Advances in computer-assisted historical reconstruction

Michael Cysouw
Philipps-Universität Marburg

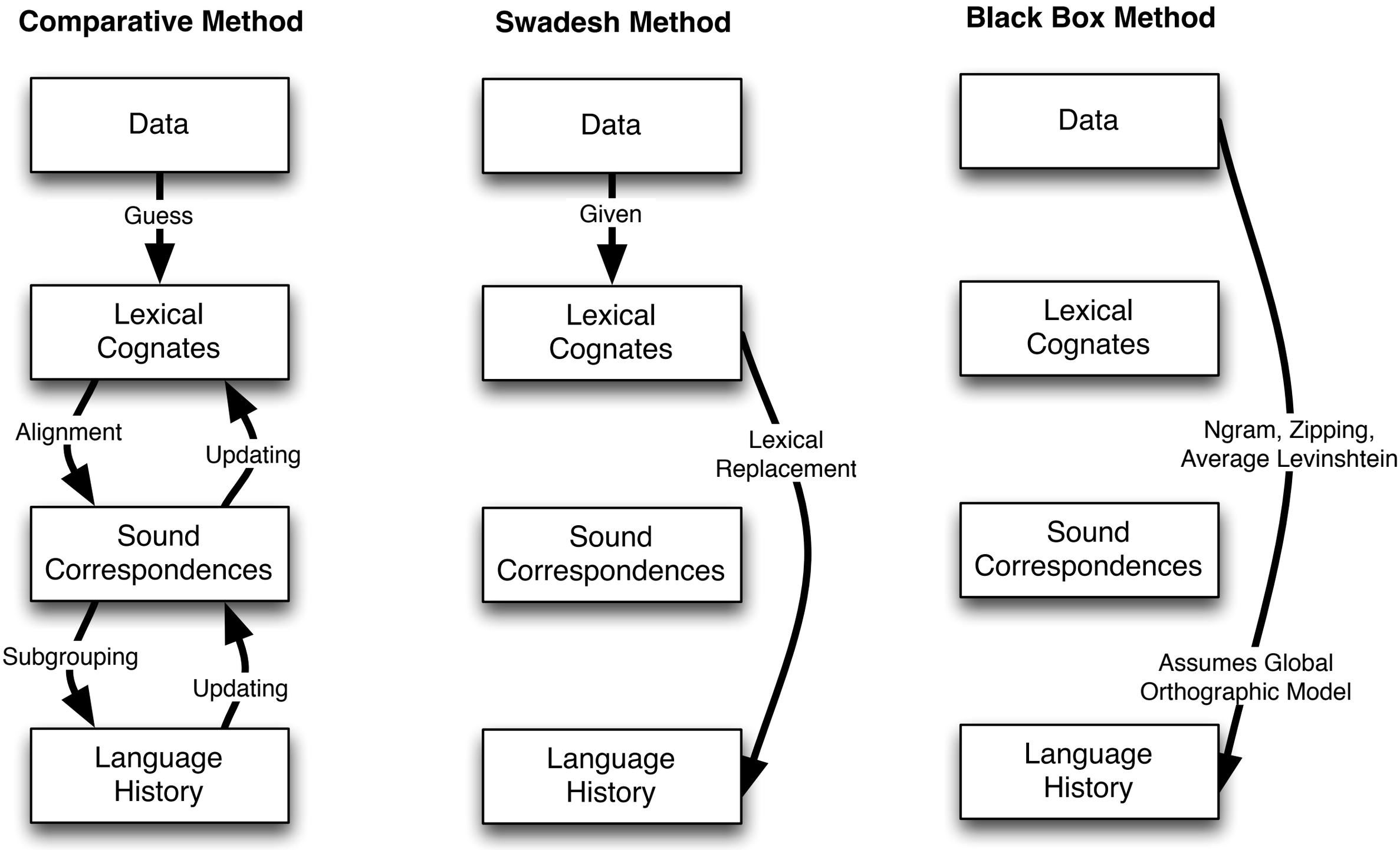
Philipps Universität Marburg

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
Data
Guess
Lexical Cognates
Alignment
Updating
Sound Correspondences
Subgrouping
Updating
Language History
Swadesh Method
Data
Given
Lexical Cognates
Lexical Replacement
Sound Correspondences
Language History
Black Box Method
Data
Lexical Cognates
Ngram, Zipping, Average Levinshtein
Sound Correspondences
Assumes Global Orthographic Model
Language History

