ArchiMob: Ein multidialektales Korpus schweizerdeutscher Spontansprache

Yves Scherrer, Universitäten Genf & Helsinki

Tanja Samardžić, CorpusLab, Universität Zürich

Prof. Elvira Glaser, Deutsches Seminar, Universität Zürich

19. Arbeitstagung zur Alemannischen Dialektologie

Freiburg, 12. Oktober 2017

ArchiMob: Vom Projekt zum Korpus

Das ArchiMob-Projekt

Das ArchiMob-Projekt

Ein von Historikern gegründeter Verein mit dem Ziel, die neuere Schweizer Geschichte zu erschliessen (http://www.archimob.ch)

555 Interviews mit Zeitzeugen des Zweiten Weltkriegs in der Schweiz

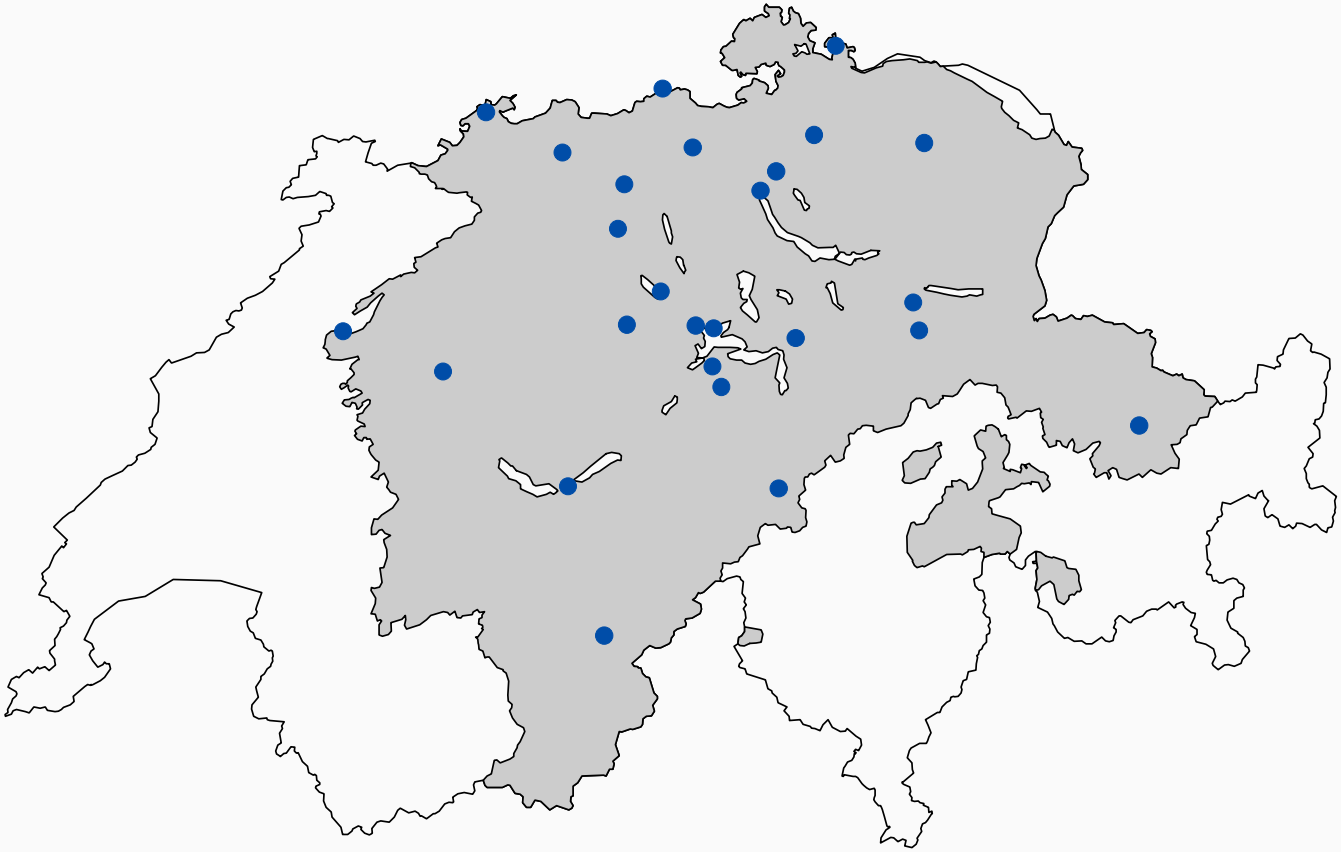
Zeitzeugen aus unterschiedlichen sozialen Klassen und Sprachregionen

Interviews geführt 1999-2001, auf Video aufgezeichnet

Dokumentauswahl

43 Interviews wurden für das Korpus ausgewählt:

- Genügende Tonqualität der Aufnahmen
- Schweizerdeutsch, Abdeckung verschiedener Dialekte (Janine Richner-Steiner, Matthias Friedli, Elvira Glaser)



Annotation

Vier Annotationsebenen:

- Transkription (manuell)
- Alignierung mit Tonquelle (halbautomatisch)
- Normalisierung (halbautomatisch)
- Tagging (halbautomatisch)

Download

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

Transkr.	Norm.	Tagging
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

Annotation

Vier Annotationsebenen:

- Transkription (manuell)
- Alignierung mit Tonquelle (halbautomatisch)
- Normalisierung (halbautomatisch)
- Tagging (halbautomatisch)

Download

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

Transkr.	Norm.	Tagging
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

Transkription

Transkription – Vier Phasen

1. 16 Dokumente (2004-2012)
    - Nisus Writer
    - Keine Segmentation in Äusserungen (nur Turns)
    - Keine Alignierung mit der Tonquelle
    - Nachträglich konvertiert, segmentiert und aligniert
2. 7 Dokumente (2012-2014)
    - FOLKER (Schmidt 2012)
    - Segmentiert in Äusserungen von 4-10 Sekunden
    - Produziert XML-Dateien mit Ton-Alignierung
3. 11 Dokumente (2015)
    - EXMARaLDA (Schmidt 2012, Weiterentwicklung von FOLKER)
    - Zusammenarbeit mit Spitch
4. 9 Dokumente (2016-2017)
    - EXMARaLDA

Transkription – Vier Phasen

Release 1.0 (2016): ~500 000 Tokens

1. 16 Dokumente (2004-2012)
   - Nisus Writer
   - Keine Segmentation in Äusserungen (nur Turns)
   - Keine Alignierung mit der Tonquelle
   - Nachträglich konvertiert, segmentiert und aligniert
2. 7 Dokumente (2012-2014)
   - FOLKER (Schmidt 2012)
   - Segmentiert in Äusserungen von 4-10 Sekunden
   - Produziert XML-Dateien mit Ton-Alignierung
3. 11 Dokumente (2015)
   - EXMARaLDA (Schmidt 2012, Weiterentwicklung von FOLKER)
   - Zusammenarbeit mit Spitch
4. 9 Dokumente (2016-2017)
   - EXMARaLDA

Transkription – Vier Phasen

Release 2.0 (2017-18): ~650 000 Tokens

1. 16 Dokumente (2004-2012)
   - Nisus Writer
   - Keine Segmentation in Äusserungen (nur Turns)
   - Keine Alignierung mit der Tonquelle
   - Nachträglich konvertiert, segmentiert und aligniert
2. 7 Dokumente (2012-2014)
   - FOLKER (Schmidt 2012)
   - Segmentiert in Äusserungen von 4-10 Sekunden
   - Produziert XML-Dateien mit Ton-Alignierung
3. 11 Dokumente (2015)
   - EXMARaLDA (Schmidt 2012, Weiterentwicklung von FOLKER)
   - Zusammenarbeit mit Spitch
4. 9 Dokumente (2016-2017)
   - EXMARaLDA

Transkription – Richtlinien

- Dieth-Schreibung, schrittweise vereinfacht
- Äusserung ist Basiseinheit der Transkription
- Sprecherwechsel werden nicht explizit annotiert (sind aber von XML-Annotation ableitbar)
- Pausen, Wiederholungen, unverständliche Passagen werden als solche annotiert
- 1 Transkriptor pro Dokument (keine Parallelannotation)

Normalisierung

Normalisierung – Warum?

- Inkonsistenzen in der Transkription:
  Transkriptoren, Richtlinien, Transkriptionswerkzeuge
- Dialektale Variation
- Sprecherinterne Variation
- Code-switching

Transkription	*min*	*maa*	*het*	*immer*	*gsaait*
Varianten im selben Text	*mi* *mii* *miin*	*ma*	*hät*		*gsait*
Varianten in anderen Texten	*mine*		*hed* *hèd* *hèt*	*ime* *imer* *emmer*	*gsäit* *gsääit*
Code-switching	*määin* *mäin* *main*	*man*			

Normalisierung – Warum?

