NLP for Dialectology
The Swiss German oral history corpus ArchiMob

Yves Scherrer, Language Technology, University of Helsinki

In collaboration with:
Tanja Samardžić, CorpusLab, University of Zurich
Prof. Elvira Glaser, Department of German, University of Zurich

The ArchiMob project

555 interviews with contemporary witnesses of the WWII period

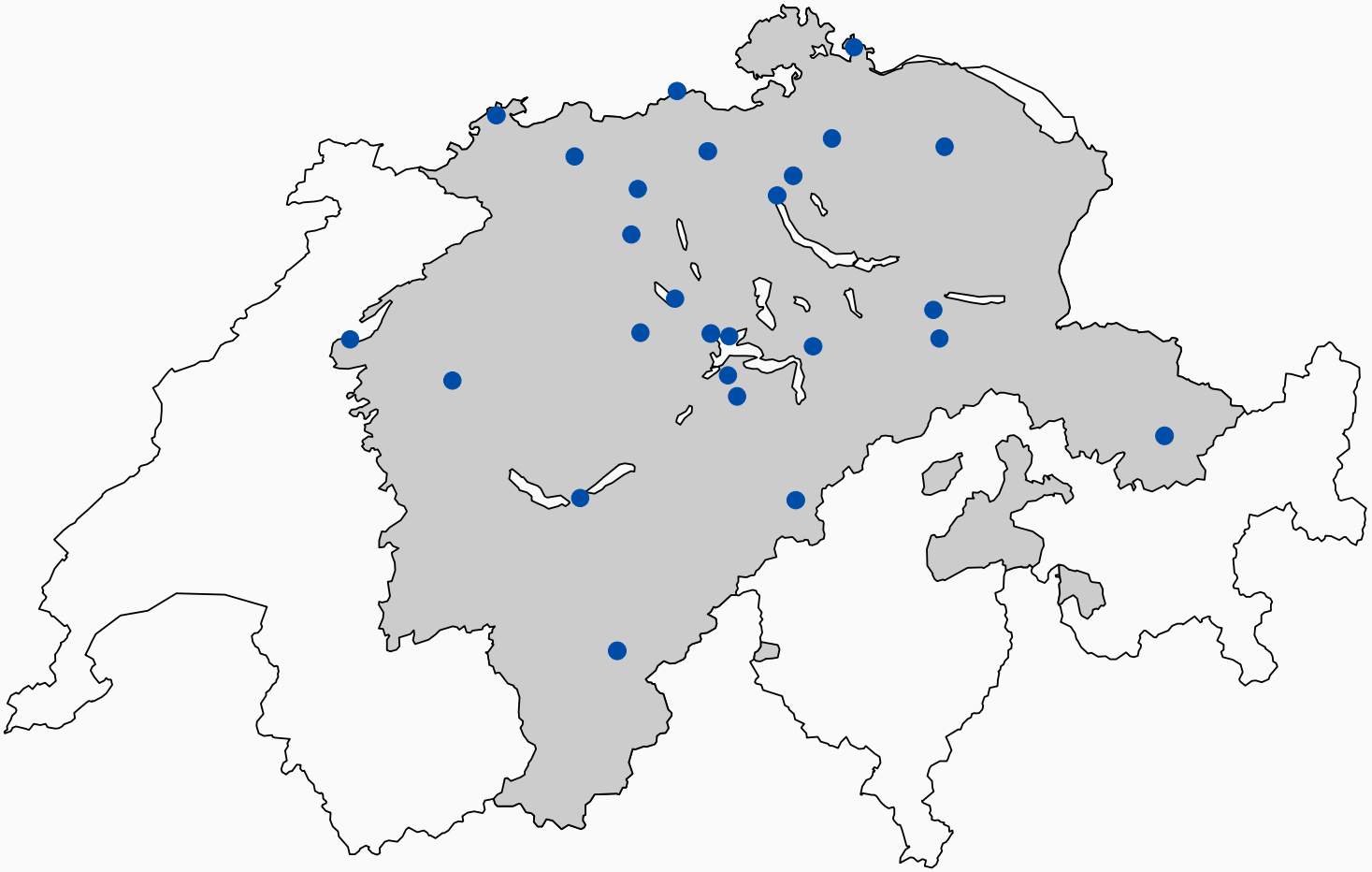
Witnesses of various social and linguistic backgrounds