Computational pipeline for comparative method

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

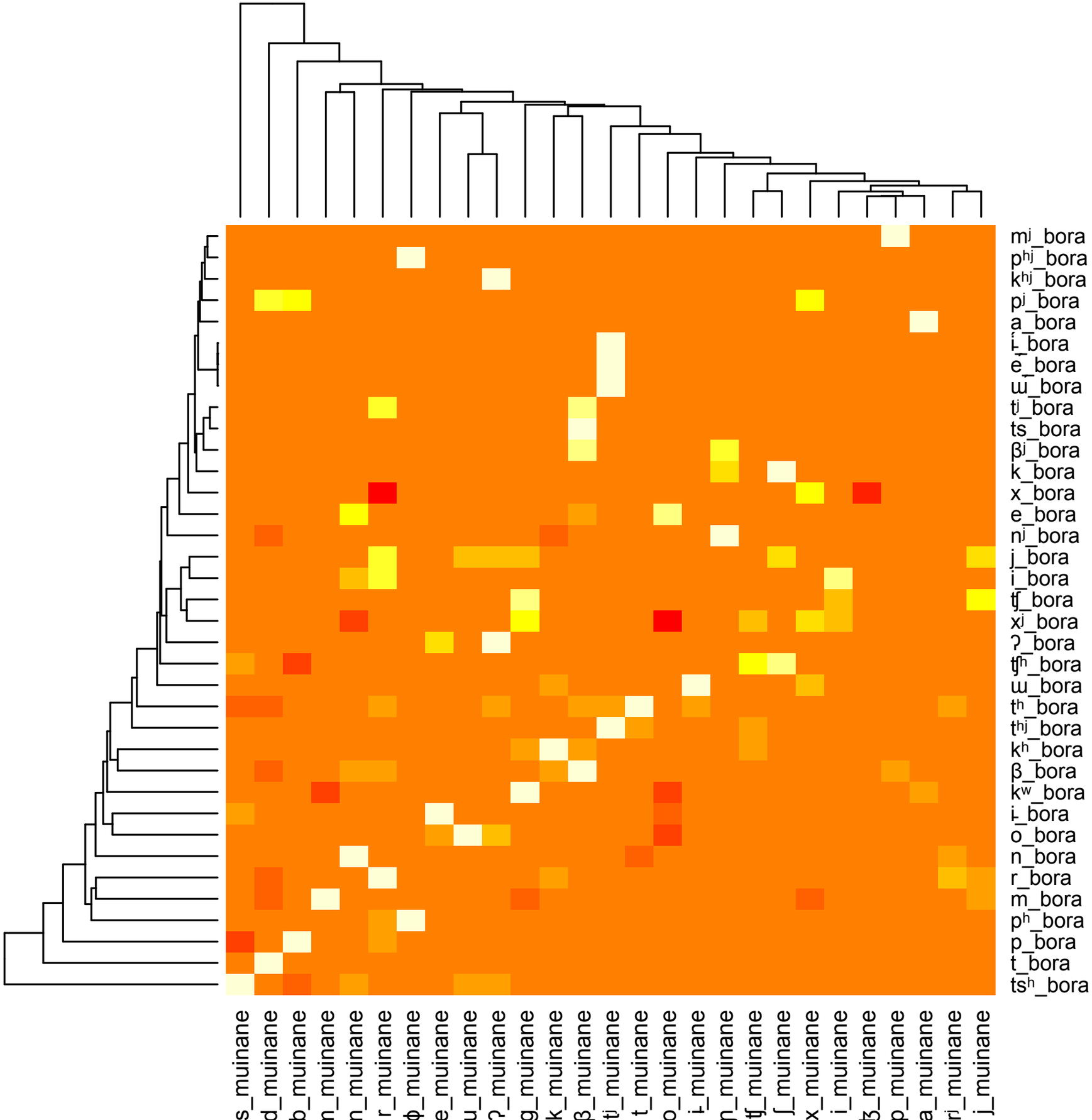
Open-source Tools

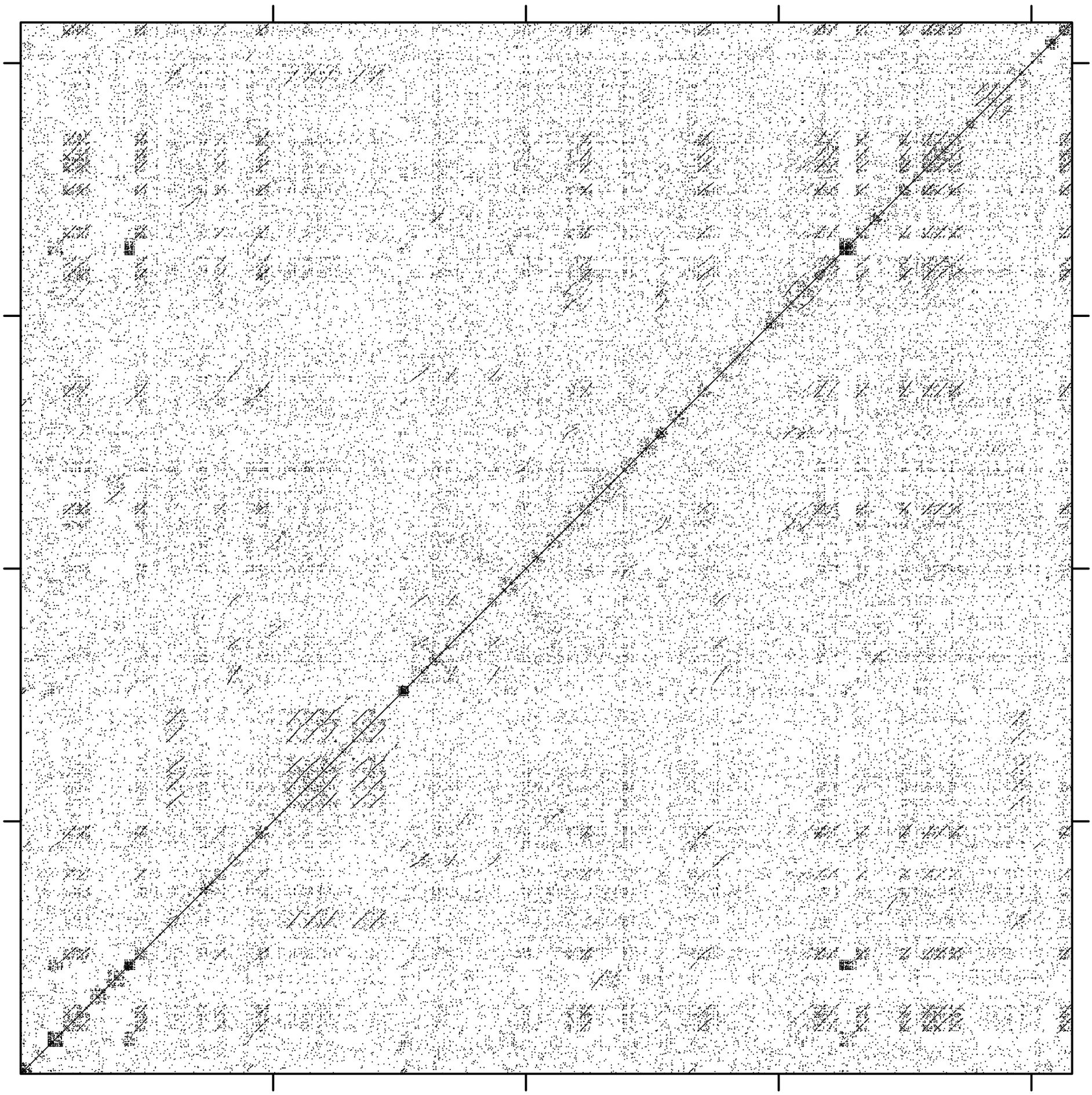
- **Orthography Processing** (with Steven Moran)
  github.com/cysouw/qlcTokenize (R, also available in Python)
- **Guessing regular correspondences:**
  github.com/cysouw/qlcMatrix (R)
- **Alignments (and more!) with LingPy** (Johann-Mattis List)
  github.com/lingpy (Python)
- **Alignment Editor** (with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

