

Advances in computer-assisted historical reconstruction

Michael Cysouw
Philipps-Universität Marburg

Philipps



Universität
Marburg

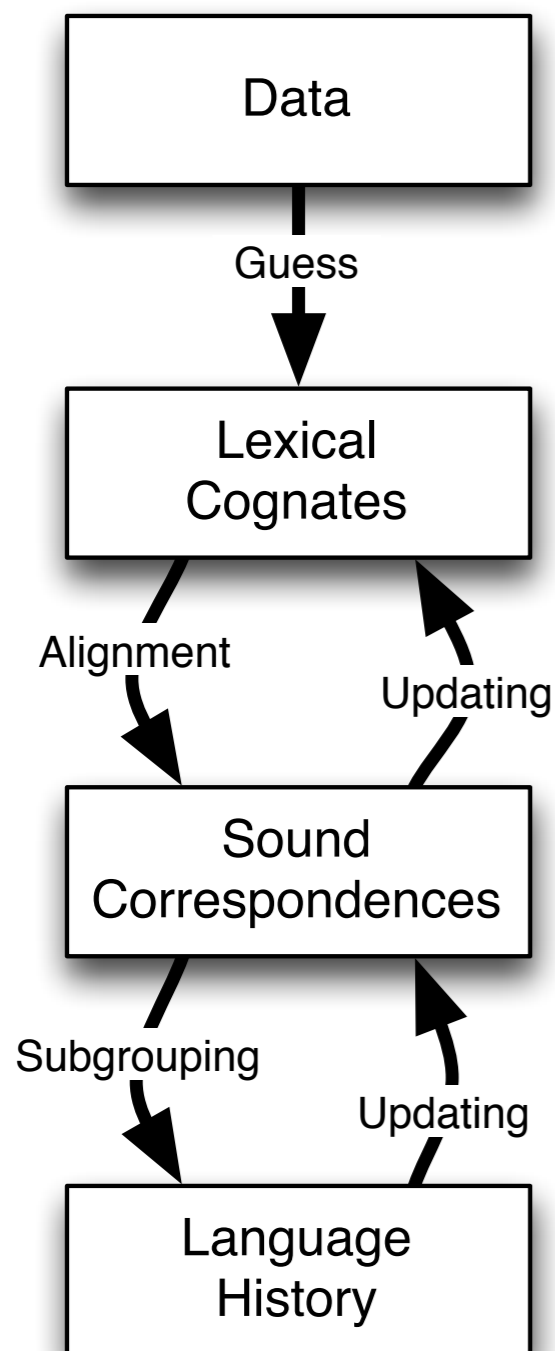
Preliminaries

- In this talk is about **reconstructing sound change**, but historical reconstruction includes of course many more phenomena!
- Focus on tools to **combine automatic quantitative approaches with manual decisions**
- Using a **multitude of smaller tools**, always based on simple CSV-type formats.
Interoperability is still not perfect

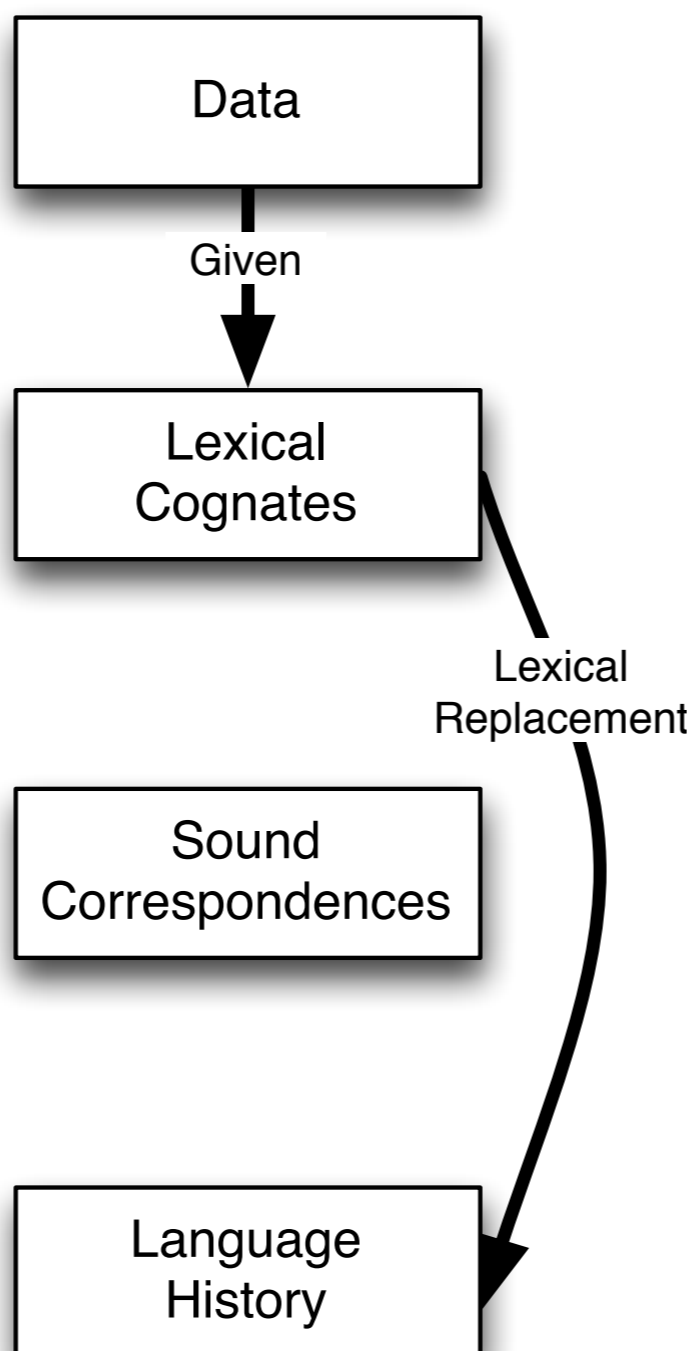
Background

- Growing interest (again) into quantitative historical linguistics
- Many recent approaches work like a ‘black box’ for ordinary historical linguists
- Although the computational approaches are often highly sophisticated, the linguistically interpretable output is meagre