- Inkonsistenzen in der Transkription:
  Transkriptoren, Richtlinien, Transkriptionswerkzeuge
- Dialektale Variation
- Sprecherinterne Variation
- Code-switching

Transkription	*min*	*maa*	*het*	*immer*	*gsaait*
Varianten im selben Text	*mi* *mii* *miin*	*ma*	*hät*		*gsait*
Varianten in anderen Texten	*mine*		*hed* *hèd* *hèt*	*ime* *imer* *emmer*	*gsäit* *gsääit*
Code-switching	*määin* *mäin* *main*	*man*			

Normalisierung – Ziele und Richtlinien

Ziel: Zusätzliche Annotationsebene, um Formen zu gruppieren, die „dasselbe Wort“ darstellen

Transkription	Normalisierung
jaa	*ja*
de	*dann*
het	*hat*
me	*man*
no	*noch*
gluegt	*gelugt*
tänkt	*gedacht*
dasch	*das ist*
ez	*jetzt*
de	*der*
genneraal	*general*

Normalisierung – Ziele und Richtlinien

Ziel: Zusätzliche Annotationsebene, um Formen zu gruppieren, die „dasselbe Wort“ darstellen

Transkription	Normalisierung
jaa	*ja*
de	*dann*
het	*hat*
me	*man*
no	*noch*
gluegt	*gelugt*
tänkt	*gedacht*
dasch	*das ist*
ez	*jetzt*
de	*der*
genneraal	*general*

Normalisierung – Ziele und Richtlinien

Ziel: Zusätzliche Annotationsebene, um Formen zu gruppieren, die „dasselbe Wort“ darstellen

Transkription	Normalisierung
jaa	*ja*
de	*dann*
het	*hat*
me	*man*
no	*noch*
gluegt	*gelugt*
tänkt	*gedacht*
dasch	*das ist*
ez	*jetzt*
de	*der*
genneraal	*general*

Normalisierung – Ziele und Richtlinien

Ziel: Zusätzliche Annotationsebene, um Formen zu gruppieren, die „dasselbe Wort“ darstellen

Transkription	Normalisierung
jaa	*ja*
de	*dann*
het	*hat*
me	*man*
no	*noch*
gluegt	*gelugt*
tänkt	*gedacht*
dasch	*das ist*
ez	*jetzt*
de	*der*
genneraal	*general*

Normalisierung – Ziele und Richtlinien

Ziel: Zusätzliche Annotationsebene, um Formen zu gruppieren, die „dasselbe Wort“ darstellen

Transkription	Normalisierung
jaa	*ja*
de	*dann*
het	*hat*
me	*man*
no	*noch*
gluegt	*gelugt*
tänkt	*gedacht*
dasch	*das ist*
ez	*jetzt*
de	*der*
genneraal	*general*

Normalisierung – Ziele und Richtlinien

Ziel: Zusätzliche Annotationsebene, um Formen zu gruppieren, die „dasselbe Wort“ darstellen

Transkription	Normalisierung
jaa	*ja*
de	*dann*
het	*hat*
me	*man*
no	*noch*
gluegt	*gelugt*
tänkt	*gedacht*
dasch	*das ist*
ez	*jetzt*
de	*der*
genneraal	*general*

Normalisierung – Automatisierung

- 6 Dokumente wurden manuell normalisiert
- Aufwand: 30-60 Stunden pro Dokument
- Können wir die verbleibenden Texte automatisch normalisieren?

Lösung: Buchstabenbasierte maschinelle Übersetzung

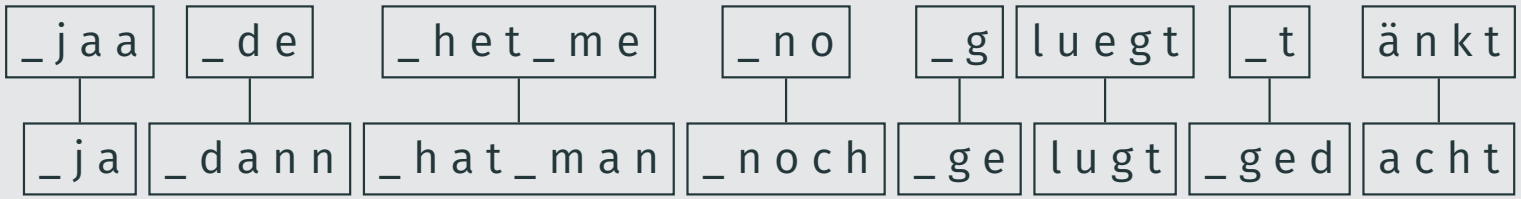
Beste Lösung (Scherrer & Ljubešić 2016):

- Kontext erweitert auf gesamte Äusserung
- Zusätzliches Sprachmodell: deutsche Filmuntertitel
- Eindeutige Wörter werden direkt kopiert
- 90.46% korrekte Normalisierung für vergleichbare Texte

Normalisierung – Automatisierung

- 6 Dokumente wurden manuell normalisiert
- Aufwand: 30-60 Stunden pro Dokument
- Können wir die verbleibenden Texte automatisch normalisieren?

Lösung: Buchstabenbasierte maschinelle Übersetzung



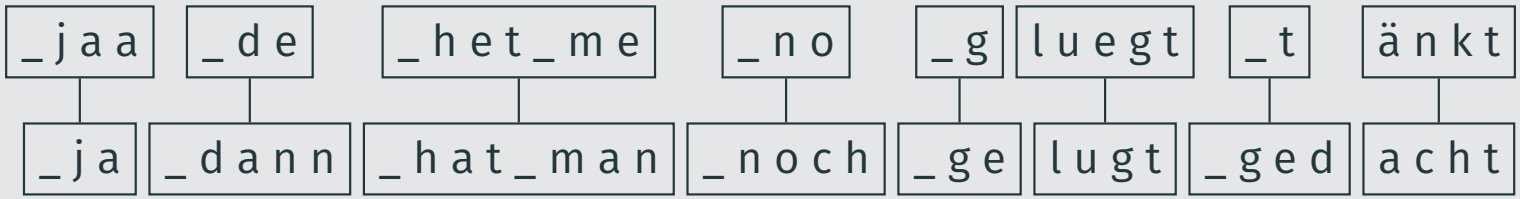
Beste Lösung (Scherrer & Ljubešić 2016):

- Kontext erweitert auf gesamte Äusserung
- Zusätzliches Sprachmodell: deutsche Filmuntertitel
- Eindeutige Wörter werden direkt kopiert
- 90.46% korrekte Normalisierung für vergleichbare Texte

Normalisierung – Automatisierung

- 6 Dokumente wurden manuell normalisiert
- Aufwand: 30-60 Stunden pro Dokument
- Können wir die verbleibenden Texte automatisch normalisieren?

Lösung: Buchstabenbasierte maschinelle Übersetzung



Beste Lösung (Scherrer & Ljubešić 2016):

- Kontext erweitert auf gesamte Äusserung
- Zusätzliches Sprachmodell: deutsche Filmuntertitel
- Eindeutige Wörter werden direkt kopiert
- 90.46% korrekte Normalisierung für vergleichbare Texte

Tagging

Tagging

Transkr.	Norm.	Tagging
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

1. Basistagger
   - Trainiert mit *NOAH's Corpus* (Hollenstein & Aepli 2014)
   - Geschriebenes Schweizerdeutsch
   - STTS+-Tagset
2. Bootstrapping
   - Einzelne ArchiMob-Dokumente werden getaggt, manuell korrigiert und hinzugefügt
   - 90.09% korrekte Annotation nach 4 Runden

Tagging

Transkr.	Norm.	Tagging
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

1. Basistagger
   - Trainiert mit *NOAH's Corpus* (Hollenstein & Aepli 2014)
   - Geschriebenes Schweizerdeutsch
   - STTS+-Tagset
2. Bootstrapping
   - Einzelne ArchiMob-Dokumente werden getaggt, manuell korrigiert und hinzugefügt
   - 90.09% korrekte Annotation nach 4 Runden

Tagging

Transkr.	Norm.	Tagging
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

1. Basistagger
    - Trainiert mit *NOAH's Corpus* (Hollenstein & Aepli 2014)
    - Geschriebenes Schweizerdeutsch
    - STTS+-Tagset
2. Bootstrapping
    - Einzelne ArchiMob-Dokumente werden getaggt, manuell korrigiert und hinzugefügt
    - 90.09% korrekte Annotation nach 4 Runden