Video recordings

Document selection

43 interviews (~650 000 tokens) selected for corpus:

- Good audio quality
- Coverage of various Swiss German dialects



Annotation

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	en	PPER
ez	jetzt	ADV

Four annotation layers:

- Transcription (manual)
- Alignment with sound source (semi-automatic)
- Normalization (semi-automatic)
- Part-of-speech tagging (semi-automatic)

Support from Language Technology?

Dialectological research results?

Annotation

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	en	PPER
ez	jetzt	ADV

Four annotation layers:

- Transcription (manual)
- Alignment with sound source (semi-automatic)
- Normalization (semi-automatic)
- Part-of-speech tagging (semi-automatic)

Support from Language Technology?

Dialectological research results?

Transcription

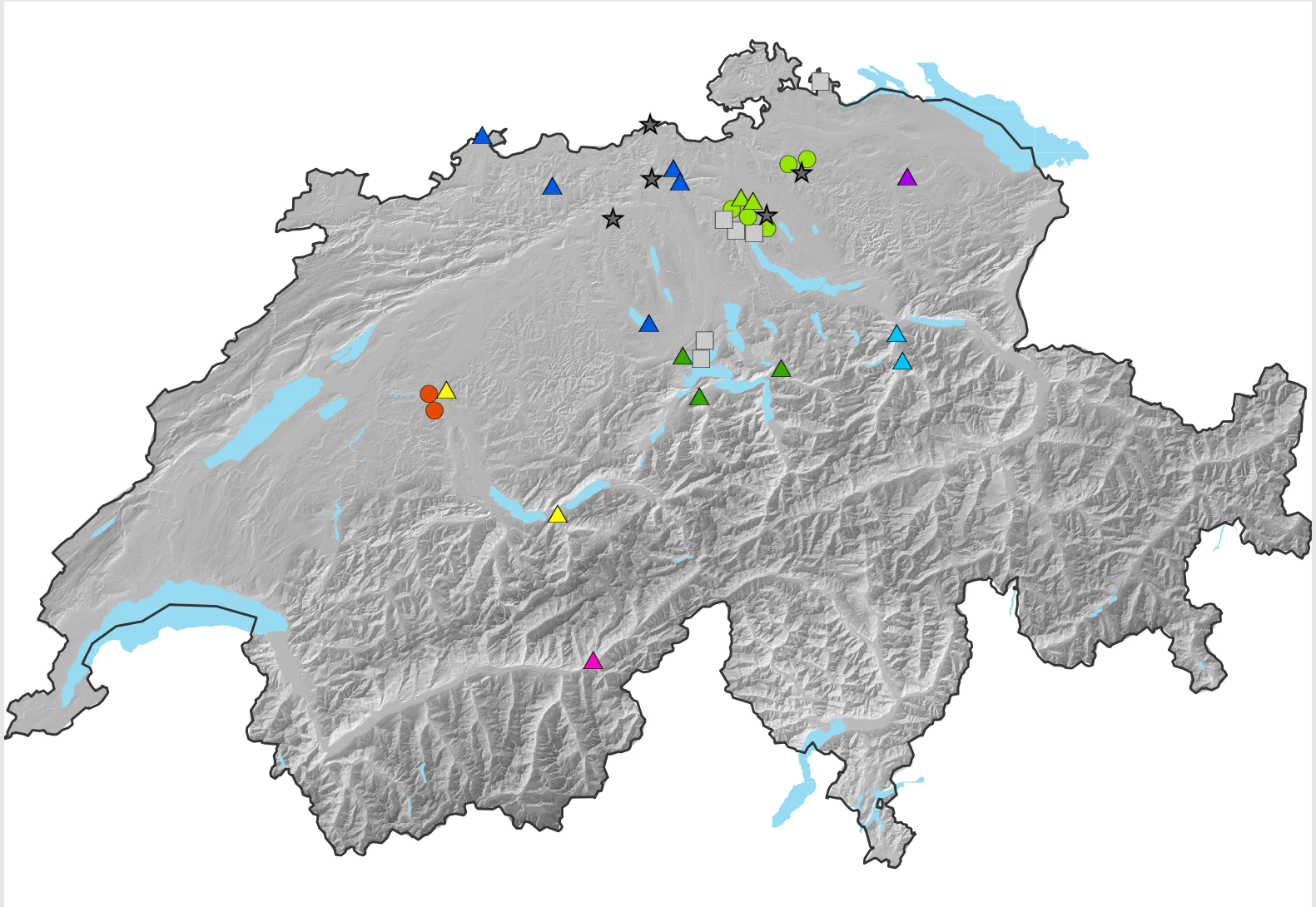
Support from LT

1. Language modelling:
   Train language model on text A, compute its perplexity on text B
   Cluster texts according to distance from others
2. Collocation analysis:
   Find word associations

Results

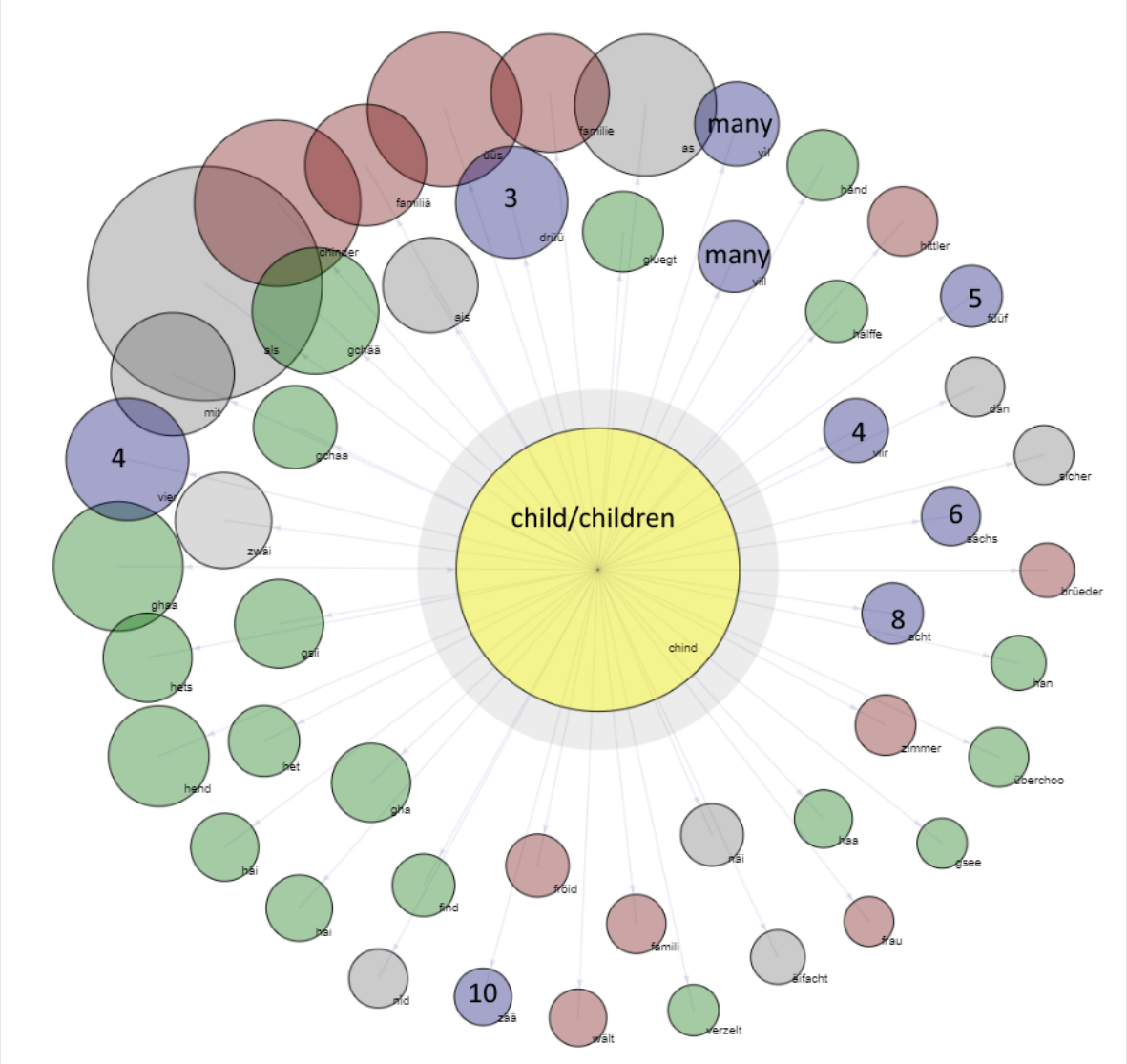
1. Dialect and transcriber clustering
2. Content analysis