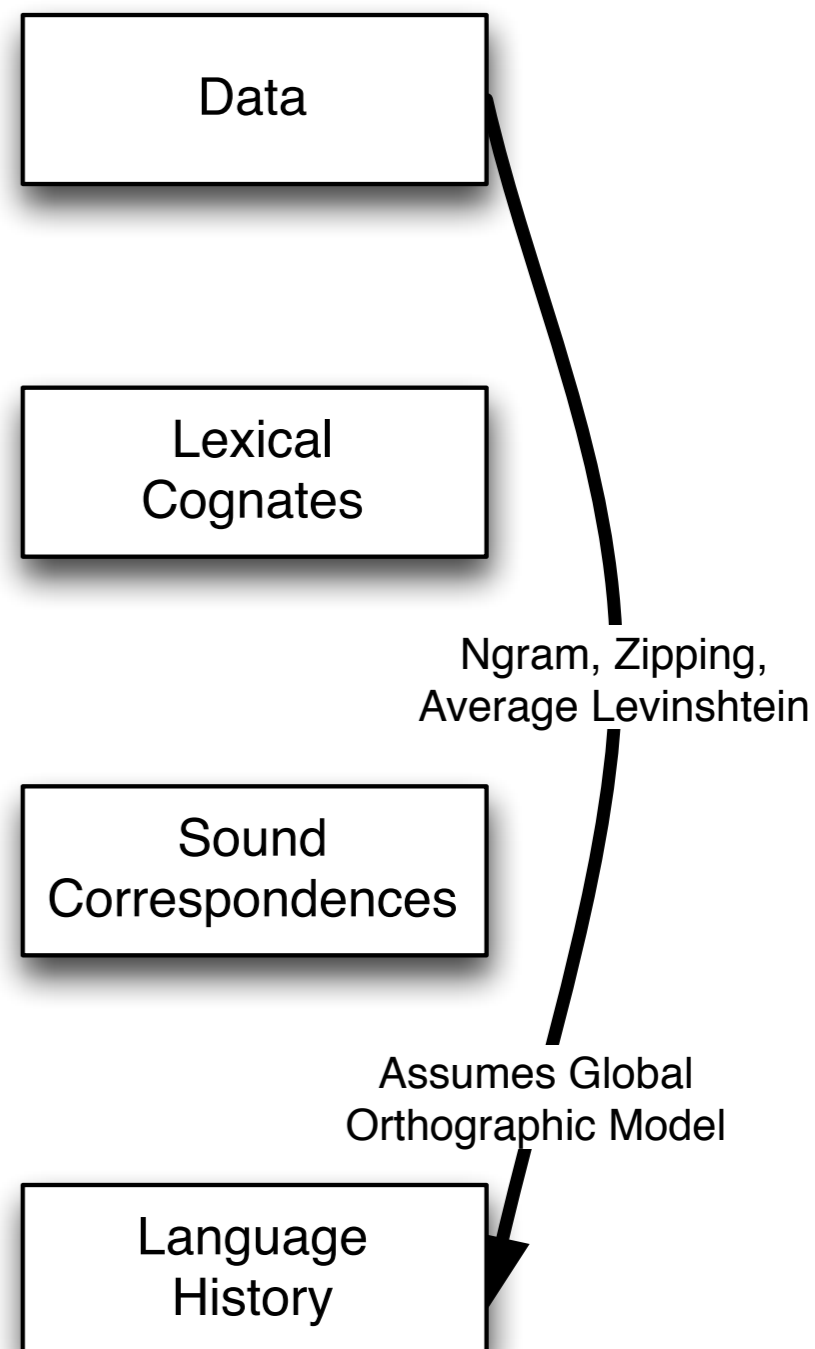
Comparative Method



Swadesh Method



Black Box Method



Computational pipeline for comparative method

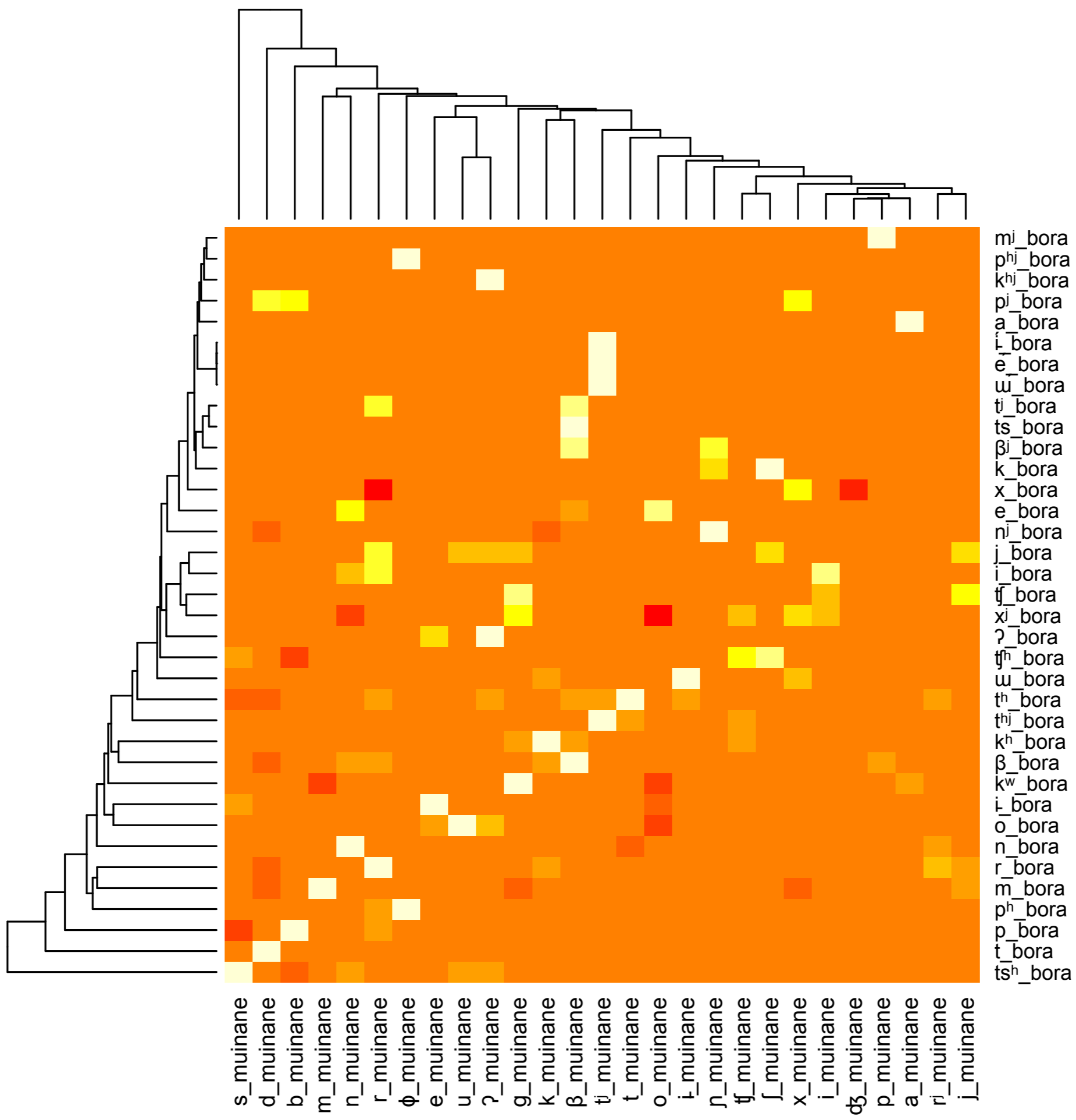
1. Cognate guessing using regularity
2. Alignment of sounds (“correspondences”)
3. Clustering of correspondence sets
4. Sound change modelling