Verfügbarkeit

Verfügbarkeit – Daten im XML-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">genneraal</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>
```

Audiodaten auf Anfrage

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

Verfügbarkeit – Korpusanalysewerkzeuge

Die ArchiMob-Daten sind auf SketchEngine und ANNIS durchsuchbar (Links auf Projektseite).

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 gsïi ùnd das (isch an)	üüs	/uns	überggangen äigentlich / ich ha düch 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 gsïi 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

Anwendung 1: Dialektale Variation

Dialektale Variation

Phonologische dialektale Variation:
Ein Graphem auf Normalisierungsebene, aber verschiedene Grapheme auf Transkriptionsebene

Beispiel:
Aus welchen Transkriptionen entsteht normalisiertes *ck*?

Transkr. → Norm.	Dokument 1	Dokument 2
k → ck	37.0%	95.2%
gg → ck	63.0%	2.4%
ch → ck	—	2.4%

- Kartierung aller Dokumente für eine Variante
- Vergleich mit SDS-Karte

Dialektale Variation

Phonologische dialektale Variation:
Ein Graphem auf Normalisierungsebene, aber verschiedene Grapheme auf Transkriptionsebene

Beispiel:
Aus welchen Transkriptionen entsteht normalisiertes *ck*?

Transkr. → Norm.	Dokument 1	Dokument 2
k → ck	37.0%	95.2%
gg → ck	63.0%	2.4%