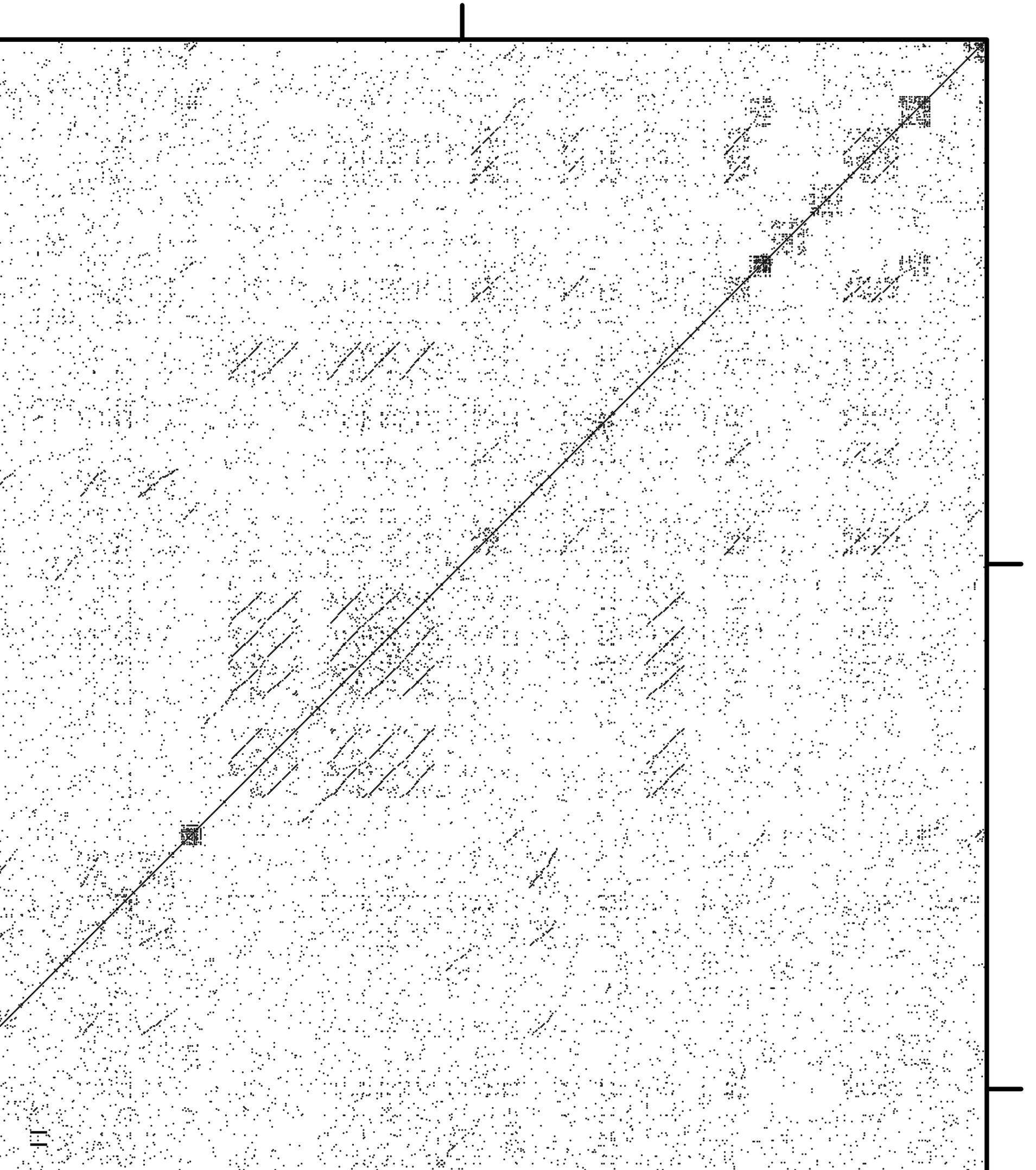
Step 1: Guessing by regularity

- Given some dictionaries, we can produce lists of words with (roughly) 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