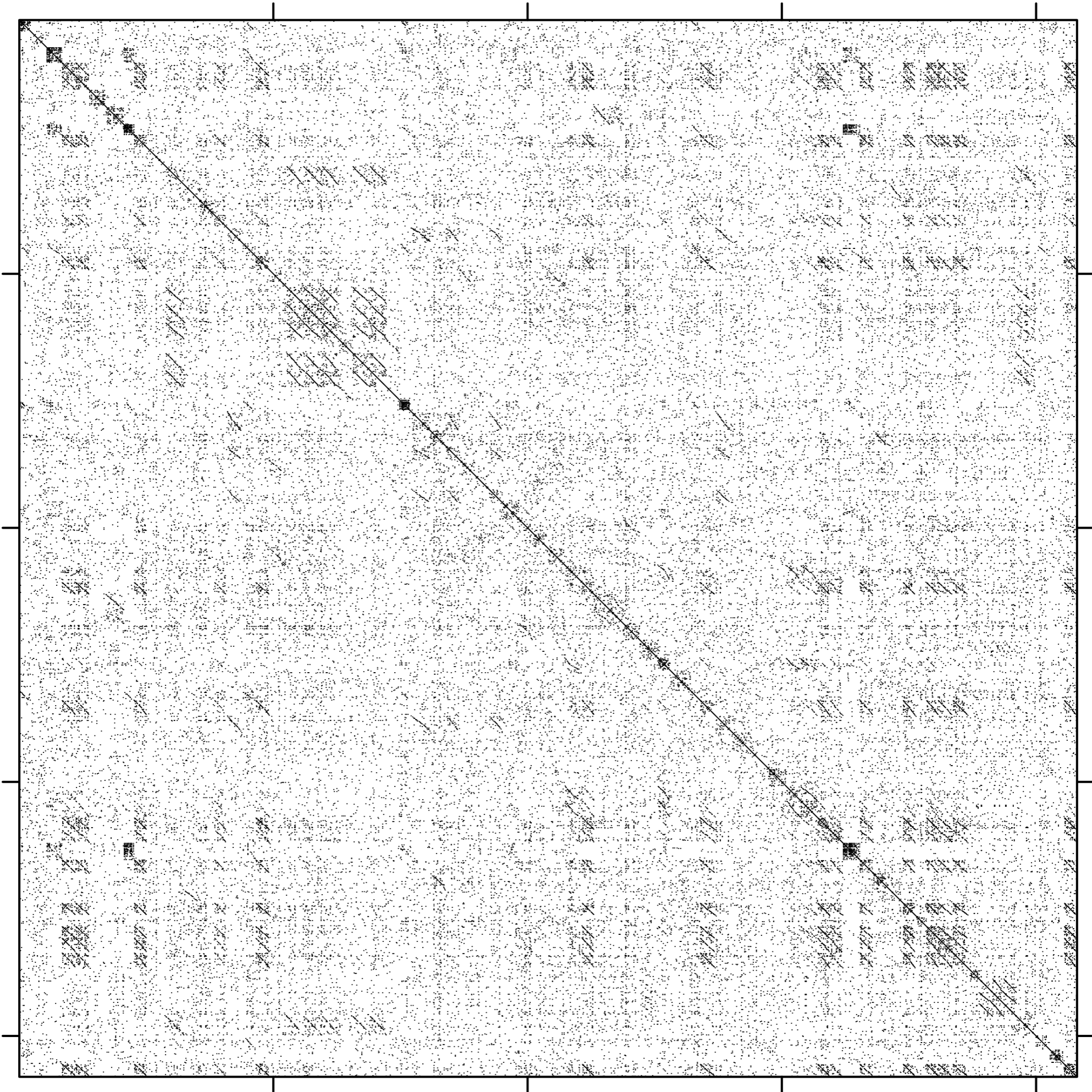
Open-source Tools

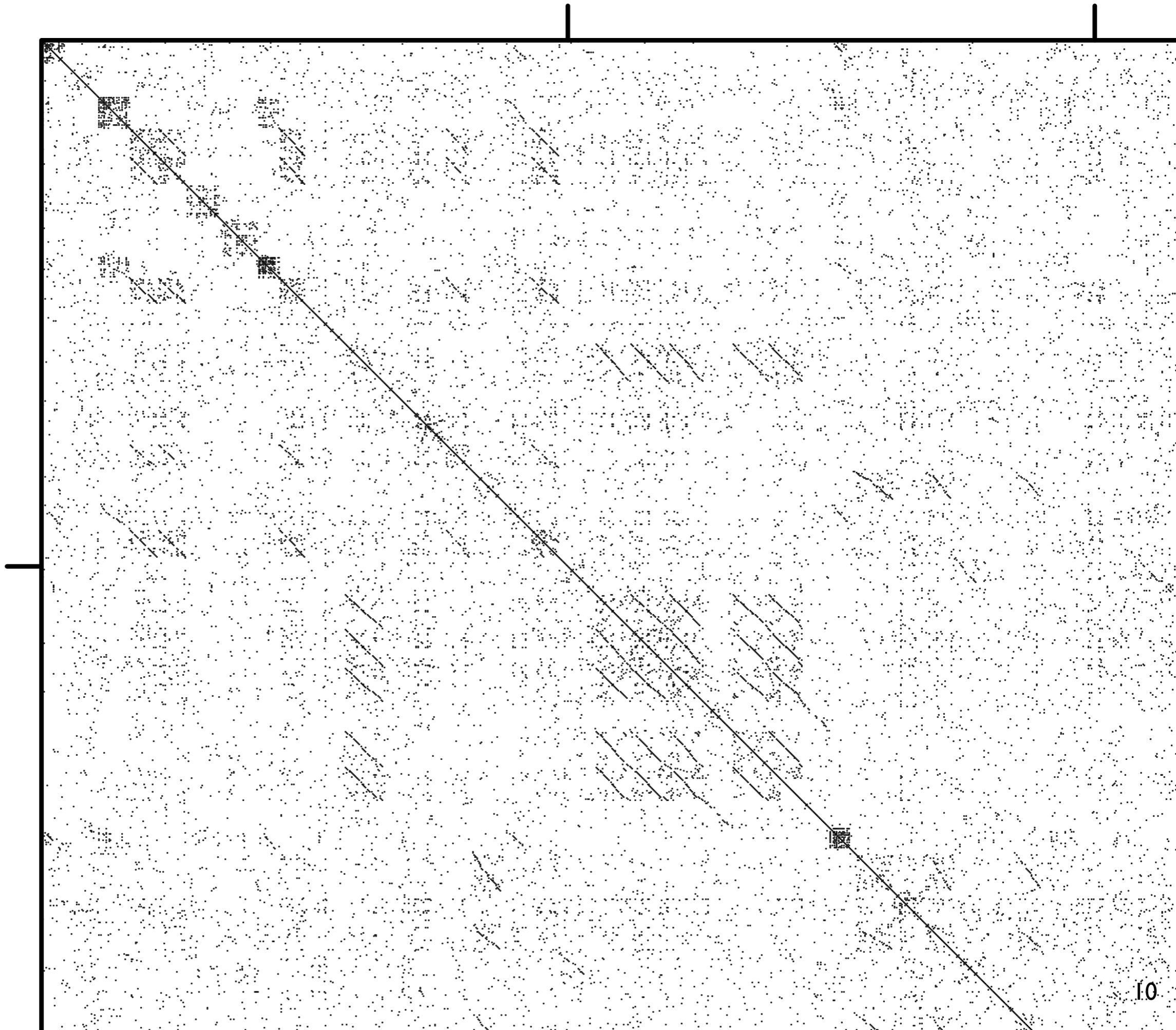
- **Orthography Processing** (with Steven Moran)
github.com/cysouw/qlcTokenize (R, also available in Python)
- **Guessing regular correspondences:**
github.com/cysouw/qlcMatrix (R)
- **Alignments (and more!) with LingPy** (Johann-Mattis List)
github.com/lingpy (Python)
- **Alignment Editor** (with Frank Nagel)
github.com/digitallinguists/msa-editor (Javascript in Browser)
- **Visualisation of Alignments**
github.com/cysouw/qlcVisualize (R, planned in Javascript)

Step 1: Guessing by regularity

- Given some dictionaries, we can produce lists of words with the same meaning
- Are there cognates in such a list?
Not just lookalikes!
- There is much noise, but the same sounds correspondences should occur regularly
- Sparse Matrix Algebra in R:
github.com/cysouw/glcMatrix







Step 2: Alignment

Lfd. Nr.	Stichwort		Transkription	Bemerkungen
1	Winter	178	'vɪntə ^h ʁ	ʁ = kontinuierlich ə bereits velarisierter
* 2	fliegen	56	'flaɪ→ə ^h tə	'fliegen die', Segmentierung unklar - kein geminiertes [t]
3	Blätter	23	'blɛ:də ^h ʁ	
4	Luft	103	lʊ ^h ft ^h	ʁ = kontinuierlich
5	hört	89	hɪ ^h t ^h	ʁ = kontinuierlich
6	gleich	78	ˌklɛ ₊ ɪk _~	folgt P
7	schneien	130	'ʃnɛɪən	
8	Wetter	174	/	statt dem 'vɪtəʁɔgə
9	tu	151	də→v	

Ort der Mundart/Kreis Astfeld/ Gronau	Aufnahme-Nr. I/62	Transkribent Angelika Braun	Listentyp A
---	----------------------	--------------------------------	----------------

Phonetischer Atlas von Deutschland

- Wenker-sentences recorded in the 1960s (with additions in the 1970s)
- Selected words from the recordings were transcribed on paper in the 1980s
- A joint project between Marburg and Groningen digitized the data in the 2000s
- In total 29530 words distributed over 183 locations and 186 cognate sets

Lfd. Nr.	Stichwort	Transkription	Bemerkungen
1	Winter	178 vɪntɐ	b = kontinuant vɛrɛɪ̯k vɛlɑtɪsɪs
2	fliegen	56 flaɪ̯ -> aɪ̯	'fliegen die', Sequenzierung unklar - kein gemischtes [tʃ]
3	Blatter	23 blɑtɐ	
4	Luft	103 lʊft	= kontinuant
5	hört	89 hœrt	b = kontinuant
6	gleich	130 glɛɪ̯ç	
7	schneien	130 ʃnɛɪ̯ən	
8	Wetter	178 vɛtɐ	statt dem 'vɛtəʀɔgə
9	tu	151 dɛv	

Ort der Mundart/Kreis Astfeld/ Gandersheim	Aufnahme-Nr. I/62	Transkribent Angelika Trauer	Listentyp A
Planrechteck X 29	Aufnahmedatum 20.11.1965	Transkribiert von 14.6.1985 bis 24.7.1985	

Besprochen von 24.07.1985
25.07.1985 UStv

Lfd. Nr.	Stichwort	Transkription	Bemerkungen
1	Winter 178	'vɪntə ^h ə	ɸ = kontinuant ə bereit velarisier
* 2	fliegen 56	'flaɪ→ə ^h ə	'fliegen die', Sequenzie- rung unklar - kein geminiertes [t]
3	Blätter 23	'blɛ:də ^h ə	
4	Luft 103	lu ^h ft ^h	ɸ = kontinuant
5	hört 89	hɪ ^h t ^h	ɸ = kontinuant
6	gleich 78	klɛ ^h →ɪk ^h	folgt P
7	schneien 130	ʃnɛ→ɪən	
8	Wetter 174	/	statt demoa 'vɪtə ^h ʀɔŋə
9	tu 151	dɛ→u	

Ort der Mundart/Kreis	Aufnahme-Nr.	Transkribent	Listentyp
-----------------------	--------------	--------------	-----------

Phonetischer Atlas von Deutschland

X 27	20.11.1965	bis 24.7.1985
------	------------	---------------

- Digitised in X-SAMPA, converted back to match original transcriptions, minor corrections for consistency of encoding

- The data is transcribed in high phonetic detail (3786 different phonetic segments)

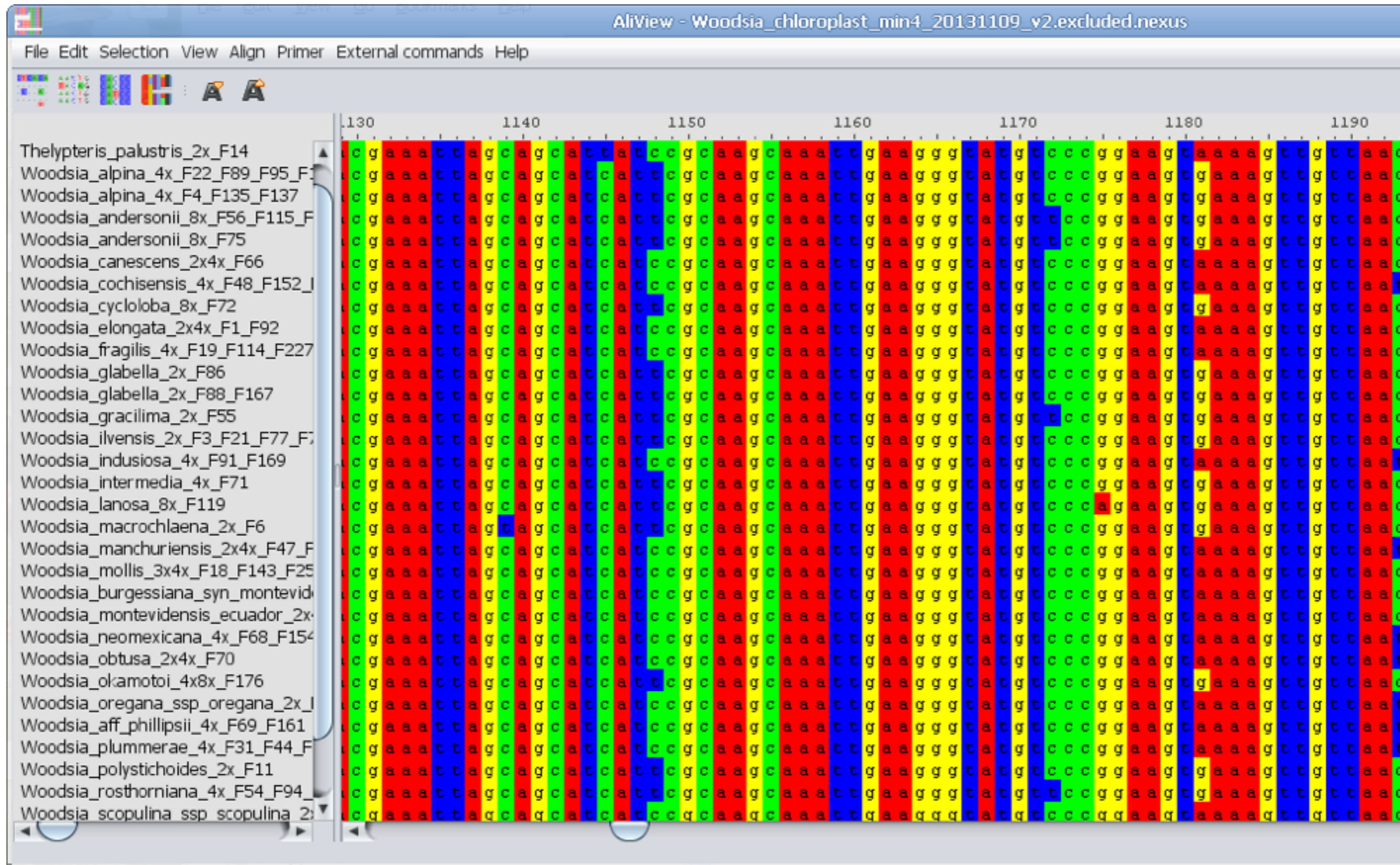
- We make the complete data available
 - ▶ electronically, separated by phonetic segments
 - ▶ as close as possible to the original source
 - ▶ including all idiosyncrasies
 - ▶ github.com/cysouw/PAD

