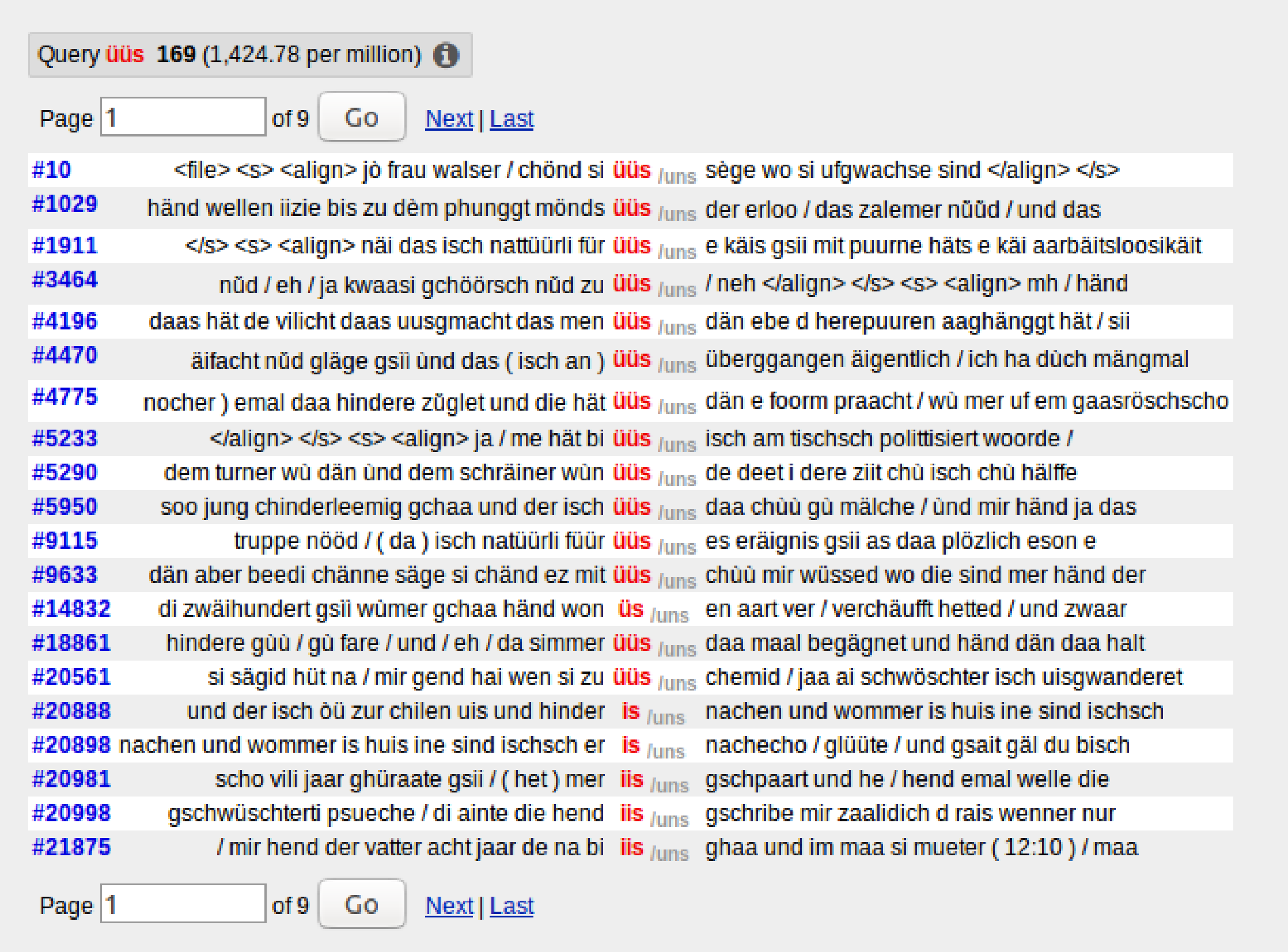
3. Audio data:

- Available on request

```
https://doi.org/
10.5281/zenodo.
1158572
```

2. Corpus query engines:

- ANNIS / Sketch Engine with flexible search:



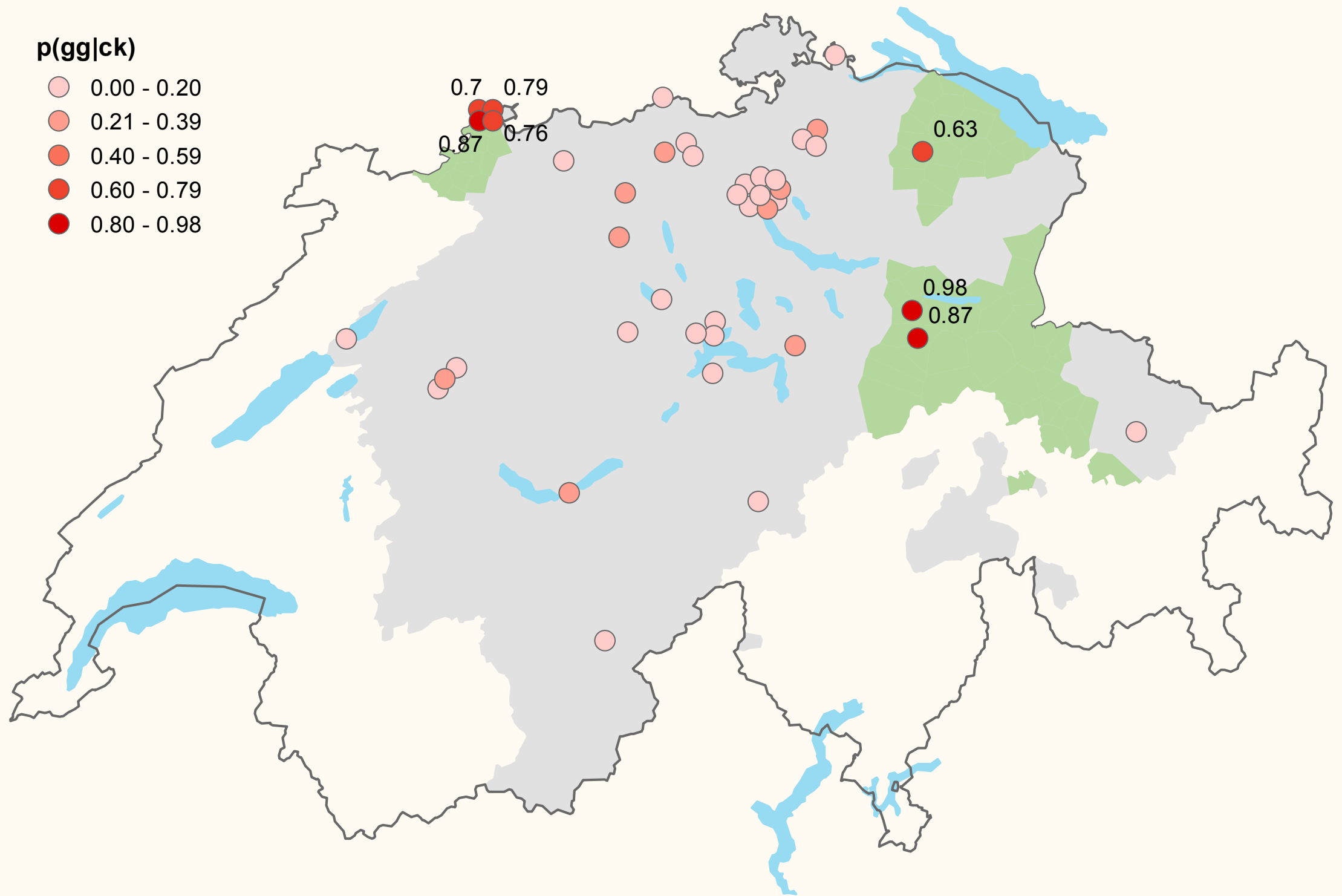
Query üüs 169 (1,424.78 per million)