	Bora	Muinane
down	tʃíɨnʲe, paári	báari, gíino
bee	íímúɯʔóexpʰi, téʔtsʰipa	nɨ́ɨ́biɨri, mɨ́ɨbiɨriʔɨ
sharp	tsʰɯ́ʔxɨβáne	sɨ́ɨ́xéβano
…	…	…


mʲ_bora
pʰʲ_bora
kʰʲ_bora
pʲ_bora
a_bora
ɨ_bora
e_bora
ɯ_bora
tʲ_bora
ts_bora
βʲ_bora
k_bora
x_bora
e_bora
nʲ_bora
j_bora
i_bora
tʃ_bora
xʲ_bora
ʔ_bora
tʃʰ_bora
ɯ_bora
tʰ_bora
tʰʲ_bora
kʰ_bora
β_bora
kʷ_bora
ɨ_bora
o_bora
n_bora
r_bora
m_bora
pʰ_bora
p_bora
t_bora
tsʰ_bora
s_muinane
d_muinane
b_muinane
m_muinane
n_muinane
r_muinane
ɸ_muinane
e_muinane
u_muinane
ʔ_muinane
g_muinane
k_muinane
β_muinane
tʲ_muinane
t_muinane
o_muinane
ɨ_muinane
ɲ_muinane
tʃ_muinane
ʃ_muinane
x_muinane
i_muinane
ʤ_muinane
p_muinane
a_muinane
rʲ_muinane
j_muinane





Step 2: Alignment

Lfd. Nr.	Stichwort		Transkription	Bemerkungen
1	Winter	178	ˈvɪntʰɐ̯ɐ̯	ɐ̯ = konsonant ɐ̯ bereits velarisiert
2	fliegen	56	ˈflɑ̯ɪ̯ɐ̯tʰə	'fliegen die', Sequenzierung unklar - kein gemeinsames [L]
3	Blätter	23	ˈblɛ̝ːdɐ̯ɐ̯̊	
4	Luft	103	lʊˈftʰ‿	ɐ̯ = konsonant
5	hört	89	hɪ̯ɐ̯t‿	ɐ̯ = konsonant
6	gleich	78	‿klɛ̝ɐ̯ɪk‿	folgt P
7	schneien	130	ˈʃnɛɐ̯ɪən	
8	Wetter	174	/	statt dessen ˈvɪtəʀʊŋə
9	tu	151	dɛɐ̯ʊ	

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

Ort der Mundart/Kreis	Aufnahme-Nr.	Transkribent	Listentyp
Astfeld/ Gandersheim	I/62	Angelika Braun	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 L/SK

Lfd. Nr.	Stichwort		Transkription	Bemerkungen
1	Winter	178	ˈvɪntʰəɐ̯	ɐ̯ = kontinuant Übergang velarisiert
2	fliegen	56	ˈflaɪ̯ɔtʰə	'fliegen die', Sequenzierung unklar - kein gemeintes [L]
3	Blätter	23	ˈblɛːdəɐ̯̊	
4	Luft	103	lʊrˈftʰ‿	ɐ̯ = kontinuant
5	hört	89	hɪɐ̯t‿	ɐ̯ = kontinuant
6	gleich	78	‿klɛɪ̯ç‿	folgt P
7	schneien	130	ˈʃnɛɪ̯ən	
8	Wetter	174	/	statt dessen ˈvɪtəʀʊŋə
9	tu	151	dɛʊ̯	