ch → ck	—	2.4%

- Kartierung aller Dokumente für eine Variante
- Vergleich mit SDS-Karte

Dialektale Variation

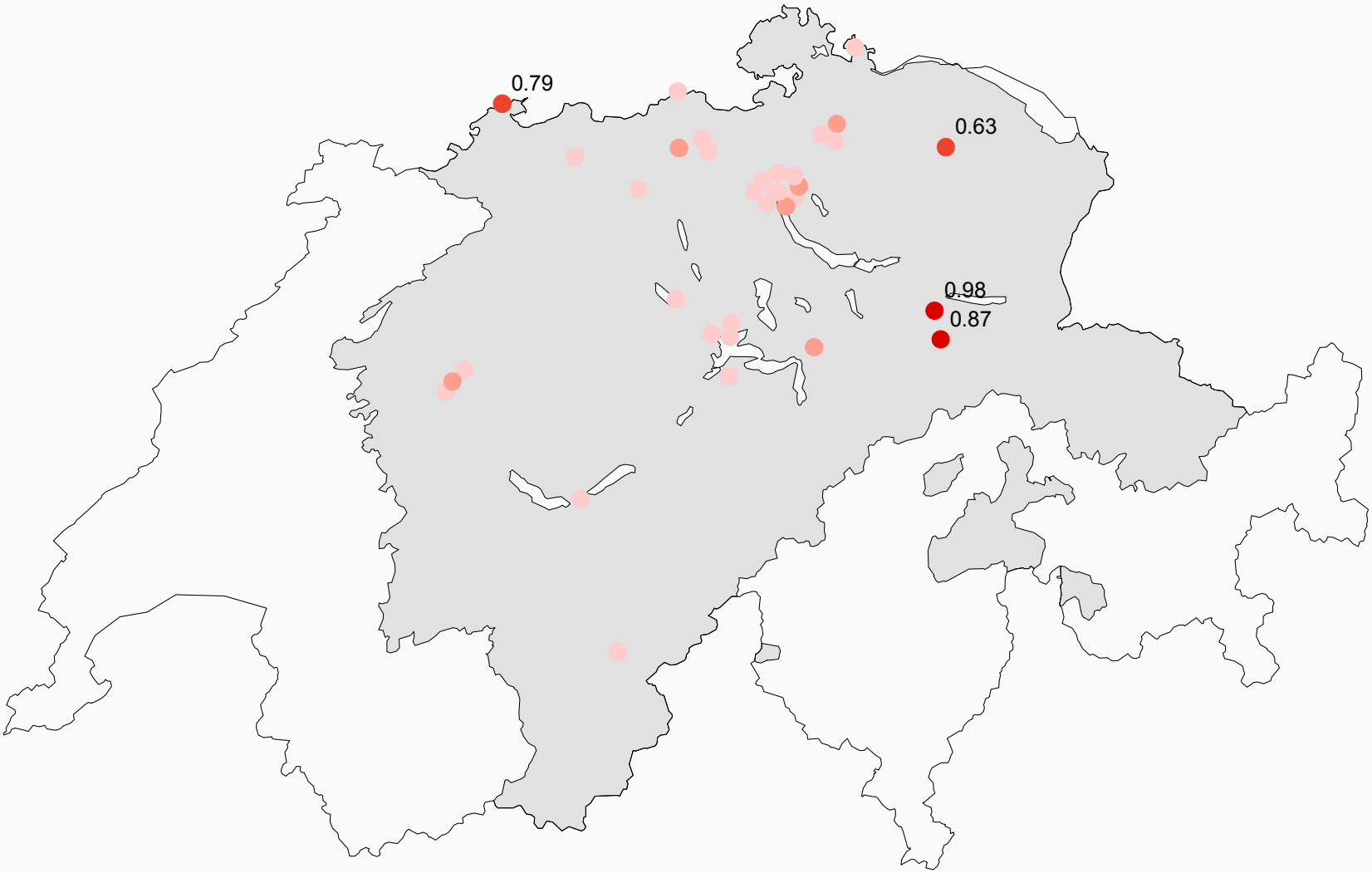
Phonologische dialektale Variation:
Ein Graphem auf Normalisierungsebene, aber verschiedene Grapheme auf Transkriptionsebene

Beispiel:
Aus welchen Transkriptionen entsteht normalisiertes *ck*?

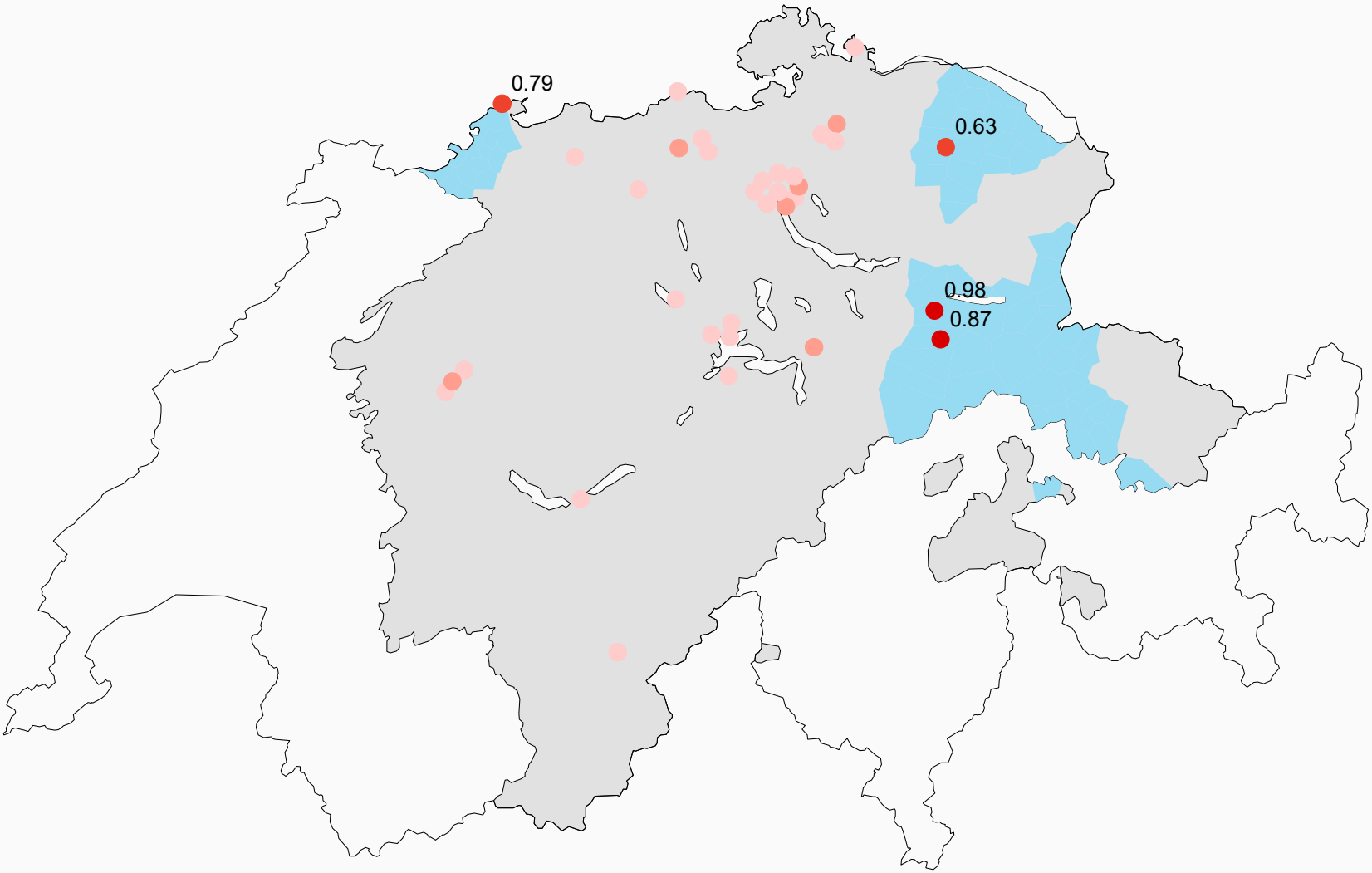
Transkr. → Norm.	Dokument 1	Dokument 2
k → ck	37.0%	95.2%
gg → ck	63.0%	2.4%
ch → ck	—	2.4%

- Kartierung aller Dokumente für eine Variante
- Vergleich mit SDS-Karte

Phonologische dialektale Variation: gg → ck

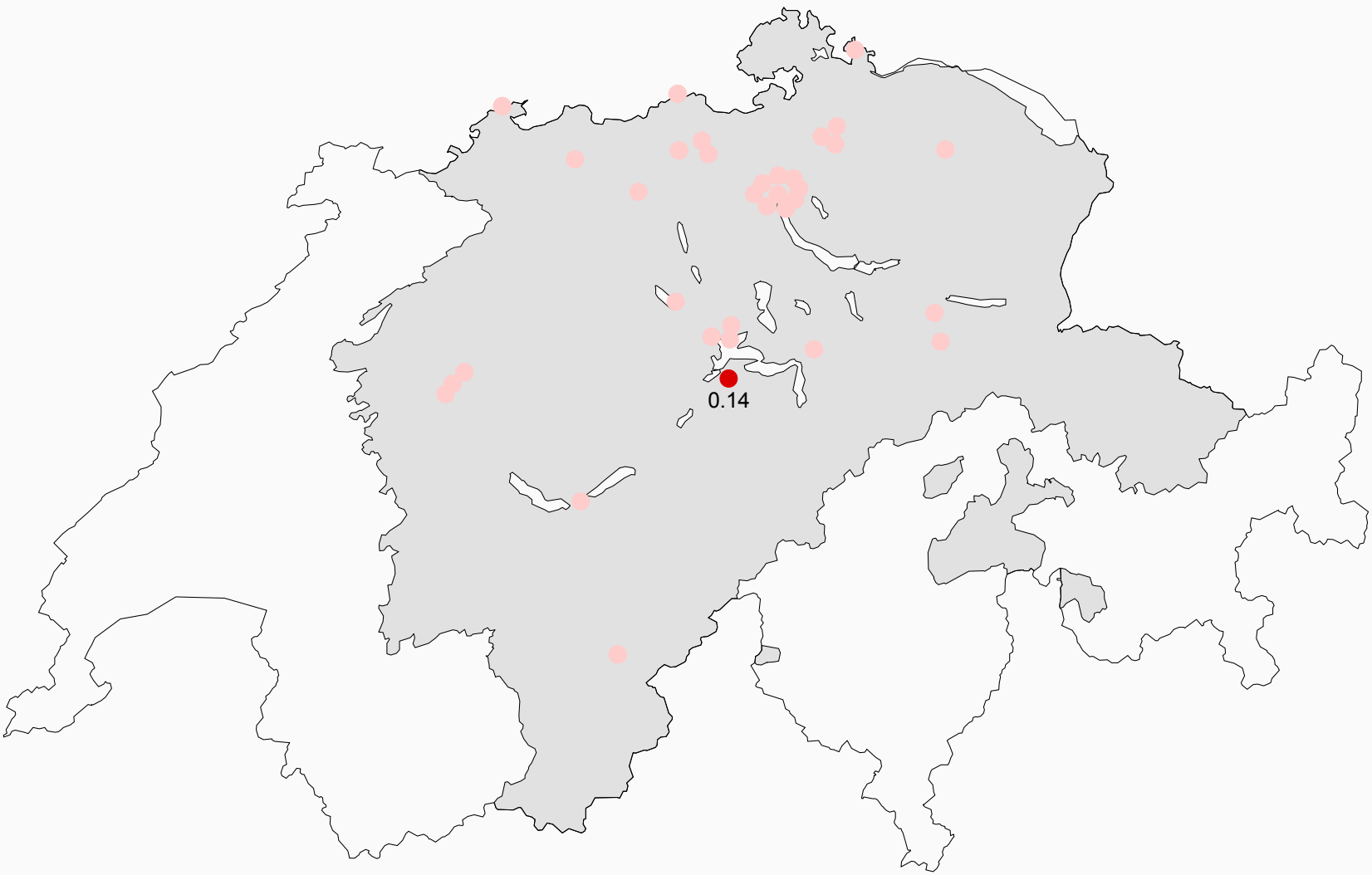


Phonologische dialektale Variation: gg → ck

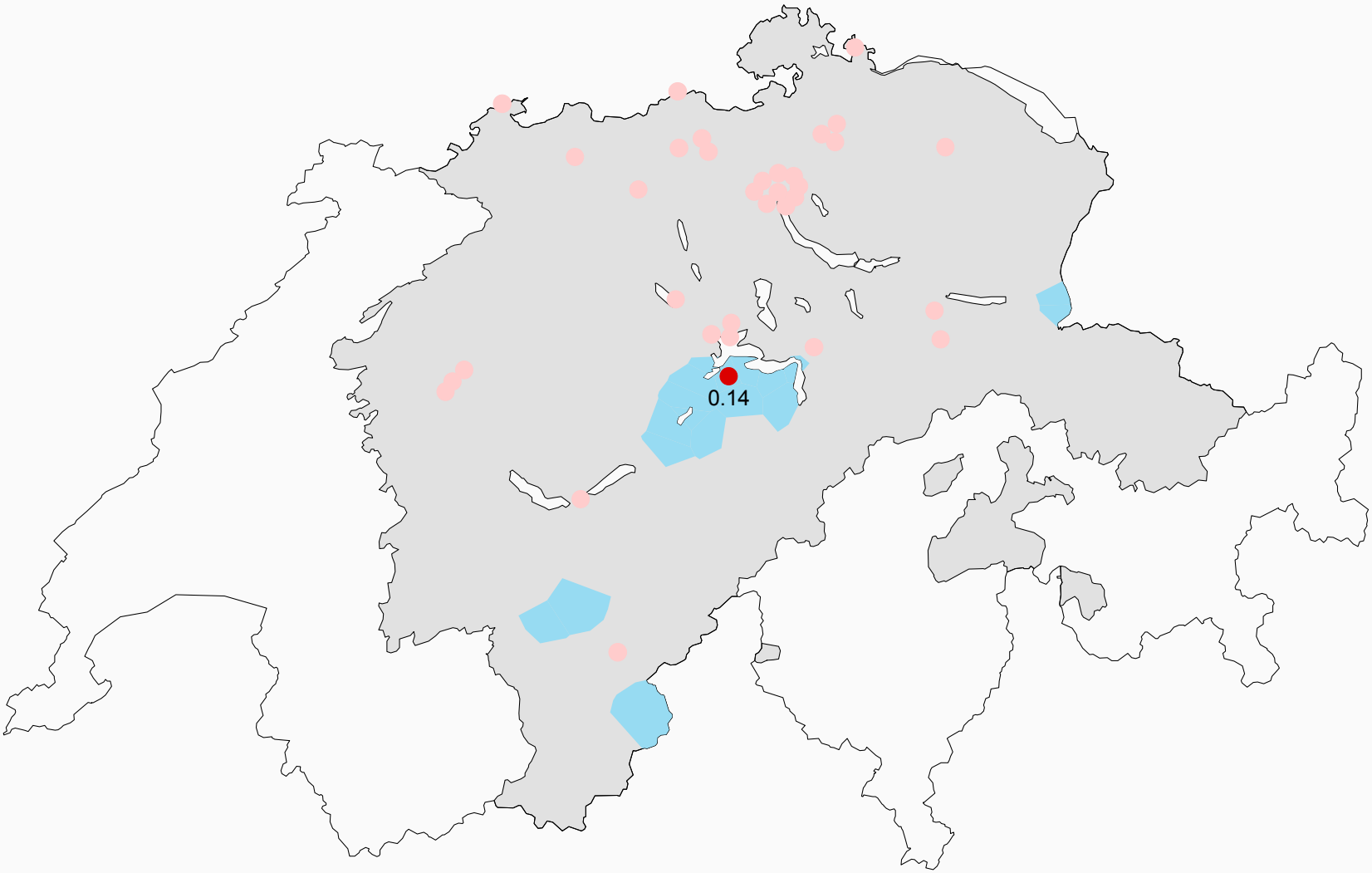


Blau: Verbreitungsgebiet der Variante *-gg-* in SDS 2/095 „drücken“

Phonologische dialektale Variation: ui → au

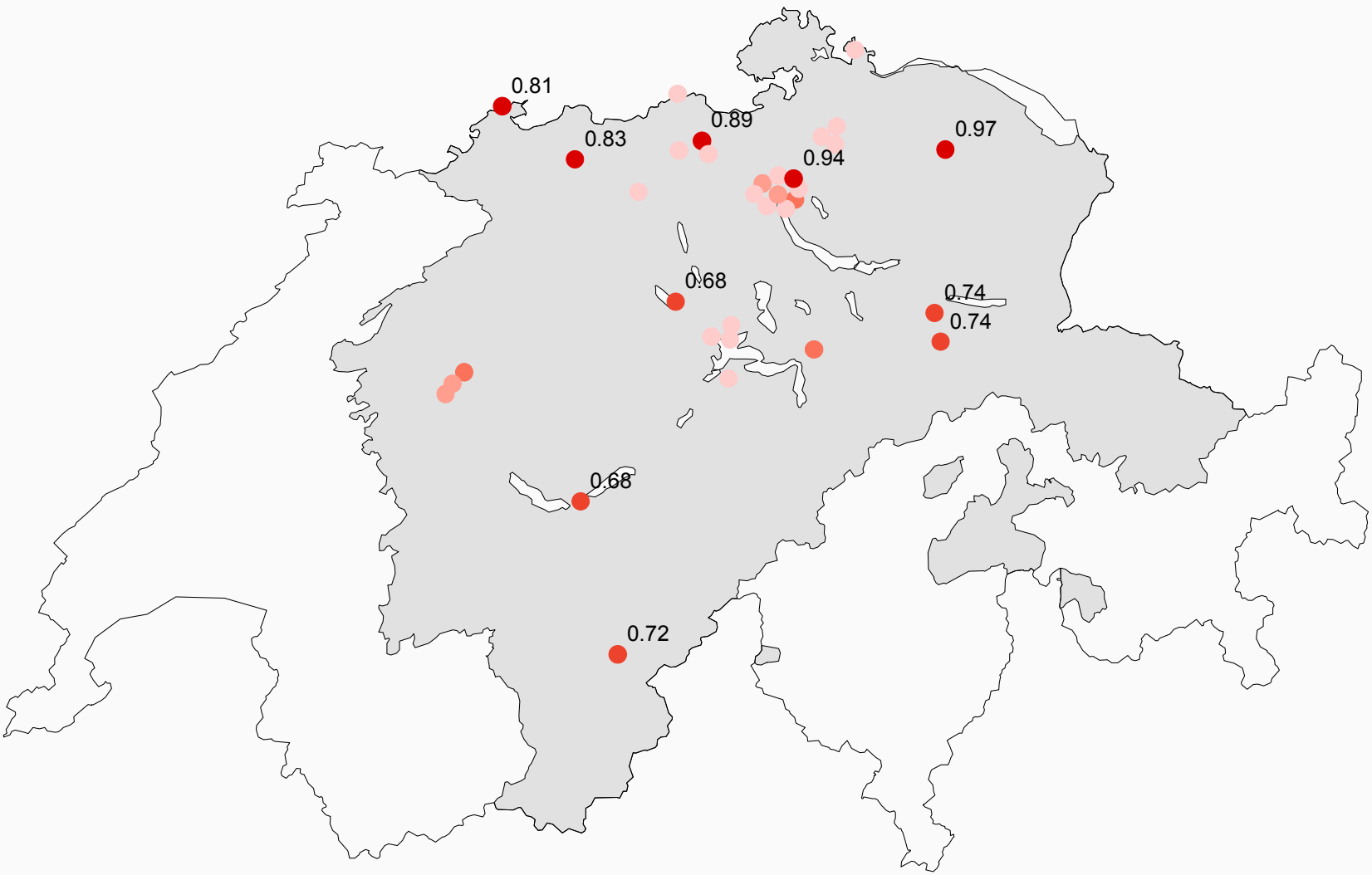


Phonologische dialektale Variation: ui → au

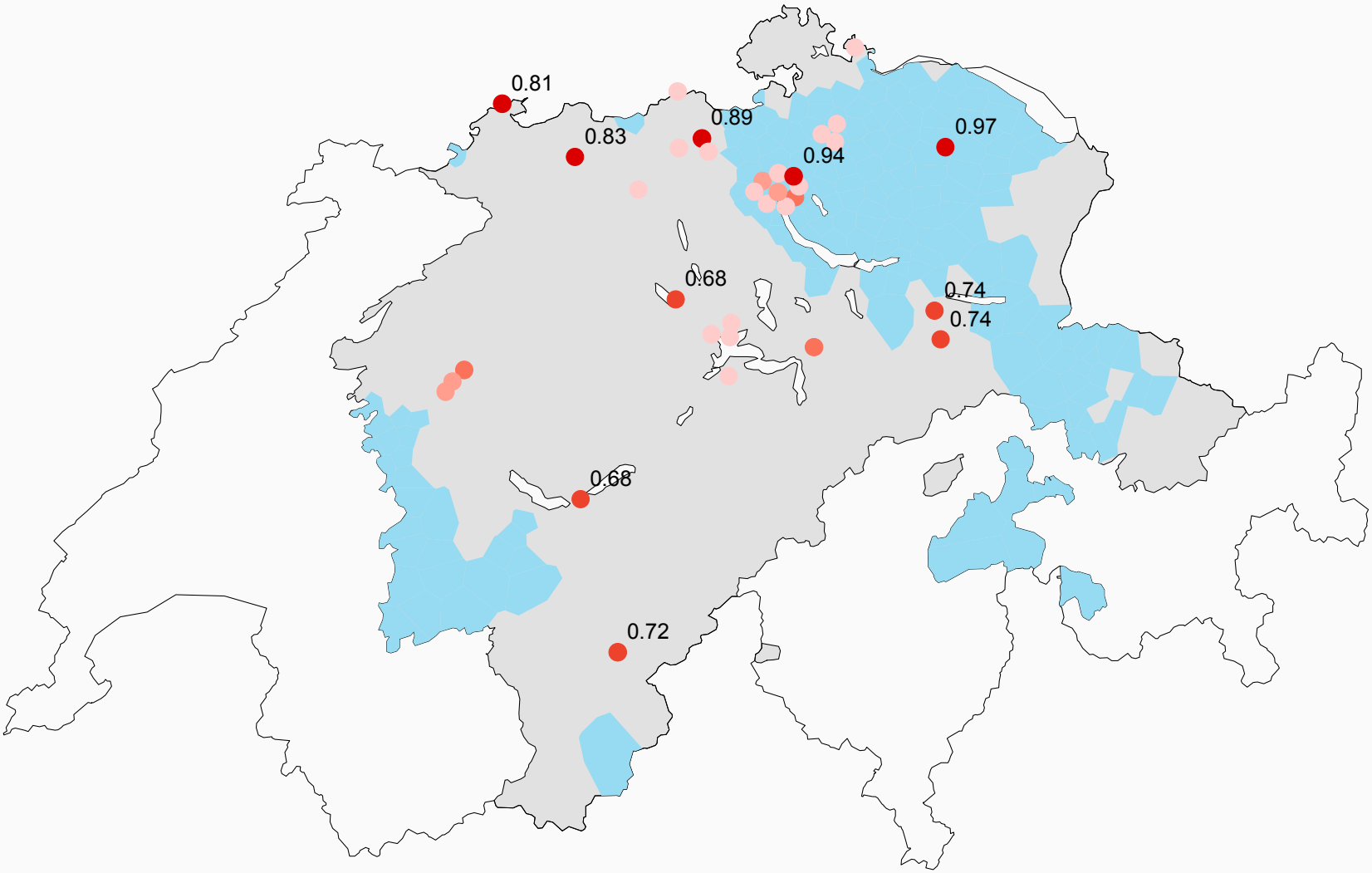


Blau: Verbreitungsgebiet der Variante *Muis* in SDS 1/106 „Maus“

Phonologische dialektale Variation: n → nn



Phonologische dialektale Variation: n → nn



Blau: Verbreitungsgebiet der Variante *Tane* in SDS 2/179 „Tanne“

Anwendung 2: Dialektidentifizierung

Dialektidentifizierung

VarDial Evaluation Campaign on Similar Languages, Varieties and Dialects