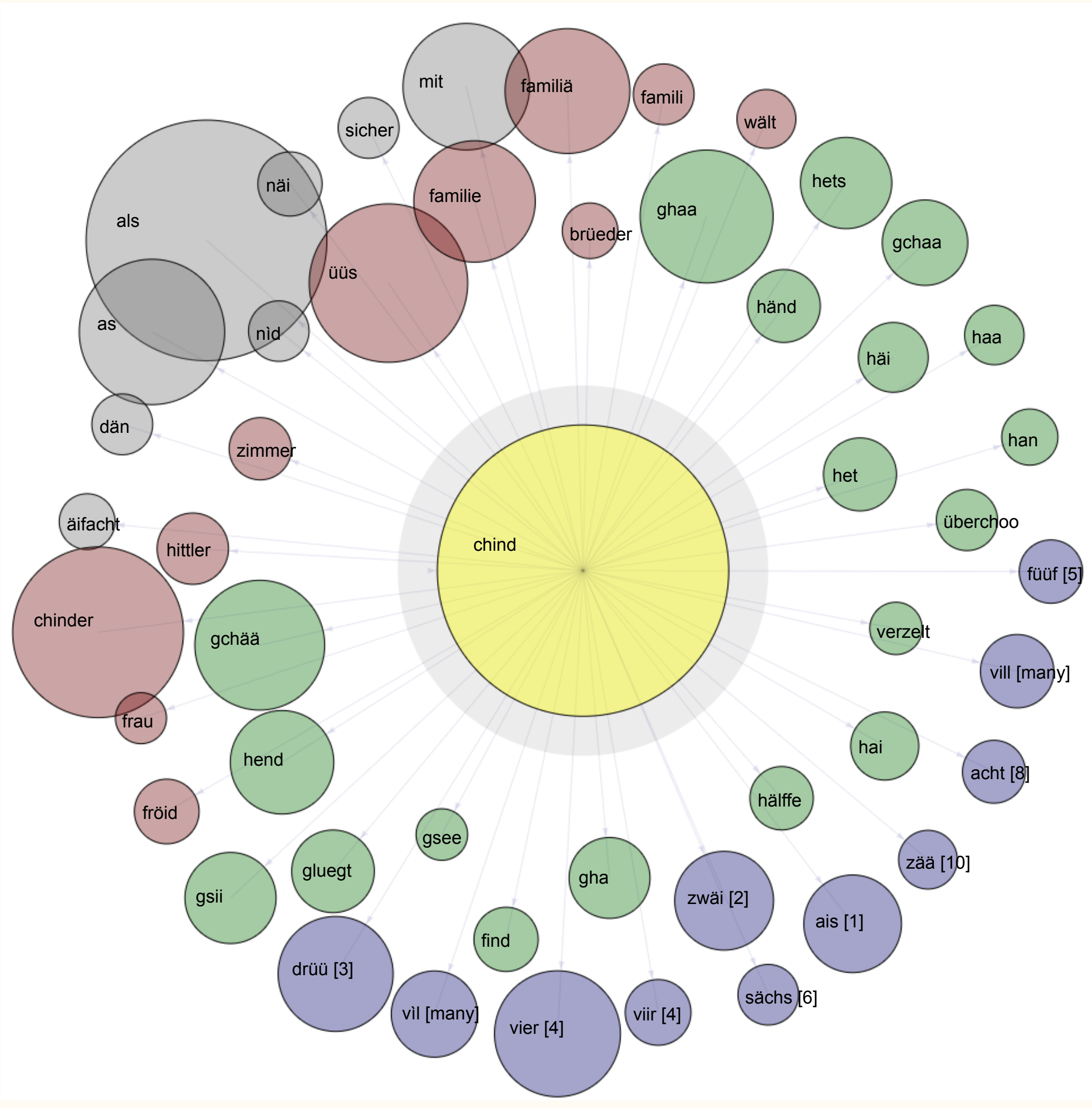
Page 1 of 9 Go Next | Last

#10	<file> <s> <align> jò frau walser / chönd si	üüs	/uns	sège wo si ufgwachse sind </align> </s>
#1029	händ wellen iizie bis zu dèm phunggt mönds	üüs	/uns	der erloo / das zalemer nüüd / und das
#1911	</s> <s> <align> näi das isch nattüürli für	üüs	/uns	e käis gsii mit puurne häts e käi aarbäitsloosikäit
#3464	nüd / eh / ja kwaasi gchöörsch nüd zu	üüs	/uns	/ neh </align> </s> <s> <align> mh / händ
#4196	daas hät de vilicht daas uusgmacht das men	üüs	/uns	dän ebe d herepuuren aaghänggt hät / sii
#4470	äifacht nüd gläge gsii ùnd das (isch an)	üüs	/uns	überggangen äigentlich / ich ha dùch mängmal
#4775	nocher) emal daa hindere züglet und die hät	üüs	/uns	dän e foorm praacht / wù mer uf em gaasröschscho
#5233	</align> </s> <s> <align> ja / me hät bi	üüs	/uns	isch am tischsch polittisiert woorde /
#5290	dem turner wù dän ùnd dem schräiner wùn	üüs	/uns	de deet i dere ziit chù isch chù hälffe
#5950	soo jung chinderleemig gchaa und der isch	üüs	/uns	daa chùù gù mälche / ùnd mir händ ja das
#9115	truppe nööd / (da) isch natüürli füür	üüs	/uns	es eräignis gsii as daa plözlich eson e