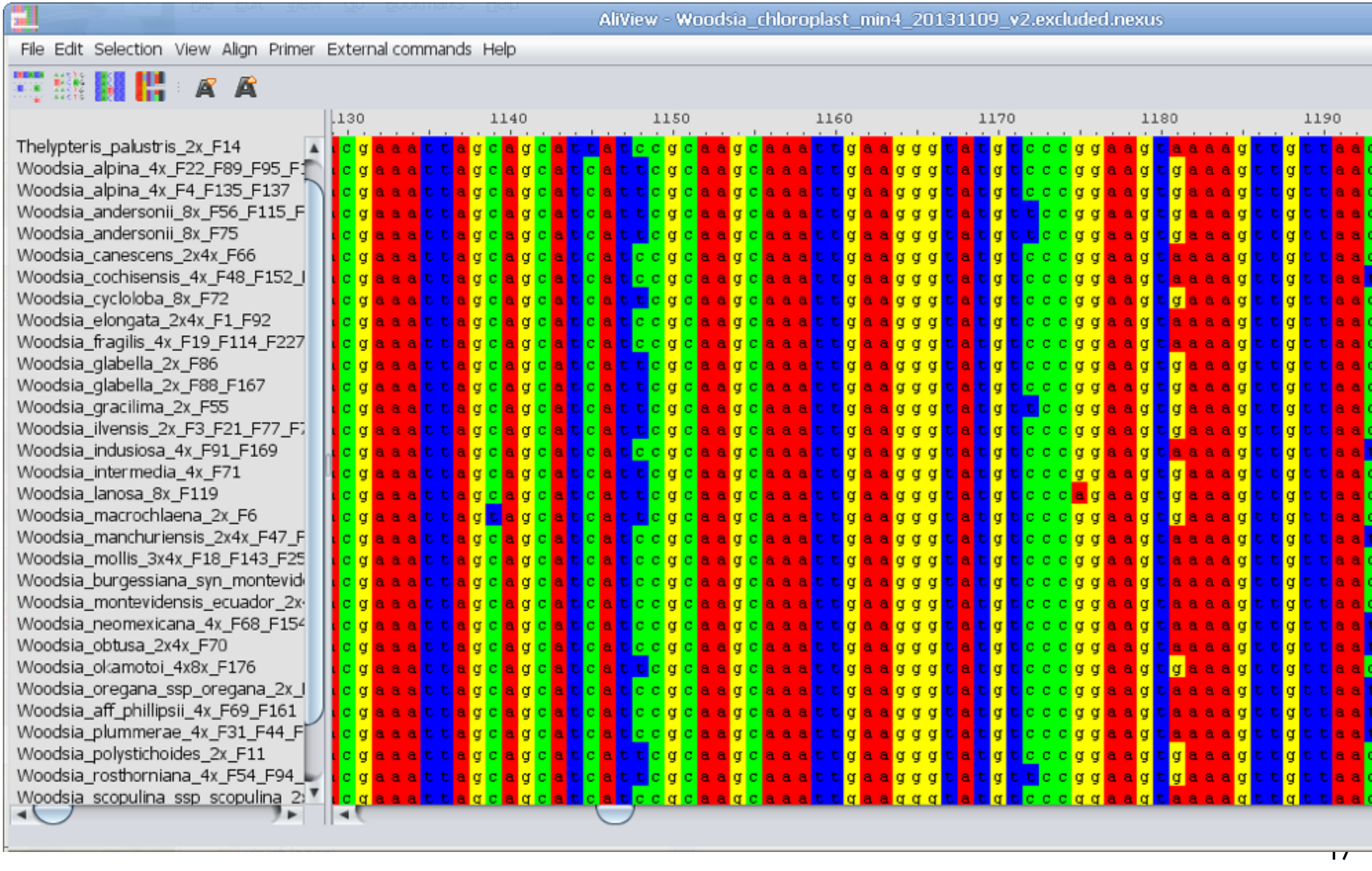
Phonetischer Atlas von Deutschland

- 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

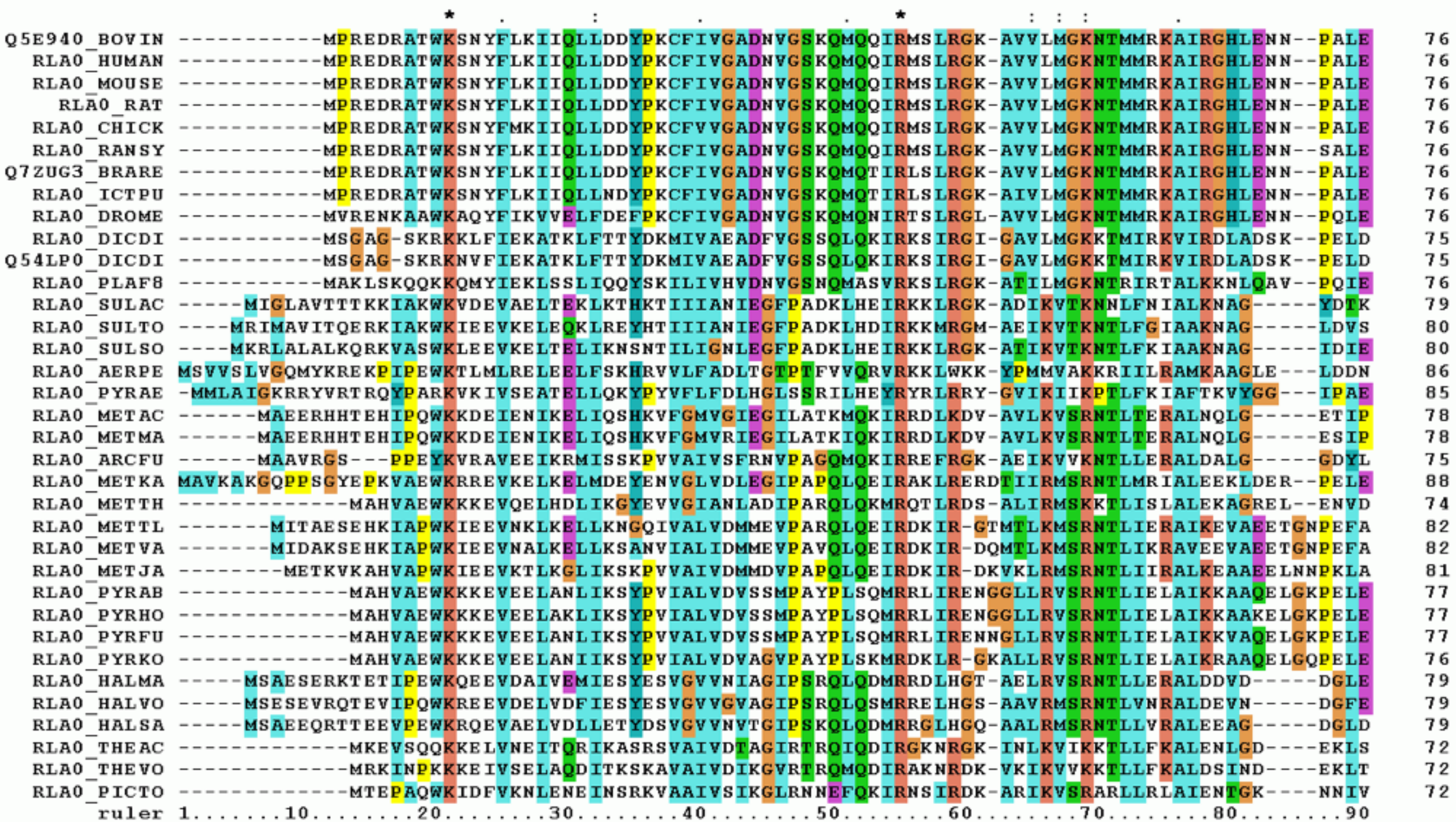
Multiple Sequence Alignment

- 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)



LOCATION	WORD
Aachen	a̠ːph
Adorf	ɑ̟ːbʱə
Ahrbergen	ɒ̝→ɔ̟phə̟
Albersloh	ɑːpᵝ̈ə̈
Allna	ɑɸh
Altenberg	ʌfɛ̈
Altentrüdin	a̠f
Altlandsberg	ɑʾfəʾ
Altwarp	oːph
Astfeld	ɒˤːpʰə
Atzendorf	afɛ̠
Ballhausen	ʌʾfə
Bardenfleth	ɔːp͡ɸ
Barssel	ɒːpʱə
Bempflingen	afː
Bennin	ɔ̝pʰ
Billingsbach	a̠f
Bockelwitz	ʌ̟və̈
Bonn	ɑːpʾ
Borstendorf	ɤ̝fː
Breddin	ɒːph
Brelingen	ɑ̟f͡βə
Bremscheid	ɒˤːpḫə̟
…	…