Lfd. Nr.	Stichwort	Transkription	Bemerkungen
1	Winter	vɪntə	kontinuant
2	fliegen	flaɪ→ətʰ	2 boreifve forisier fliegen, neutric nung 'uklar - kau semi-waites, I, I
3	Plätter	ble: dɔ	kontinuant
4	Luft	luɔftʰ	
5	hört	hɔɪt	kontinuant
6	gleich	kleɪk	folgt
7	gleich	kleɪk	
8	gleich	kleɪk	staf deman 'vɪtə Rɔŋə
9	tu	dɔv	

Multiple Sequence Alignment

- Just a fancy name for sound correspondences
- Each sound correspondence is “aligned” in a column, possibly adding empty cells
- It is a useful and consistent way to represent comparative data (both between languages or dialects)

Multi-alignment of nucleotides (4-letter alphabet)



Multi-alignment of amino-acids (20-letter alphabet)

Accession	Species	Sequence	Length
Q5E940	BOVIN	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0	HUMAN	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0	MOUSE	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0	RAT	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0	CHICK	MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKOMQQIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0	RANSY	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMSGKNTMMRKAIRGHLENN--SALE	76
Q7ZUG3	BRARE	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQTIKRLSLRGK-AVVLMSGKNTMMRKAIRGHLENN--PALE	76
RLA0	ICTPU	MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKOMQTIKRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0	DROME	MVRENKAAWKAQYFIKVVLFDFPKCFIVGADNVGSKOMQNIKRTSLRGL-AVVLMSGKNTMMRKAIRGHLENN--PQLE	76
RLA0	DICDI	MSGAG-SKRKRLFIEKATKLFITYDKMIVAEADFGVSSQLQKIRKSIRGI-GAVLMGKKTMRKVIKIRDLADSK--PELD	75
Q54LP0	DICDI	MSGAG-SKRKNVFIEKATKLFITYDKMIVAEADFGVSSQLQKIRKSIRGI-GAVLMGKKTMRKVIKIRDLADSK--PELD	75
RLA0	PLAF8	MAKLSKQKKQMYIEKLSLIQQYSKILIVHVDNVGSKNQMASVRKSLRGK-ATILMSGKNTMRIRKALKKKNLQAV--PQIE	76
RLA0	SULAC	MIGLAVTTTKKIAKWKVDEVAELTEKLTHTKTIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG----YDTK	79
RLA0	SULTO	MRIMAVITQERKIAKWKIEEVKELEOKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG----LDVS	80
RLA0	SULSO	MKRLALALKQKVASWKLEEVKELTELKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFGIAAKNAG----IDIE	80
RLA0	AERPE	MSVVS LVGQMYKREKPIPEWKTLMLELEELFSKHRVFLFADLTGTPTFVVRVRKKLWKK-YPMMAVAKKRIILRAMKAAGLE---LDDN	86
RLA0	PYRAE	MMLAIGKRRYVTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIKPTLFGIAAKNAG---IPAE	85
RLA0	METAC	MAEERHHTTEHIPQWKKDEIENIKELIQSHKVFGMVIEGILATKMKIRRDLDKV-AVLKVSNTLTERALNQLG----ETIP	78
RLA0	METMA	MAEERHHTTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKMKIRRDLDKV-AVLKVSNTLTERALNQLG----ESIP	78
RLA0	ARCFU	MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGOMQKIRREFRKG-AEIKVVKNTLLERALDALG----GDYL	75
RLA0	METKA	MAVKAKGQPPSGYEPKVAEWRREVKELKELMDEYENVGLVDLEGIPAPQLQEIIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE	88
RLA0	METTH	MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLLISLAEKAGREL--ENVD	74
RLA0	METTL	MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA	82
RLA0	METVA	MIDAKSEHKIAPWKIEEVNALKELLKSNVIALIDMMEVPAVQLQEIIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA	82
RLA0	METJA	METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNPKLA	81
RLA0	PYRAB	MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRLIRENGGLLRVSNTLIELAIKKAQELGKPELE	77
RLA0	PYRHO	MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRLIRENGGLLRVSNTLIELAIKKAQELGKPELE	77
RLA0	PYRFU	MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRLIRENGLLRVSNTLIELAIKKAQELGKPELE	77
RLA0	PYRKO	MAHVAEWKKKEVEELANLIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSNTLIELAIKRAAQELGQPELE	76
RLA0	HALMA	MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSNTLLEALDDVD----DGLE	79
RLA0	HALVO	MSESEVRQTEVIPQWKREEVDLDFIESYESVGVVGVAGIPSRQLQSMRRE LHGS-AAVRMSRNTLVNRAALDEVN----DGFE	79
RLA0	HALSA	MSAEEQRTTEEVPWKQEEVAELVDLLETYSVGVVNVVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG----DGLD	79
RLA0	THEAC	MKEVSQQKKELVNEITORIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLEKALLENLGD----EKLS	72
RLA0	THEVO	MRKINPKKKEIVSELAQDITKSKAVAVDIKGVTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT	72
RLA0	PICTO	MTEPAQWKIDFVKNLENEINSRKVAIVSIVKGLRNNEFQKIRNSIRDK-ARIKVSRRARLLRLAIENTGK----NNIV	72

ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90

LOCATION	WORD
Aachen	a:ph
Adorf	ɑ:b ^h ə
Ahrbergen	o→ɔphə
Albersloh	ɑ:p ^h ə
Allna	αϕh
Altenberg	Λfɛ
Altentrüdin	af
Altlandsberg	ɑ'fə'
Altwarp	o:ph
Astfeld	ɒ':p ^h ə
Atzendorf	afɛ
Ballhausen	Λ'fə
Bardenfleth	ɔ:p̄ϕ
Barssel	ɒ:p ^h ə
Bempflingen	af:
Bennin	ɔp ^h
Billingsbach	af
Bockelwitz	Λvə
Bonn	ɑ:p'
Borstendorf	ʏf:
Breddin	ɒ:ph
Brelingen	ɑfβə
Bremscheid	ɒ':p ^h ə
...	...

A	FF	E
a:	ph	-
ɑ:	b ^h	ə
o→ɔ	ph	ə
ɑ:	p ^h	ə
α	ϕh	-
Λ	f	ɛ
ɑ	f	-
ɑ'	f	ə'
o:	ph	-
ɒ':	p ^h	ə
a	f	ɛ
Λ'	f	ə
ɔ:	p̄ϕ	-
ɒ:	p ^h	ə
a	f:	-
ɔ	p ^h	-
ɑ	f	-
Λ	v	ə
ɑ:	p'	-
ʏ	f:	-
ɒ:	ph	-
ɑ	f̄β	ə
ɒ':	p ^h	ə
...

● **Workflow:**

- ▶ Tokenisation of segments (github.com/cysouw/qlcTokenize)
- ▶ Automatic alignment using **LingPy** (github.com/lingpy)
- ▶ Manual correction using **Alignment Editor** (github.com/digitallinguist/msa-editor)
- ▶ Separation of cognates (e.g. *Samstag* vs. *Sonnabend*)
- ▶ Annotation of columns (e.g. many-to-one alignments, metathesis)
- ▶ Merging of complex columns and removing boundaries

MSA Editor

Choose Files 3 files Augenblick_1013.msa View Edit Reload Save

COLUMNID	1	2	3	4	5	6	7	8	9
STANDARD	Au	g	e	n	b	l	i	ck	(e)
Adorf	æ→u	ʁ	-	ŋ	b	l	ɛ	k ^h	-
Ahrbergen	æ→õ	ʁ	ə	m	b	l	ɪ	k'	-
Albersloh	æ→u	-	-	m	β	l	ɪ	k	-
Allna	ɔɪ	-	-	-	p	l	æ ^c	x	-
Altenberg	ʁʁ	ʁ	ã	-	b	l	ɪ	k	-
Altentrüdin	æ→u	ʁ	ə	-	p	l	ɪ	g	-
Altlandsberg	ɑ'→u	ʁ	-	ŋ	b	l	ɪ	k̄x	-
Altwarp	õu	-	-	ŋ	b	l	ɪ	k	-
Astfeld	ʊɪ	ʁ	ə	m	b	l	ɪ	k ^h	ə
Ballhausen	ɑ→u	ʁ	-	ŋ	p	l	ɪ	k	-
Bardenfleth	oɪ	g	-	ŋ	b	l	ɛ	k̄x ₊	-
Barssel	oɪ	g	-	ŋ	p	l	ɪ	k ₊	-
Bempflingen	æ→u	g	ə	-	b	l	ɪ	c ^h	-
Bennin	oɪ	-	-	ŋ	b	l	ɪ	x	-
Billingsbach	ɑɪ	x	ə	-	p	l	ɪ	k̄x	-
Bockelwitz	ɑ→u	ʁ	-	ŋ	b	l	ɪ	k	-
Borstendorf	ɔɪ	ʁ̄x	-	ŋ	p	l	ɛ	k	-