#9633	dän aber beedi chänne säge si chänd ez mit	üüs	/uns	chùù mir wüssed wo die sind mer händ der
#14832	di zwäihundert gsii wùmer gchaa händ won	üs	/uns	en aart ver / verchäufft hetted / und zwaar
#18861	hindere gùù / gù fare / und / eh / da simmer	üüs	/uns	daa maal begägnet und händ dän daa halt
#20561	si sägid hüt na / mir gend hai wen si zu	üüs	/uns	chemid / jaa ai schwöschter isch uisgwanderet
#20888	und der isch öü zur chilen uis und hinder	is	/uns	nachen und wommer is huis ine sind ischsch
#20898	nachen und wommer is huis ine sind ischsch er	is	/uns	nachecho / glüüte / und gsait gäl du bisch
#20981	scho vili jaar ghüraate gsii / (het) mer	iis	/uns	gschpaart und he / hend emal welle die
#20998	gschwüschterti psueche / di ainte die hend	iis	/uns	gschribe mir zaalidich d rais wenner nur
#21875	/ mir hend der vatter acht jaar de na bi	iis	/uns	ghaa und im maa si mueter (12:10) / maa

Page 1 of 9 Go Next | Last

Using the ArchiMob corpus in digital humanities research





1. Comparison of dialectal variation patterns with atlas data

- Red dots: Proportion of normalized *ck* occurrences that are dialectally realized as *gg*
- Green zones: *gg*-areas according to the linguistic atlas of German-speaking Switzerland (SDS)