A	FF	E
a̠ː	ph	-
ɑ̟ː	bʱ	ə
ɒ̝→ɔ̟	ph	ə̟
ɑː	pᵝ̈	ə̈
ɑ	ɸh	-
ʌ	f	ɛ̈
a̠	f	-
ɑʾ	f	əʾ
oː	ph	-
ɒˤː	pʰ	ə
a	f	ɛ̠
ʌʾ	f	ə
ɔ̟ː	p͡ɸ	-
ɒ̟ː	pʱ	ə
a	fː	-
ɔ̝	pʰ	-
a̠	f	-
ʌ̟	v	ə̈
ɑː	pʾ	-
ɤ̝	fː	-
ɒː	ph	-
ɑ̟	f͡β	ə
ɒˤː	pḫ	ə̟
…	…	…

- **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	ɐ→ʊ	ʁ	-	ŋ̩	b	l	e̞	k^h	-
Ahrbergen	ə̹→o͡o̝	ɣ	ə	m	b	l	ɪ̽	k’	-
Albersloh	ɐ→ʊ	-	-	m	β̃	l	ɪ	k	-
Allna	ɔː	-	-	-	p	l	œˤ	x	-
Altenberg	ɤɣ̣	ʁ	ə̃	-	b	l	ɪ	k	-
Altentrüdin	ɐ→ʊ	ʁ	ə	-	p	l	ɪ̣	ɡ	-
Altlandsberg	ɑ’→ʊ	ɣ	-	ŋ̩	b	l	ɪ̣	k͡x	-
Altwarp	o͡ʊ	-	-	ŋ̩	b	l	ɪ	k	-
Astfeld	ʊ̗ː	ɣ	ə	m	b	l	ɪ	k^h	ə
Ballhausen	a̽→ʊ	ɣ	-	ŋ̩	p	l	ɪ̽	k	-
Bardenfleth	oː	ɡ	-	ŋ̩	b	l	ɛ	k͡x̥	-
Barssel	oː	g̃	-	ŋ̩	p	l	ɪ̗	k̥	-
Bempflingen	ɐ→ʊ	ɡ	ə	-	b	l	ɪ̣	c^h	-
Bennin	oː	-	-	ŋ̩	b	l	ɪ̗	x	-
Billingsbach	ɑ̟ː	x	ə	-	p	l	ɪ̣	k̃x	-
Bockelwitz	a̽→ɯ	ɣ	-	ŋ̩	b	l	ɪ	k	-
Borstendorf	ɔ̗ː	ɣ͡x	-	ŋ̩	p	l	e̞	k	-