- Satz/Absatz aus Korpus $\rightarrow$ Sprache/Dialekt
- **Ähnliche Sprachen:** bosnisch/kroatisch/serbisch, ...
- **Arabische Dialekte:** Ägypten, Nordafrika, Levante, Golf
- **Schweizerdeutsche Dialekte:** Basel, Bern, Luzern, Zürich

Beispiele

- u simer geng a d landi öppe zwöi drüümau
- unsere suntigschbaziirgang hètis
- da hät s dänn amel eson en linsebrei ggää
- de vatter hets natüürli glii gmèrkt ùnd dä ìsch ainisch choo ùnd nòchhär nümme
- daa bi wind und wätter im schnee

Dialektidentifizierung

VarDial Evaluation Campaign on Similar Languages, Varieties and Dialects

- Satz/Absatz aus Korpus $\rightarrow$ Sprache/Dialekt
- **Ähnliche Sprachen:** bosnisch/kroatisch/serbisch, …
- **Arabische Dialekte:** Ägypten, Nordafrika, Levante, Golf
- **Schweizerdeutsche Dialekte:** Basel, Bern, Luzern, Zürich

Beispiele

- u simer geng a d landi öppe zwöi drüümau
- unsere suntigschbaziirgang hètis
- da hät s dänn amel eson en linsebrei ggää
- de vatter hets natüürli glii gmèrkt ùnd dä ìsch ainisch choo ùnd nòchhär nümme
- daa bi wind und wätter im schnee

Dialektidentifizierung

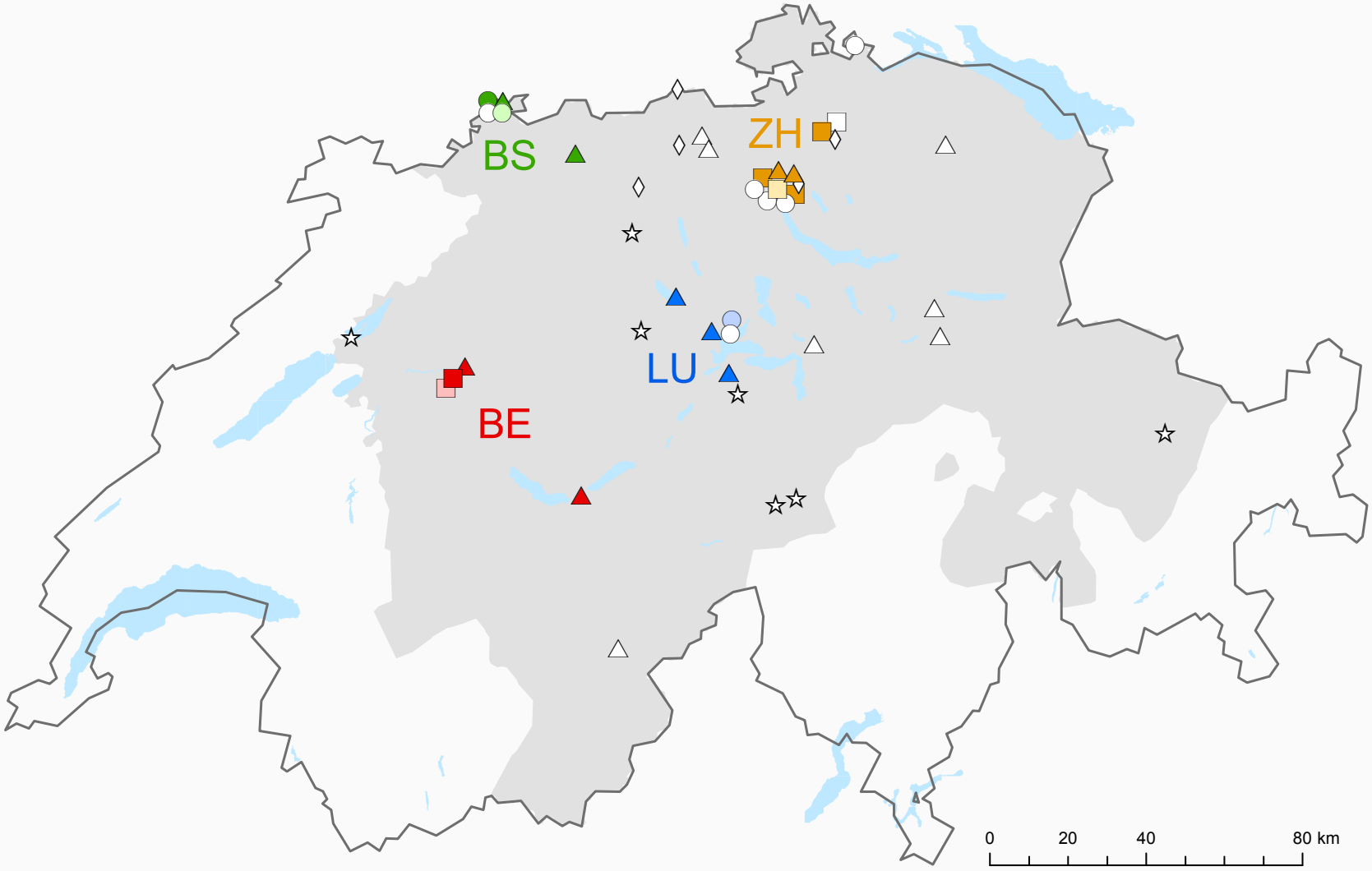
VarDial Evaluation Campaign on Similar Languages, Varieties and Dialects

- Satz/Absatz aus Korpus $\rightarrow$ Sprache/Dialekt
- **Ähnliche Sprachen:** bosnisch/kroatisch/serbisch, ...