2. Clustering of aggregate distances between documents

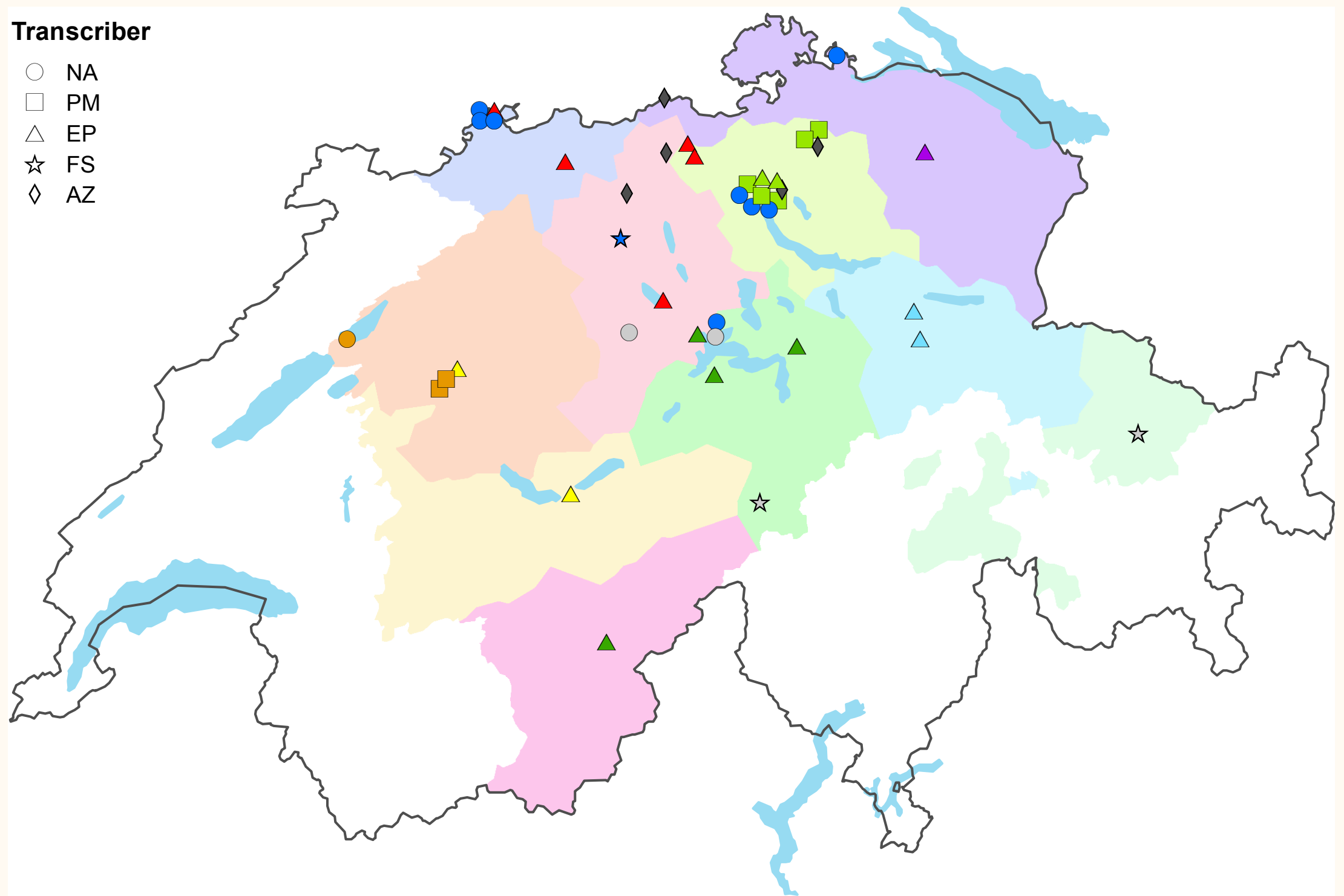
- Train a 4-gram language model on each document
- Evaluate each document with each LM using perplexity measure → distance matrix
- Colored symbols: Hierarchical clustering of documents according to distance matrix
- Colored zones: Comparison with clustering obtained from SDS atlas data

3. Collocation analysis

- Which words collocate with *chind* 'child'?
- Numerals (in blue) provide information about typical family sizes in the first half of the 20th century

4. Lexicological analysis

- How is the usage of the friendship terms *koleeg* 'colleague' and *fründ* 'friend' distributed across gender? (Schifferle 2017)
- Legend:
  - ◐ Person working in the same profession (e.g. doctor)
  - ● Person working in the same company, military colleague, school/club colleague
  - ○■ Close friend, school friend
  - □ Lover



ID/birth date	*-koleeg-*	*-fründ-*
Female speakers:		
1073 (*1908)	●	■■■■■■■■■■ ■■■■
1007 (*1912)	—	—
1063 (*1918)	◐●○○●●●	■■■□
1170 (*1918)	—	■
1212 (*1921)	—	—
1270 (*1923)	●	□
1048 (*1925)	●	■■
1261 (*1932)	○	■■■■□
Male speakers:		
1195 (*1914)	○	—
1147 (*1915)	●●●	■□
1198 (*1915)	●●●●●○○	—
1207 (*1916)	●●●	—
1142 (*1920)	—	—
1143 (*1924)	●●●●●●●●●○ ○○○	—
1057 (*1925)	●●●●●●●●●	—
1209 (*1209)	●●●	—