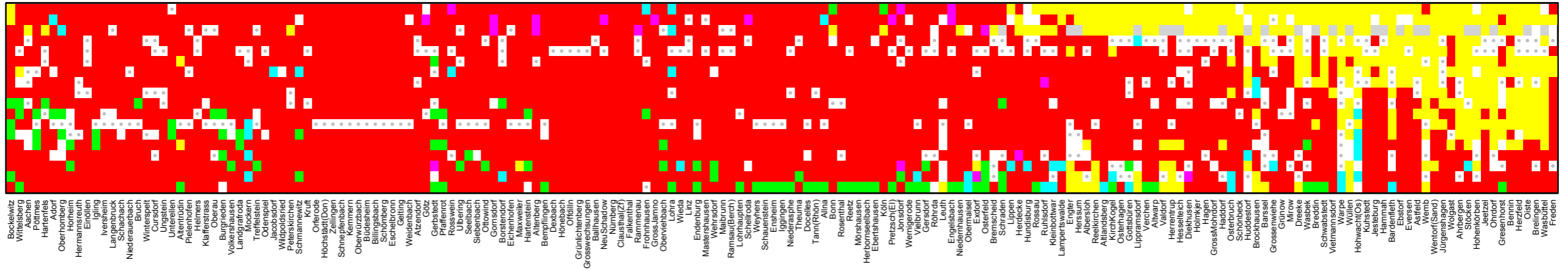
github.com/digitallinguist/msa-editor

Step 3: Correspondence Sets

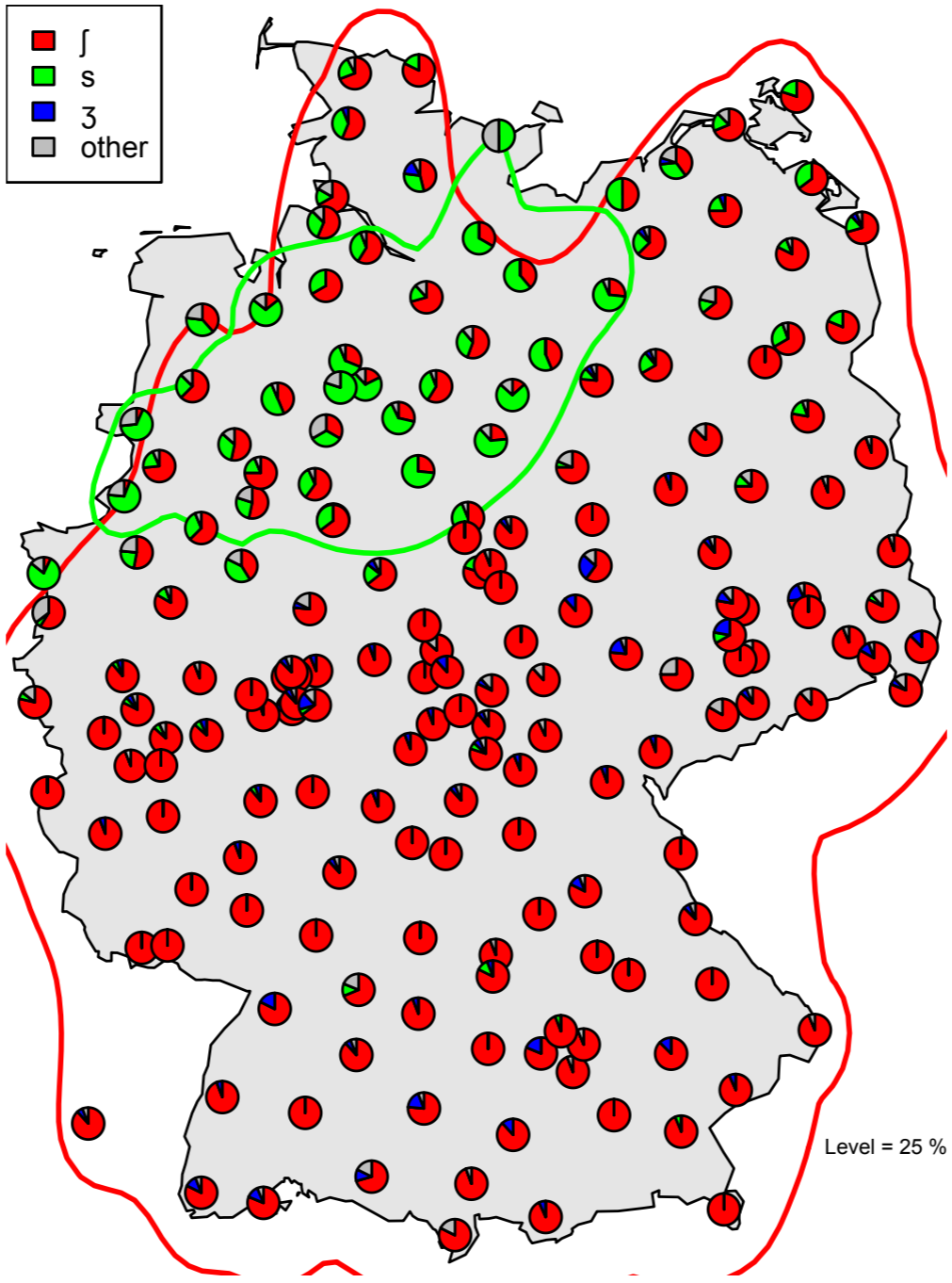
- more than 700 columns of aligned segments (“correspondences”)
- Comparative-historical linguistics uses clusters of correspondences (“correspondence sets”)
- Automatic clustering of columns is a good start, but needs correction
- Visualisations in R
github.com/cysouw/qlcVisualize

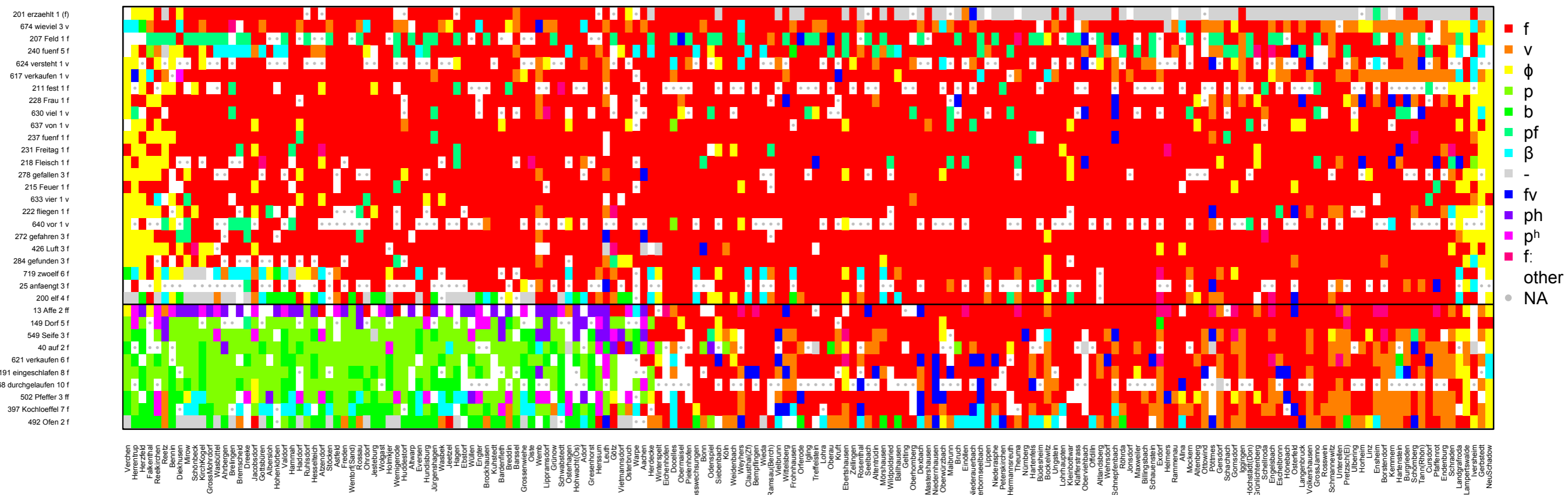


700 Wurst 5 s
 172 Durst 4 s
 142 Donnerstag 8 s
 312 gestohlen 3 s
 626 versteht 4 s
 85 bestellt 3 s
 581 Stueckchen 1 s
 319 gestorben 3 s
 534 schwarz 1 s
 188 eingeschlafen 5 sch
 522 Schnee 1 sch
 525 schneien 1 sch
 516 schlechte 1 sch
 538 Schwester 1 sch
 530 schoene 1 sch
 221 Tisch 3 sch
 587 Tisch 3 sch
 156 Dreschen 7 sch