- **Arabische Dialekte:** Ägypten, Nordafrika, Levante, Golf
- **Schweizerdeutsche Dialekte:** Basel, Bern, Luzern, Zürich

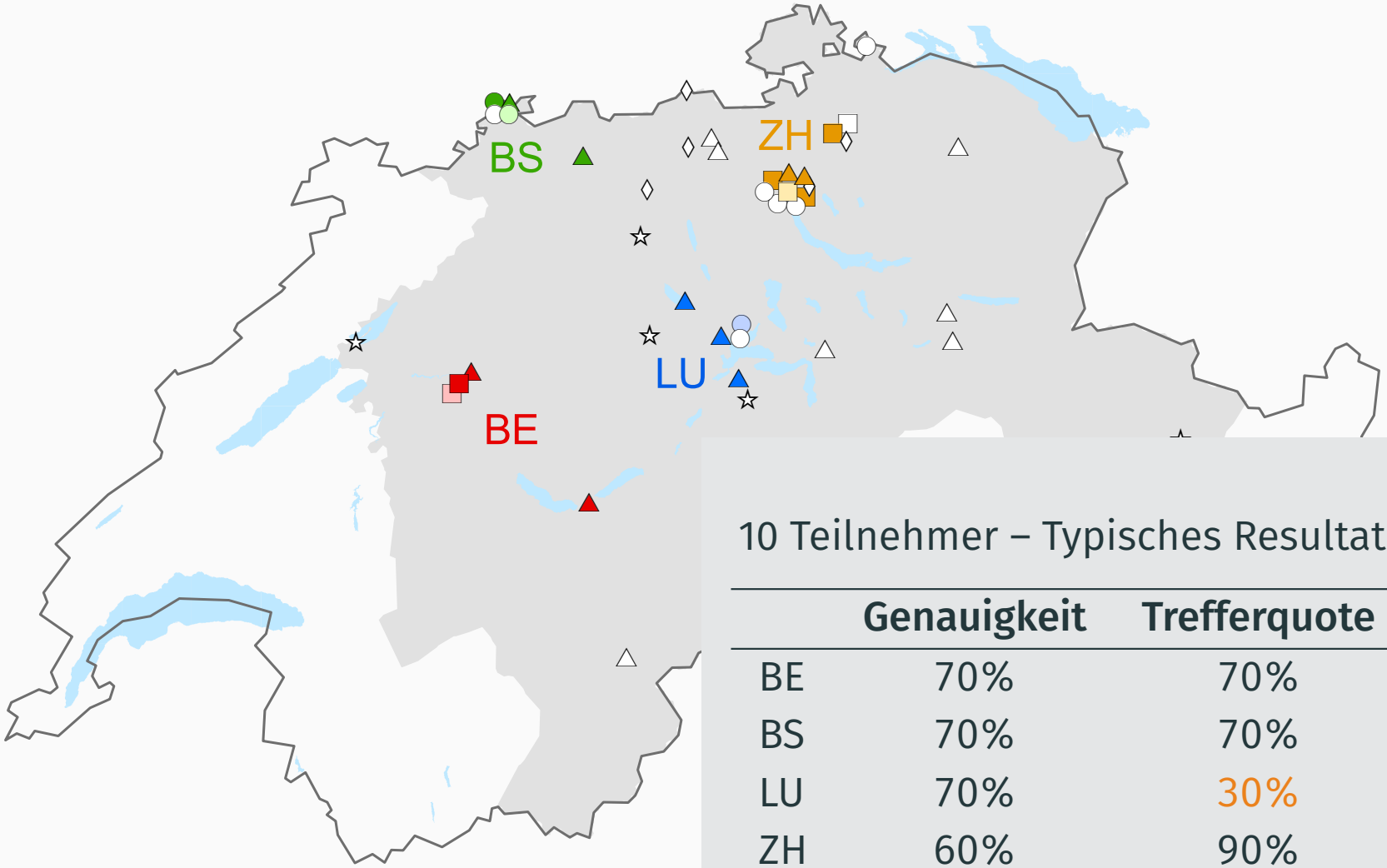
Beispiele

- u simer geng a d landi öppe zwöi drüümau
- unsere suntigschbaziirgang hètis
- da hät s dänn amel eson en linsebrei ggää
- de vatter hets natüürli glii gmèrkt ùnd dä ìsch ainisch choo ùnd nòchhär nümme
- daa bi wind und wätter im schnee

Dialektidentifizierung – Daten



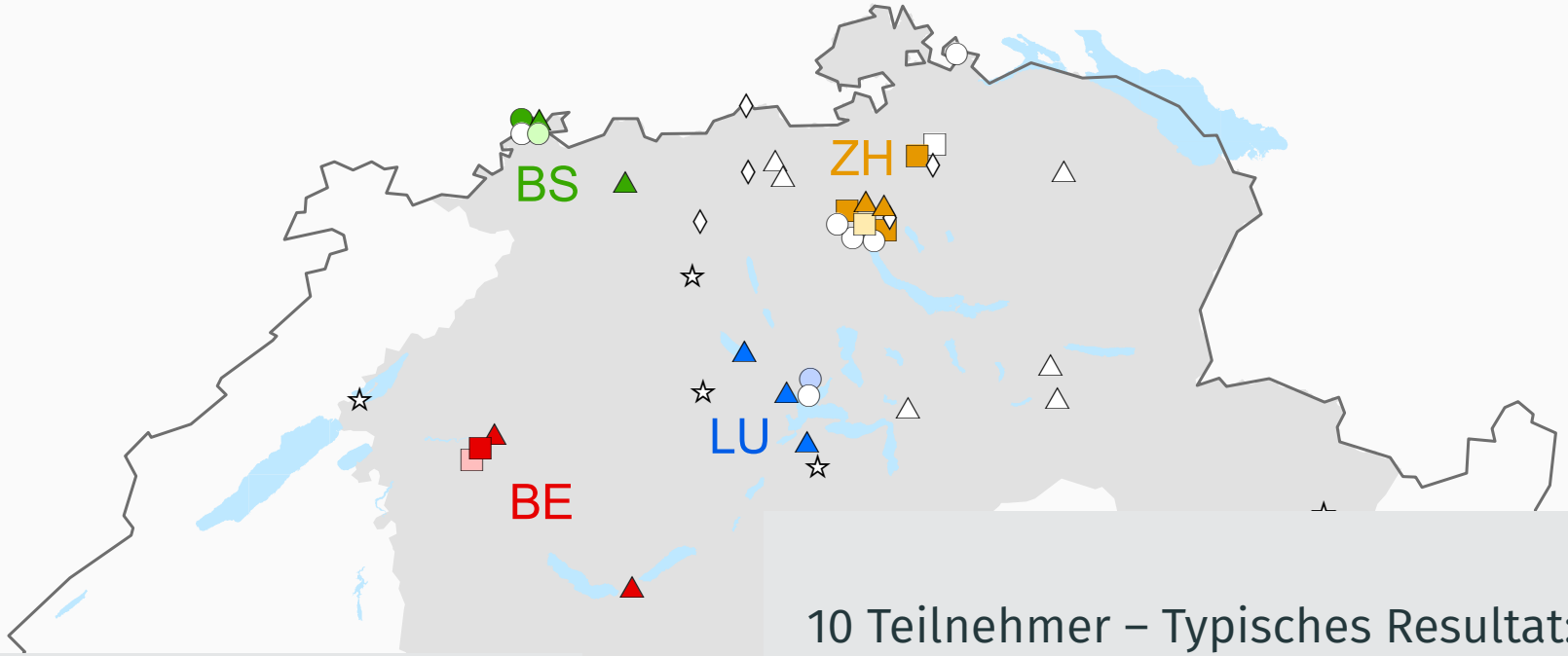
Dialektidentifizierung – Daten



10 Teilnehmer – Typisches Resultat:

	Genauigkeit	Trefferquote
BE	70%	70%
BS	70%	70%
LU	70%	30%
ZH	60%	90%

Dialektidentifizierung – Daten



- Nur 30% aller LU-Instanzen wurden tatsächlich als solche identifiziert
- Transkriptoreneffekte stärker als angenommen

10 Teilnehmer – Typisches Resultat:

	Genauigkeit	Trefferquote
BE	70%	70%
BS	70%	70%
LU	70%	30%
ZH	60%	90%

Anwendung 3: Textklassifizierung

Textklassifizierung

Idee:

1. Distanzmatrix zwischen Dokumenten erstellen

Sprachmodelle

Für jedes Dokument wird ein Sprachmodell erstellt:

- Buchstaben-4-Gramme
- Basiert nur auf Transkriptionen, nicht auf Normalisierung

Ein Sprachmodell kann ein anderes Dokument bewerten:

- Perplexität $\approx$ Distanz

2. Herkömmliche Dialektometrie-Verfahren anwenden (z.B. Clustering)