Result: dialect and transcriber clustering



Transcription

Result: collocation analysis



Normalization

Group together different transcribed forms of “the same word” due to:

- Dialectal variation — *hät, hed, hèt, hèd* → *hat*
- Intra-speaker variation — *mi, mii, miin, min* → *mein*
- Code-switching — *main, määin, mäin* → *mein*

Support from LT

- Character-level machine translation
- Trained on 5 manually normalized documents

Results

- Mapping of dialectal variation patterns
- Search for lexical or syntactic features independently of phonetic variation

Normalization

Group together different transcribed forms of “the same word” due to:

- Dialectal variation — *hät, hed, hèt, hèd* → *hat*
- Intra-speaker variation — *mi, mii, miin, min* → *mein*
- Code-switching — *main, määin, mäin* → *mein*

Support from LT

- Character-level machine translation
- Trained on 5 manually normalized documents

Results

- Mapping of dialectal variation patterns
- Search for lexical or syntactic features independently of phonetic variation

Normalization

Group together different transcribed forms of “the same word” due to:

- Dialectal variation — *hät, hed, hèt, hèd* → *hat*
- Intra-speaker variation — *mi, mii, miin, min* → *mein*
- Code-switching — *main, määin, mäin* → *mein*

Support from LT

- Character-level machine translation
- Trained on 5 manually normalized documents