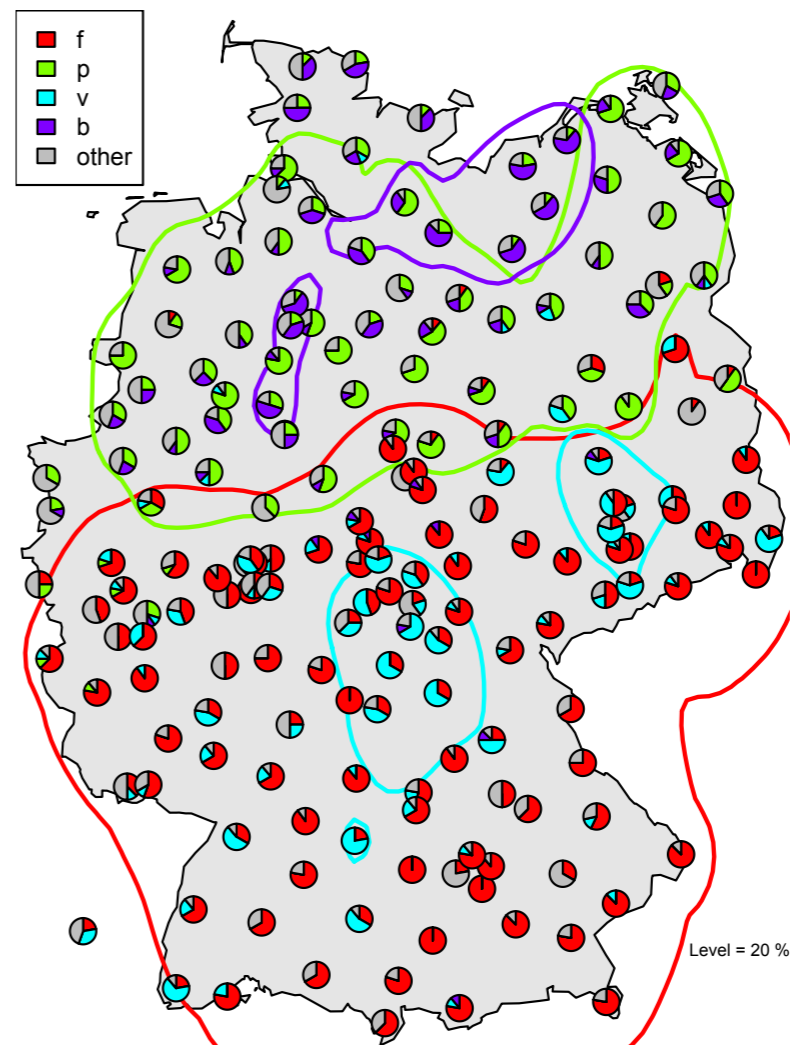
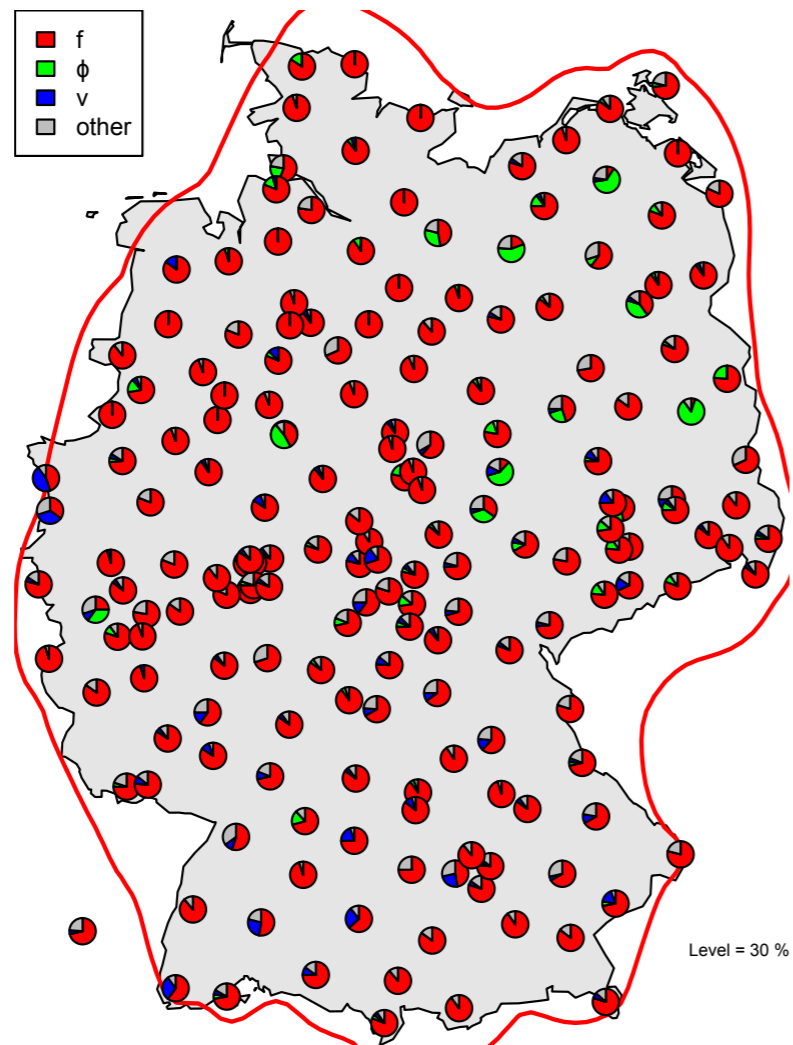
Correspondences "sch"





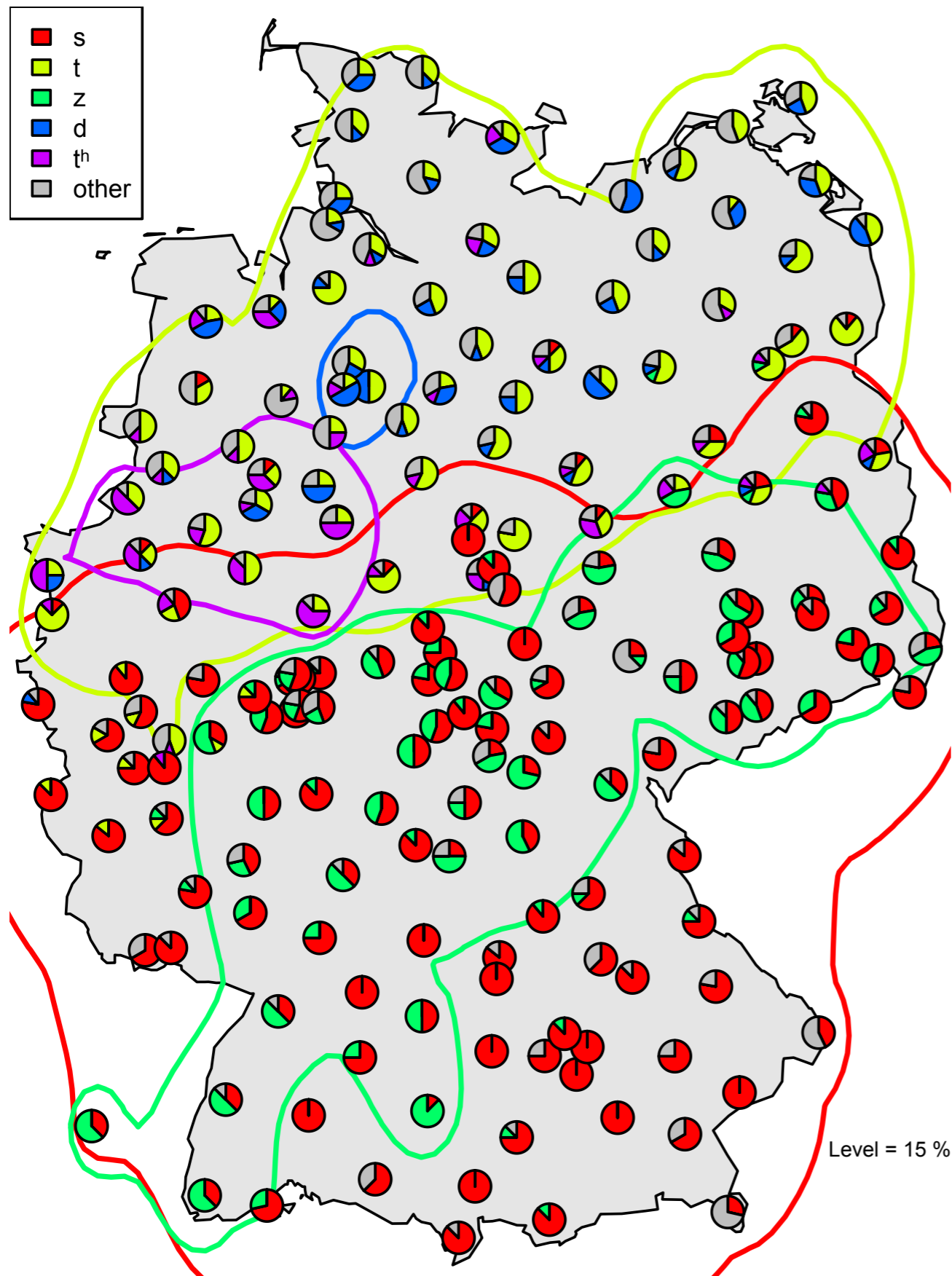
Correspondences "f"

Correspondences "f/p"