3. Analyse: Dialektale Unterschiede? Transkriptoren-unterschiede? Vergleich mit Atlasdaten

Textklassifizierung

Idee:

1. Distanzmatrix zwischen Dokumenten erstellen

Sprachmodelle

Für jedes Dokument wird ein Sprachmodell erstellt:

- Buchstaben-4-Gramme
- Basiert nur auf Transkriptionen, nicht auf Normalisierung

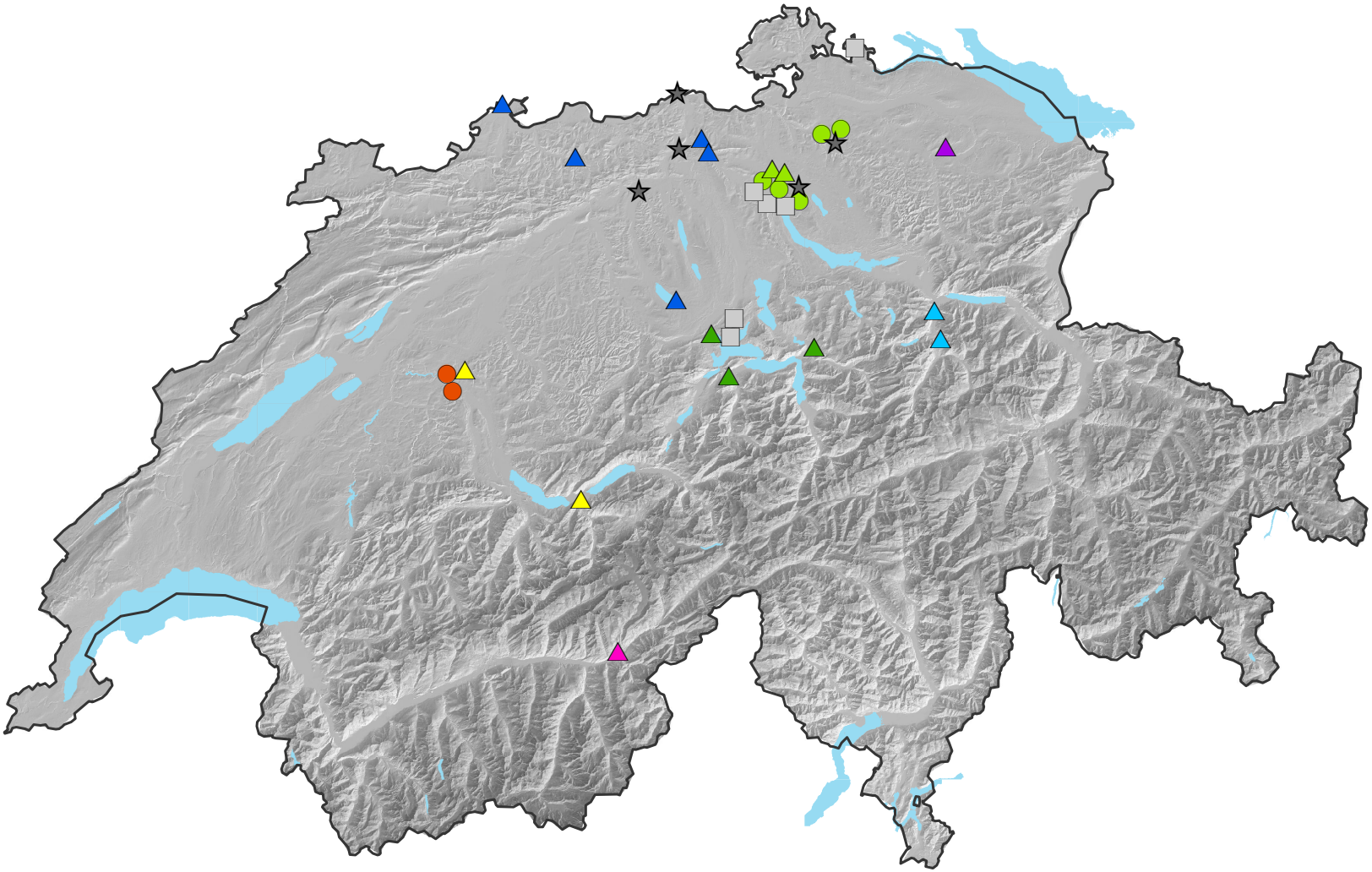
Ein Sprachmodell kann ein anderes Dokument bewerten:

- Perplexität $\approx$ Distanz

2. Herkömmliche Dialektometrie-Verfahren anwenden (z.B. Clustering)
3. Analyse: Dialektale Unterschiede? Transkriptoren-unterschiede? Vergleich mit Atlasdaten

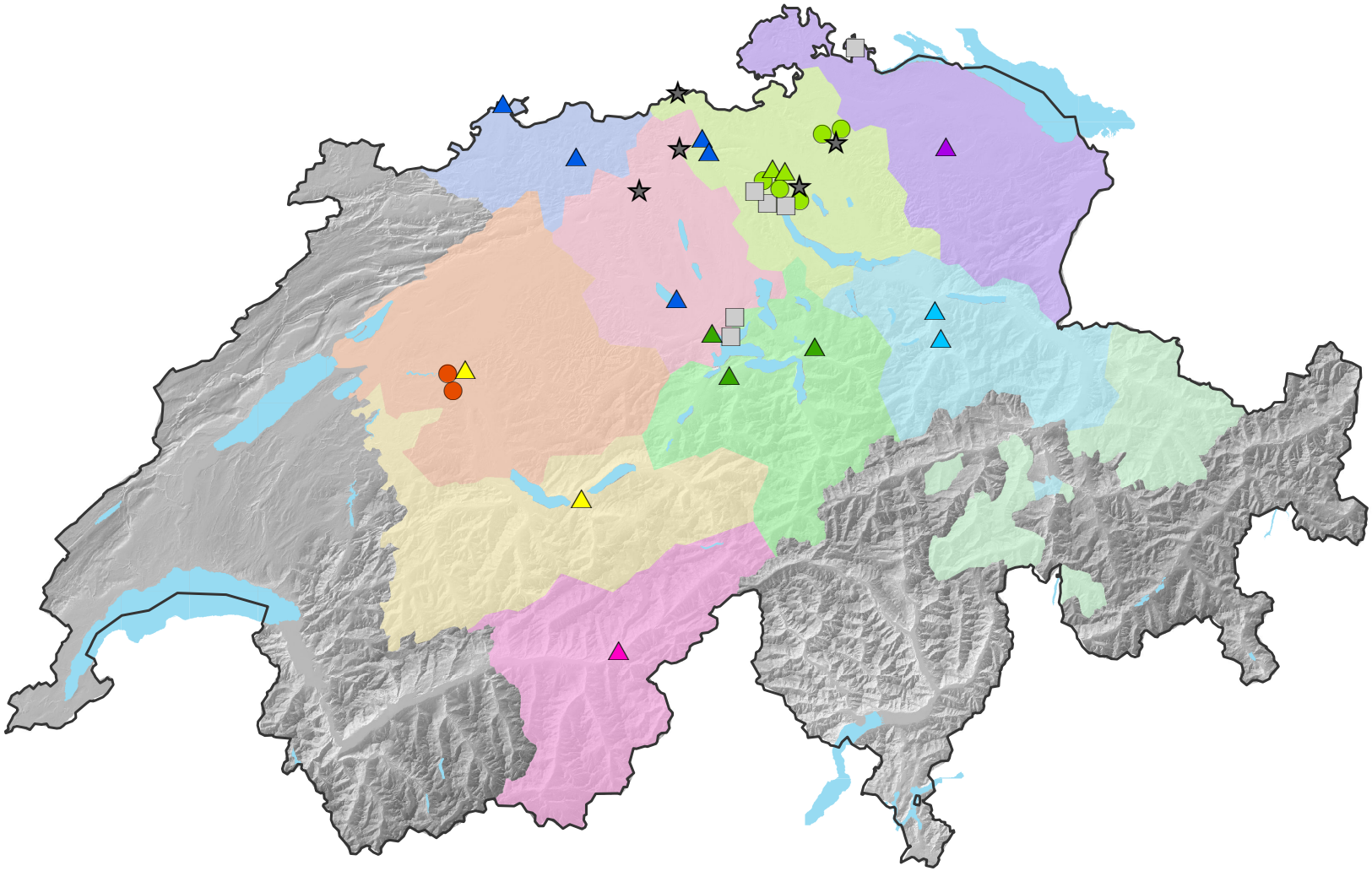
Textklassifizierung

Ward-Algorithmus, 10 Cluster

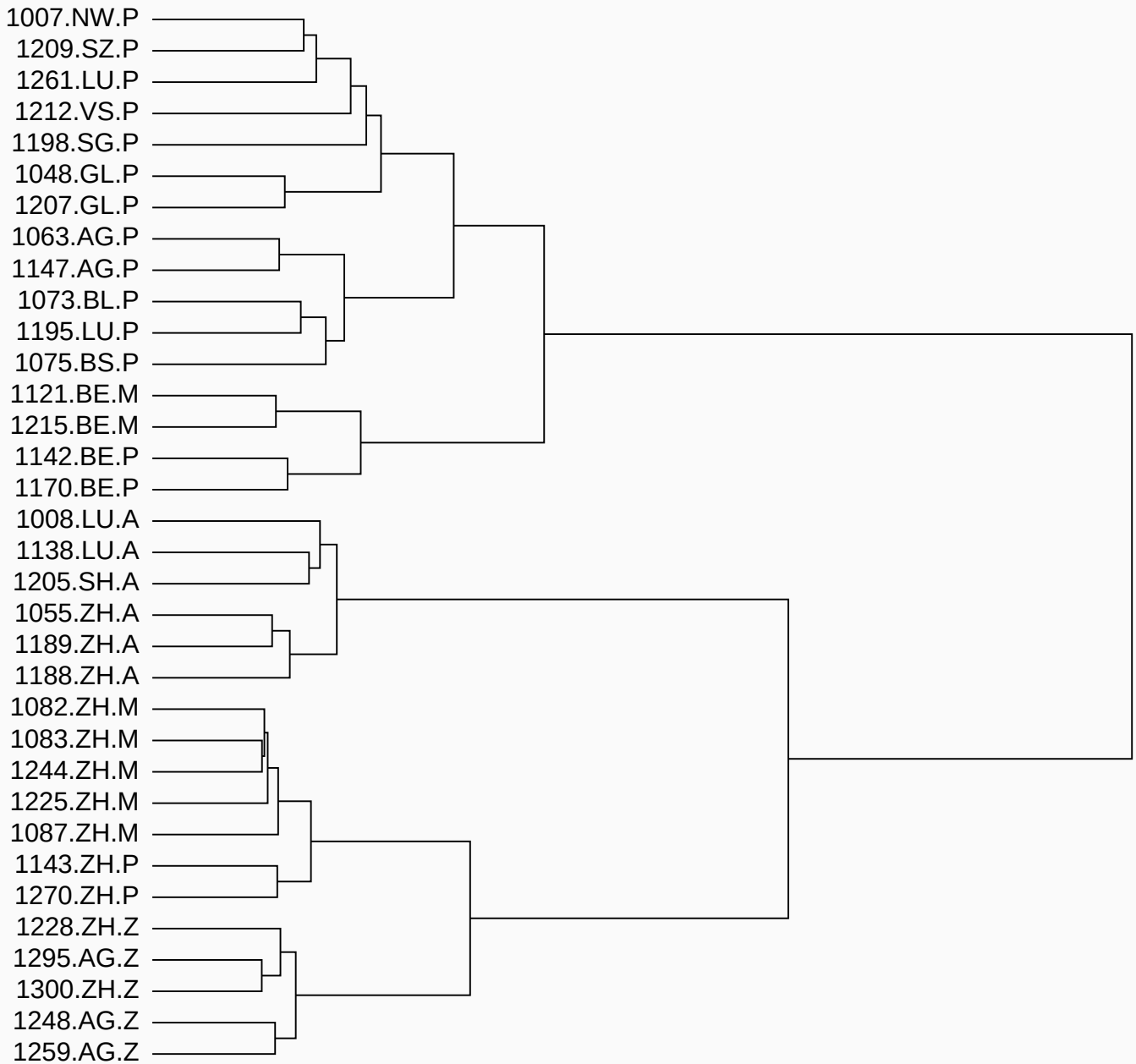


Textklassifizierung

Vergleich mit 10-Cluster-Lösung aus SDS- und SADS-Daten (Scherrer & Stoeckle 2016)



Textklassifizierung – Dendrogramm



Zusammenfassung

Zusammenfassung

Annotation: Textauswahl, Transkription, Normalisierung und Tagging

- Computerlinguistische Methoden und Werkzeuge für automatische Annotation

Anwendungen: Variationsmuster, Dialektidentifizierung, Dialektklassifizierung

- Basis: Transkription (+Normalisierung)
- Trotz starker Transkriptoreneffekte gute Grundlage für interessante dialektologische Fragestellungen

Verfügbarkeit: Release 1.0 online, Release 2.0 bald auch: `http://www.spur.uzh.ch/en/departments/korpuslab/ArchiMob.html`