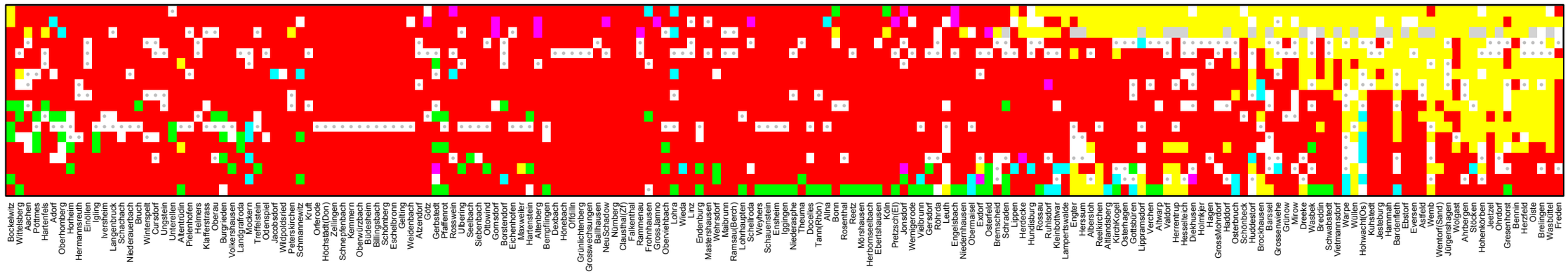
github.com/digitallinguist/msa-editor

Step 3: Correspondence Sets

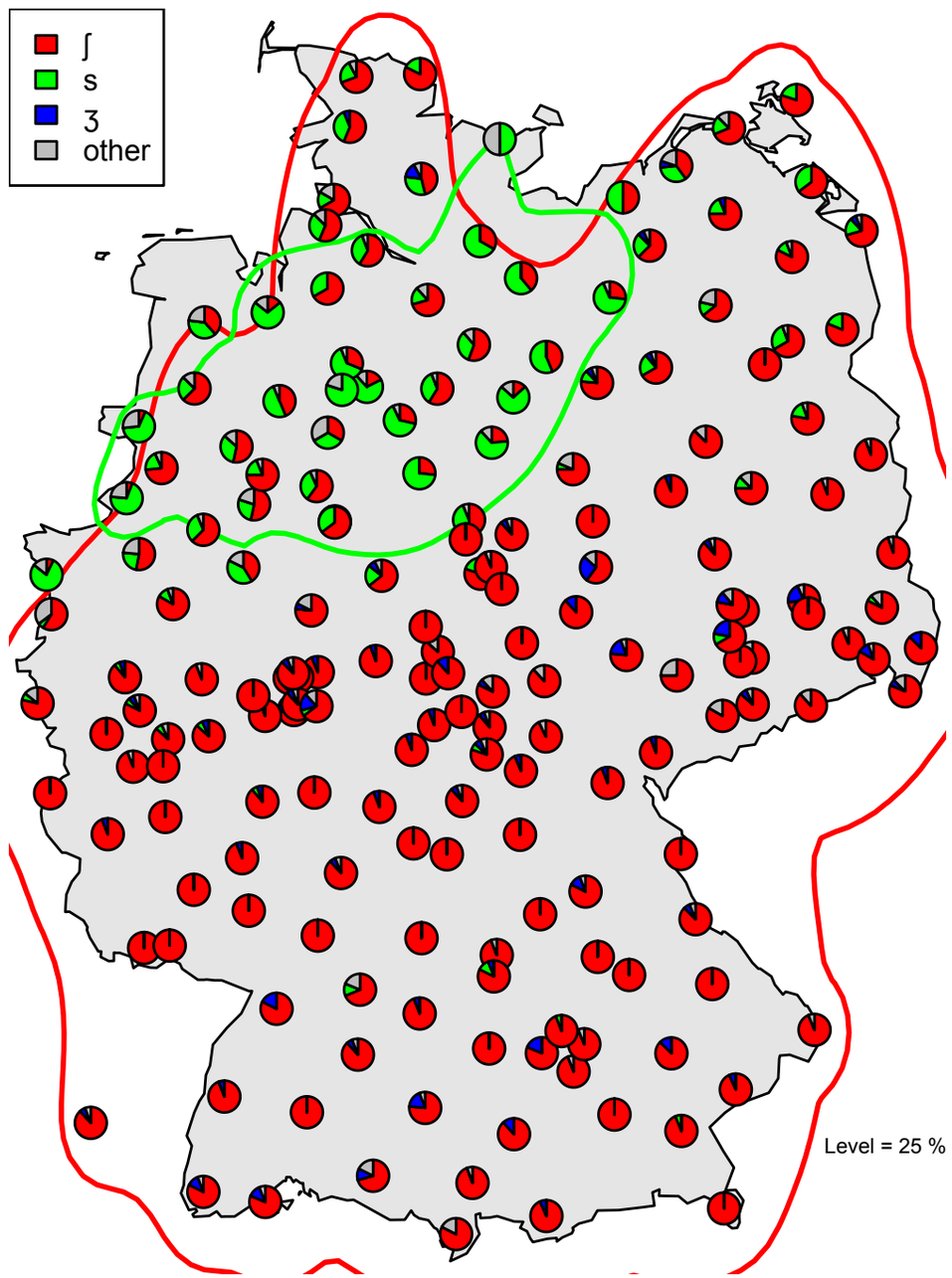
- more than 700 columns of aligned segments (“correspondences”)
- Comparative-historical linguistics uses clusters of correspondences (“correspondence sets”)
- Automatic clustering of columns is a good start, but needs correction
- Visualisations in R github.com/cysouw/qlcVisualize

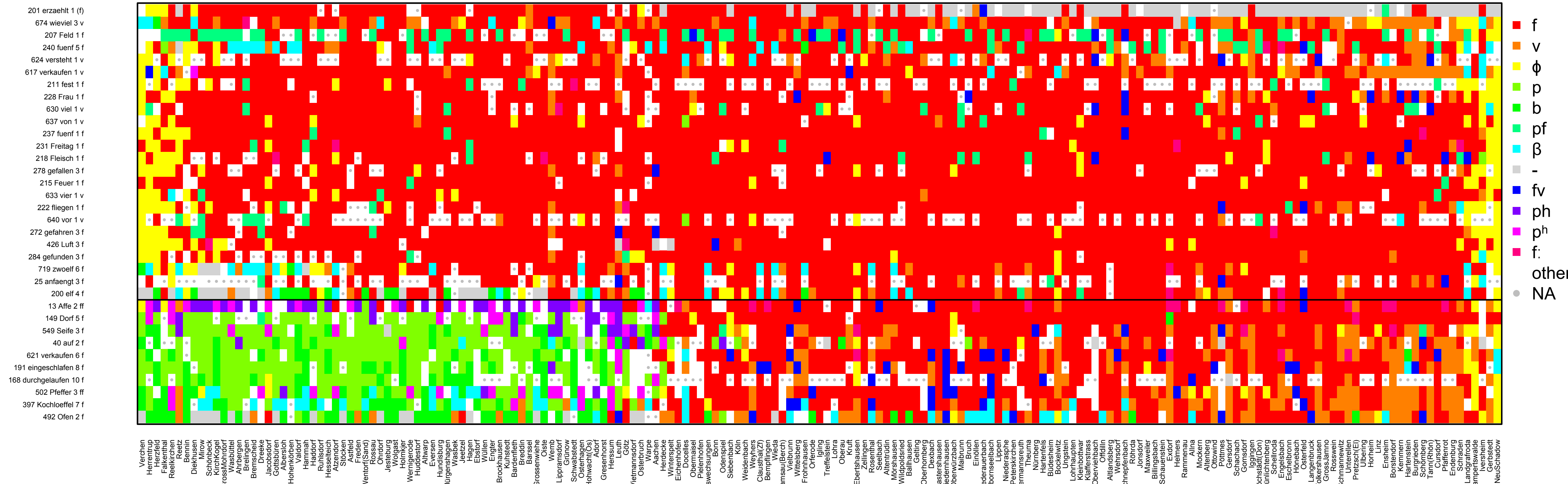

-
n
t
ɔ
l
s
ɾ
m
i
ɛ
p
ʃ
β
ɔ
k
o
d
v
a
b
e
u
th
i
ɾ
a
fi
ɐɪ
ɯ
x
z
ts
c
ŋ
ɣ
œ
e
g
kh
u
ø
ɾɪ
aɪ
ɐʊ
j
ɒ
ɤ
ɣ
æ
θ
ɸ
pf
ou
ô
ʏ
ʌ
ks
kx
l
eɪ
other
NA