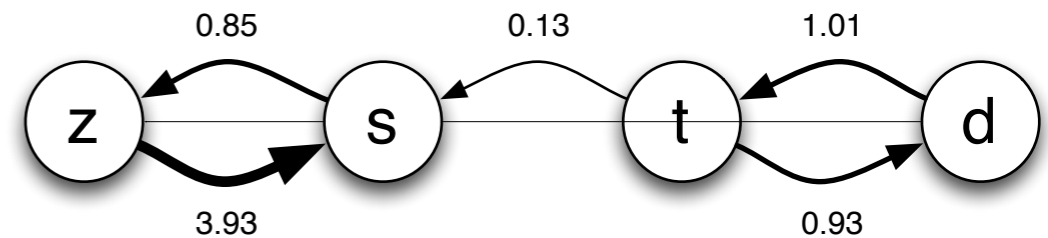


Step 4: Sound Change Modelling

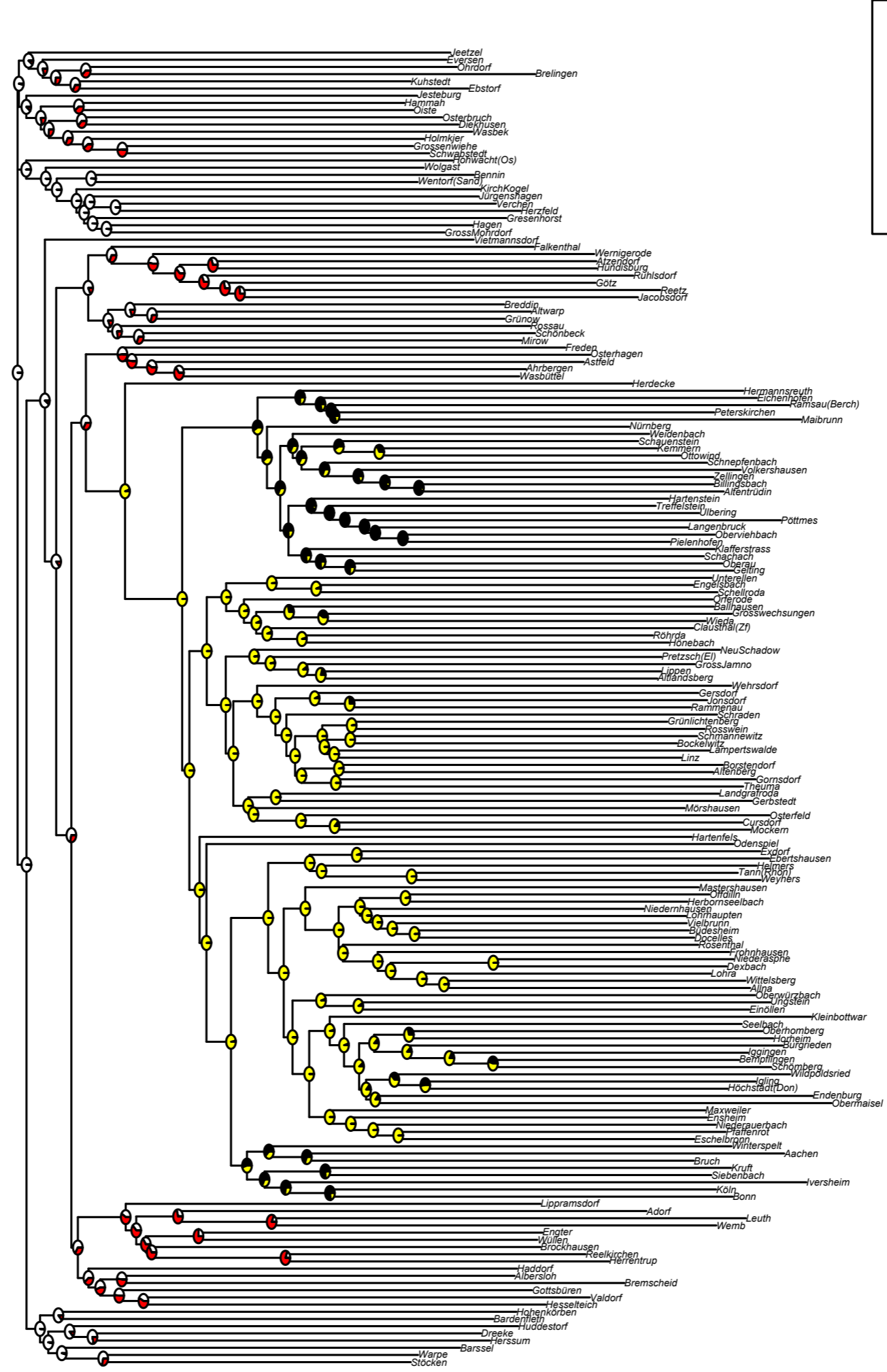
Correspondences “s/t”



- Fit a model of sound changes on an unrooted tree based on all correspondences (using *corHMM* in R)
- Continuous-time Markov Chain transition rates:



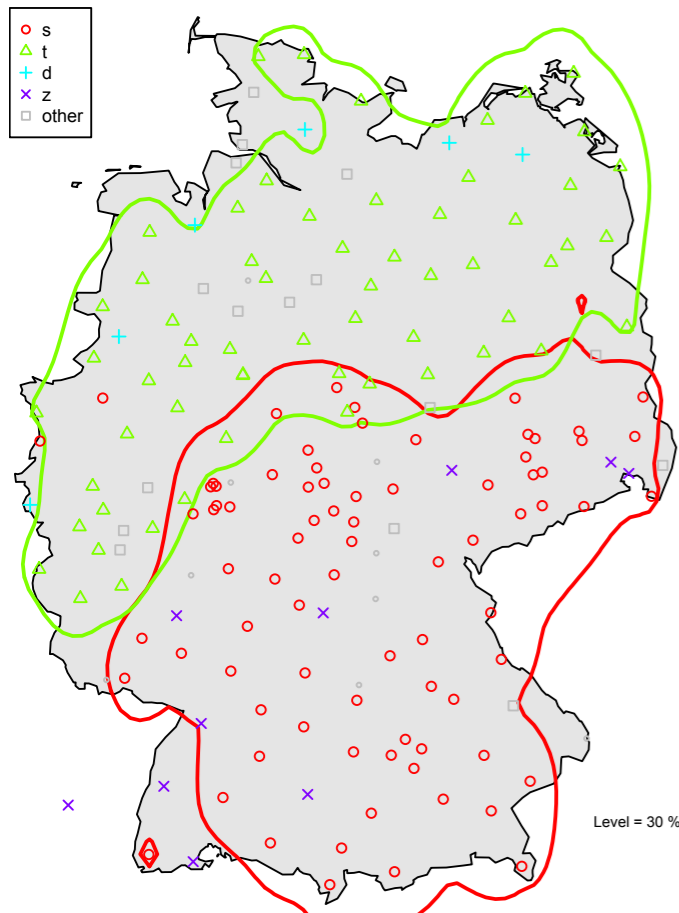
Based on /s/ in German words:
beißen, besser, das, größer, groß,
heiß, muss, Wasser, weiße



Conclusions

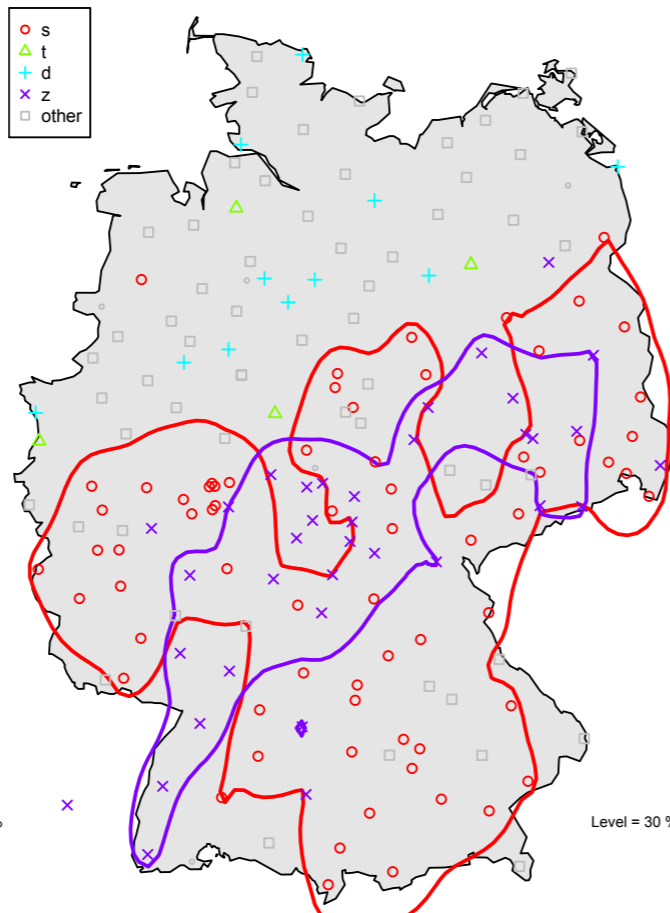
- Old-school historical linguistics is worth looking into
- Quantitative methods can augment that tradition
- Interaction between algorithmic procedures and manual decisions is (currently) needed
- Use small, one-trick, tools that can be combined (“Unix-philosophy”) instead of dreaming of large scale computational infrastructure