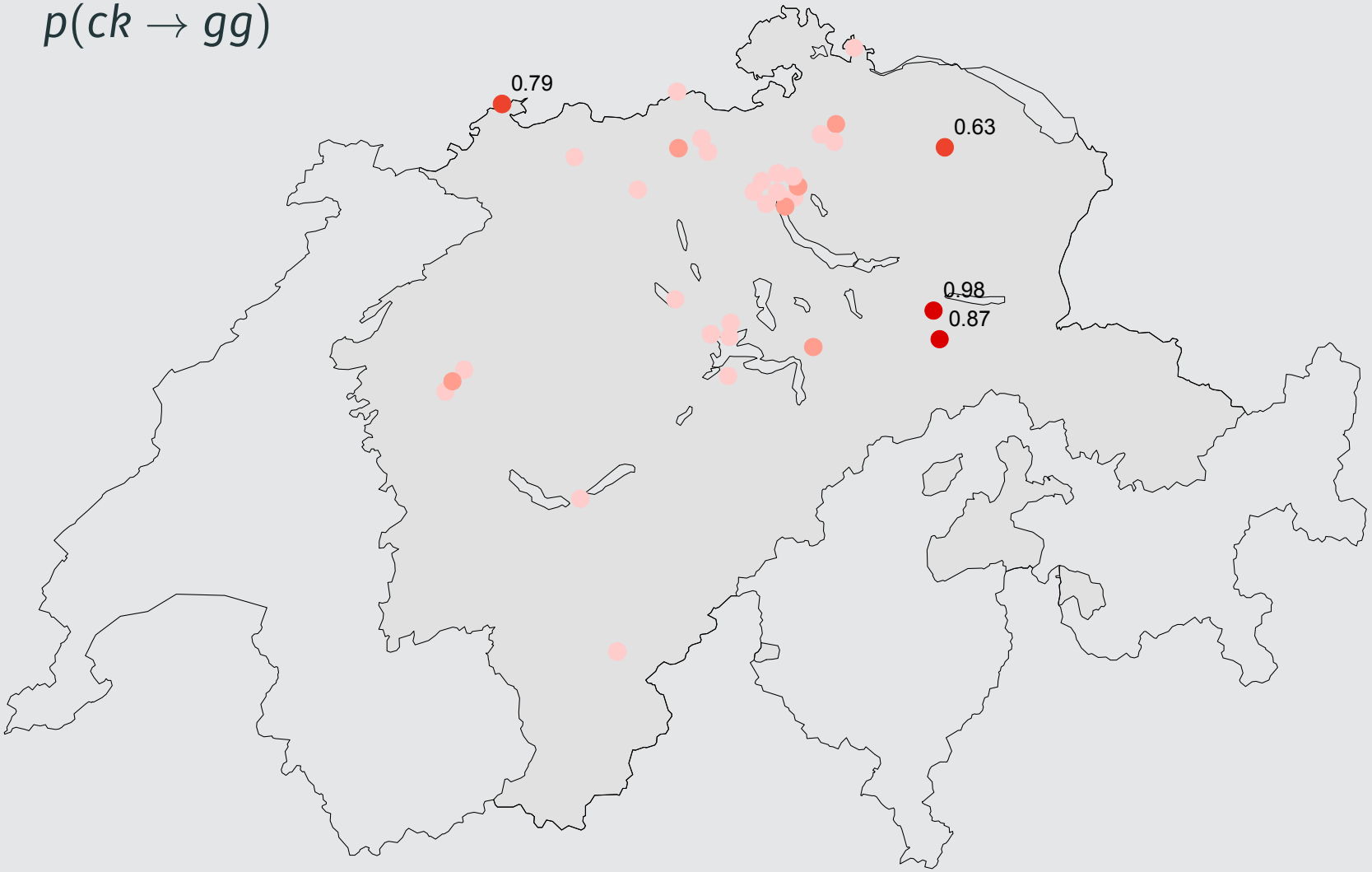
Results

- Mapping of dialectal variation patterns
- Search for lexical or syntactic features independently of phonetic variation

Normalization

Result: dialectal variation patterns

$p(ck \rightarrow gg)$



Normalization

Application: dialect-independent 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 nûûd / und das
#1911	</s> <s> <align> näi das isch nattüürli für	üüs	/uns	e käis gsii mit puurne häts e käi aarbäitsloosikäit
#3464	nûd / eh / ja kwaasi gchöörsch nûd 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 nûd gläge gsii ùnd das (isch an)	üüs	/uns	überggangen äigentlich / ich ha dùch mängmal
#4775	nocher) emal daa hindere zûglet 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 òü 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

Part-of-speech tagging

Support from LT

Domain adaptation: Use existing training corpora from spoken Standard German or from written Swiss German

Bootstrapping: Tag document, manually correct, add to training set

Result

- Facilitating search for syntactic phenomena

Availability

Original project: `http://www.archimob.ch`

Corpus homepage: `http://www.spur.uzh.ch/en/departments/korpuslab/ArchiMob.html`

- XML download
- Audio data on request
- Corpus analysis tools: SketchEngine, ANNIS