Zusammenfassung

Annotation: Textauswahl, Transkription, Normalisierung und Tagging

- Computerlinguistische Methoden und Werkzeuge für automatische Annotation

Anwendungen: Variationsmuster, Dialektidentifizierung, Dialektklassifizierung

- Basis: Transkription (+Normalisierung)
- Trotz starker Transkriptoreneffekte gute Grundlage für interessante dialektologische Fragestellungen

Verfügbarkeit: Release 1.0 online, Release 2.0 bald auch: `http://www.spur.uzh.ch/en/departments/korpuslab/ArchiMob.html`

Zusammenfassung

Annotation: Textauswahl, Transkription, Normalisierung und Tagging

- Computerlinguistische Methoden und Werkzeuge für automatische Annotation

Anwendungen: Variationsmuster, Dialektidentifizierung, Dialektklassifizierung

- Basis: Transkription (+Normalisierung)
- Trotz starker Transkriptoreneffekte gute Grundlage für interessante dialektologische Fragestellungen

Verfügbarkeit: Release 1.0 online, Release 2.0 bald auch: `http://www.spur.uzh.ch/en/departments/korpuslab/ArchiMob.html`

Dank

Finanzierung:

Hasler-Stiftung Förderbeitrag Nr. 16038
UZH UFSP Sprache und Raum
Spitch

Mitarbeit:

Noëmi Aepli
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
Hanna Ruch
Beni Ruef
Fatima Stadler
Phillip Ströbel
Simone Ueberwasser
Alexandra Zoller