121 das 3 s



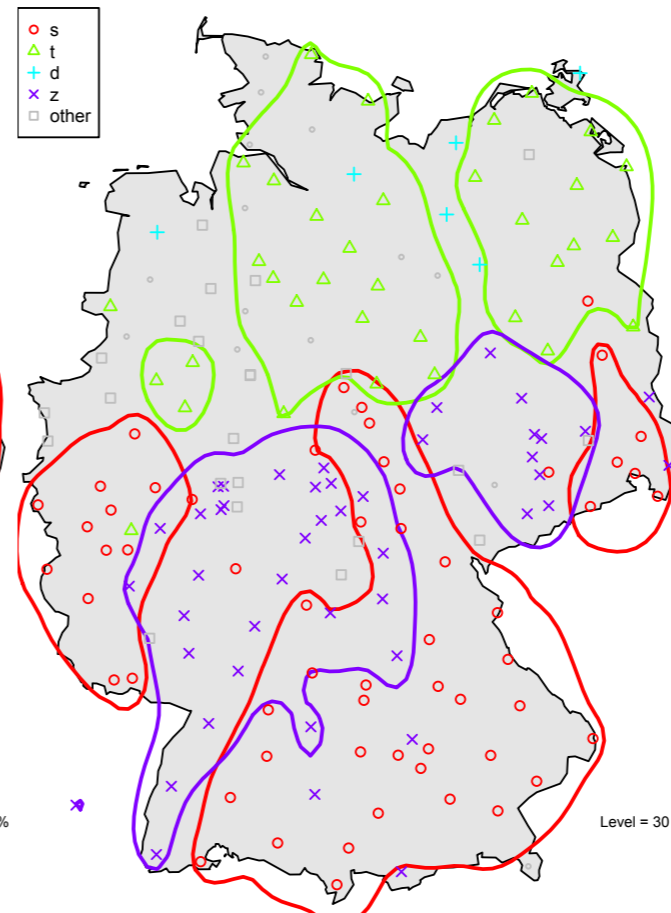
Level = 30 %

81 besser 3 ss



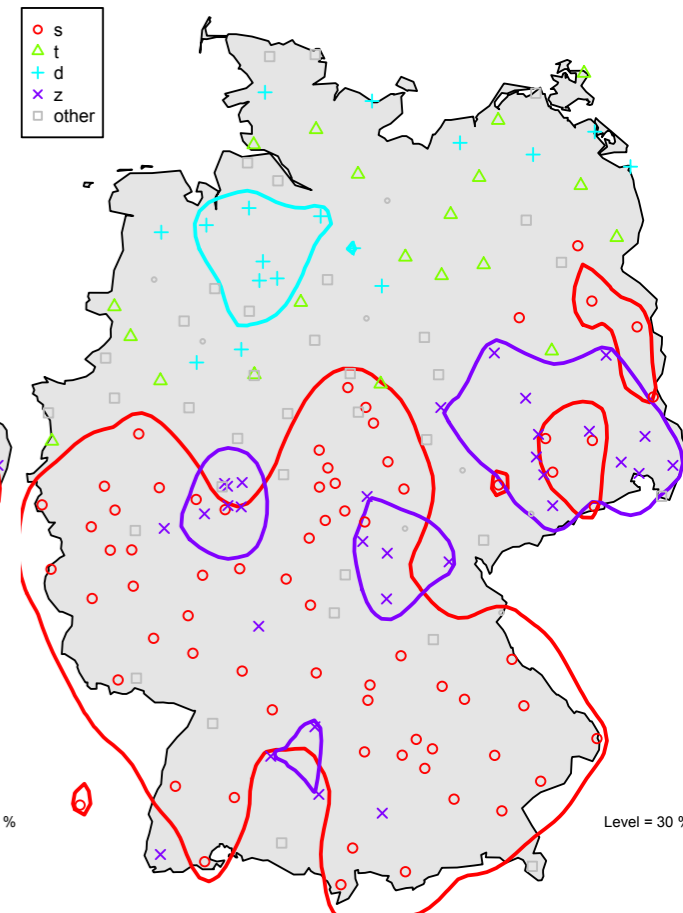
Level = 30 %

72 beissen 3 ß



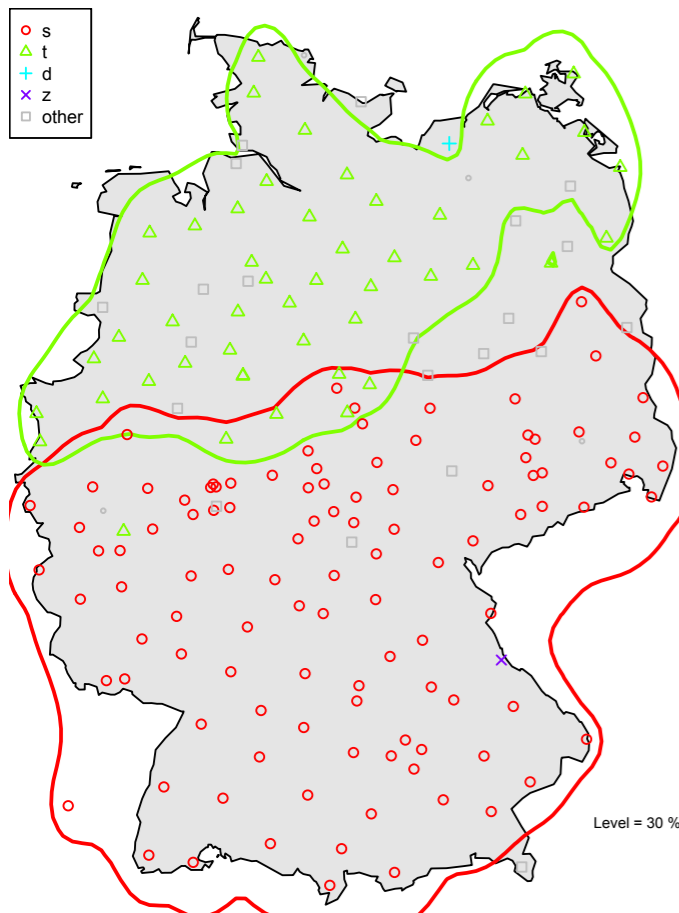
Level = 30 %

660 weisse 3 ß



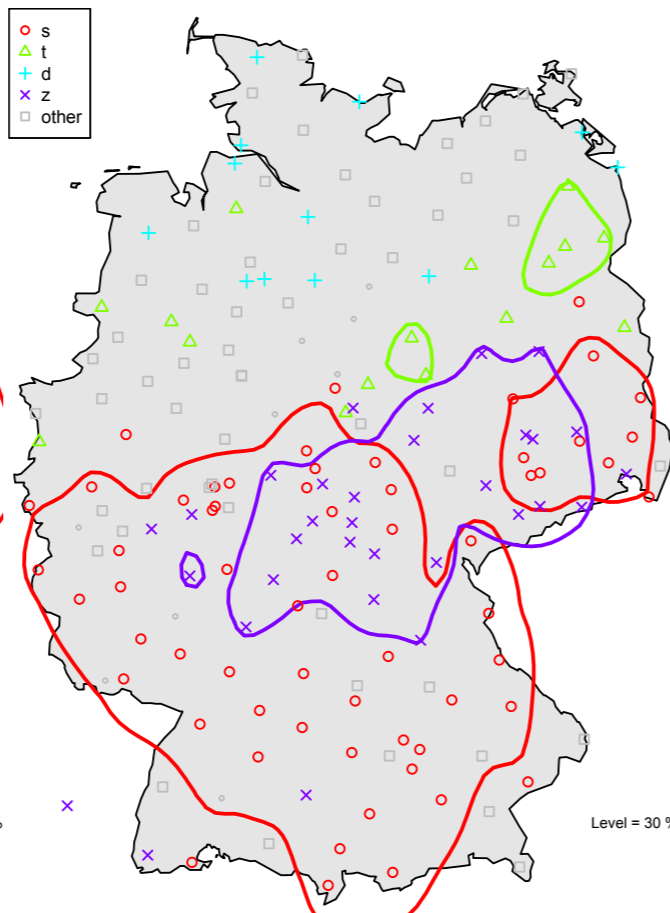
Level = 30 %

341 gross 4 ß



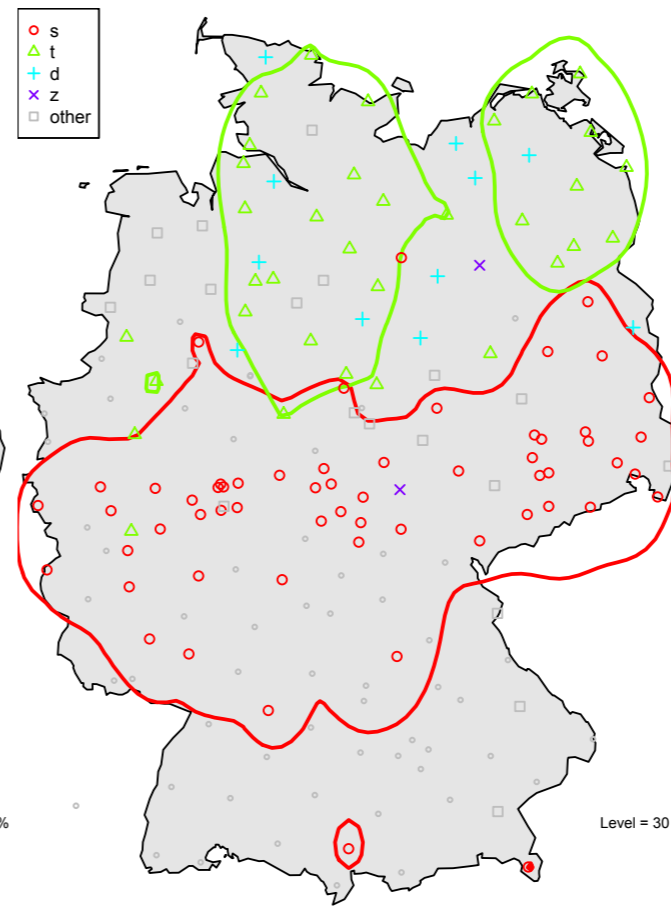
Level = 30 %

651 Wasser 3 ss



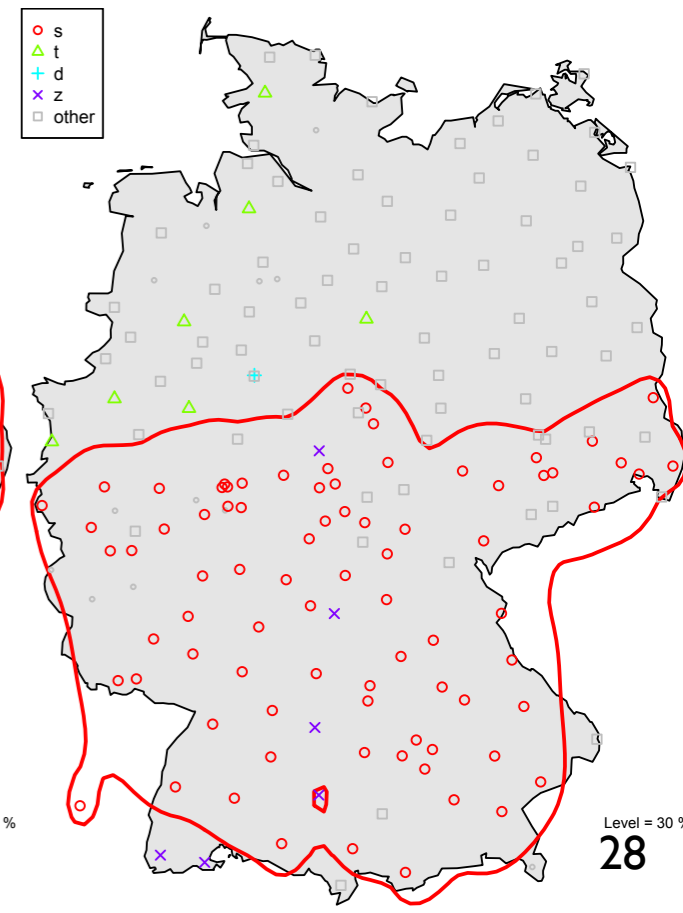
Level = 30 %

473 muss 3 ss



Level = 30 %

357 heiss 3 ss



Level = 30 %