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 sch
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 Fleisch 4 sch
587 Tisch 3 sch
156 Dreschen 7 sch

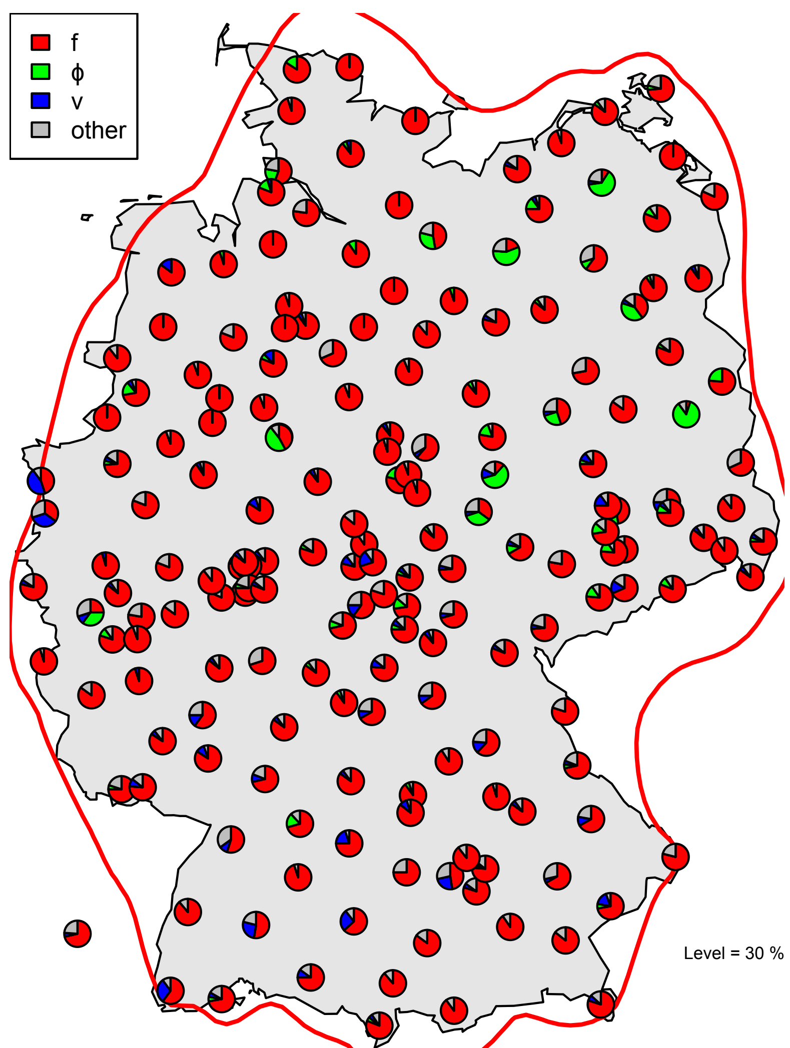


Correspondences “sch”

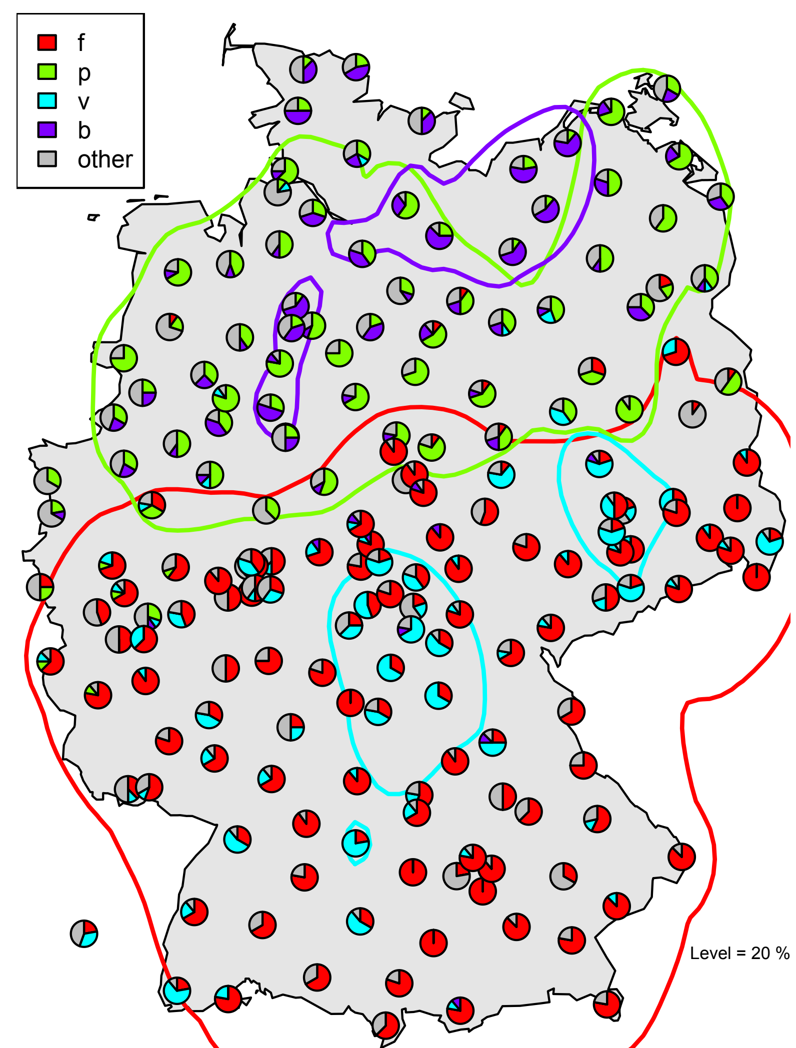




Correspondences “f”

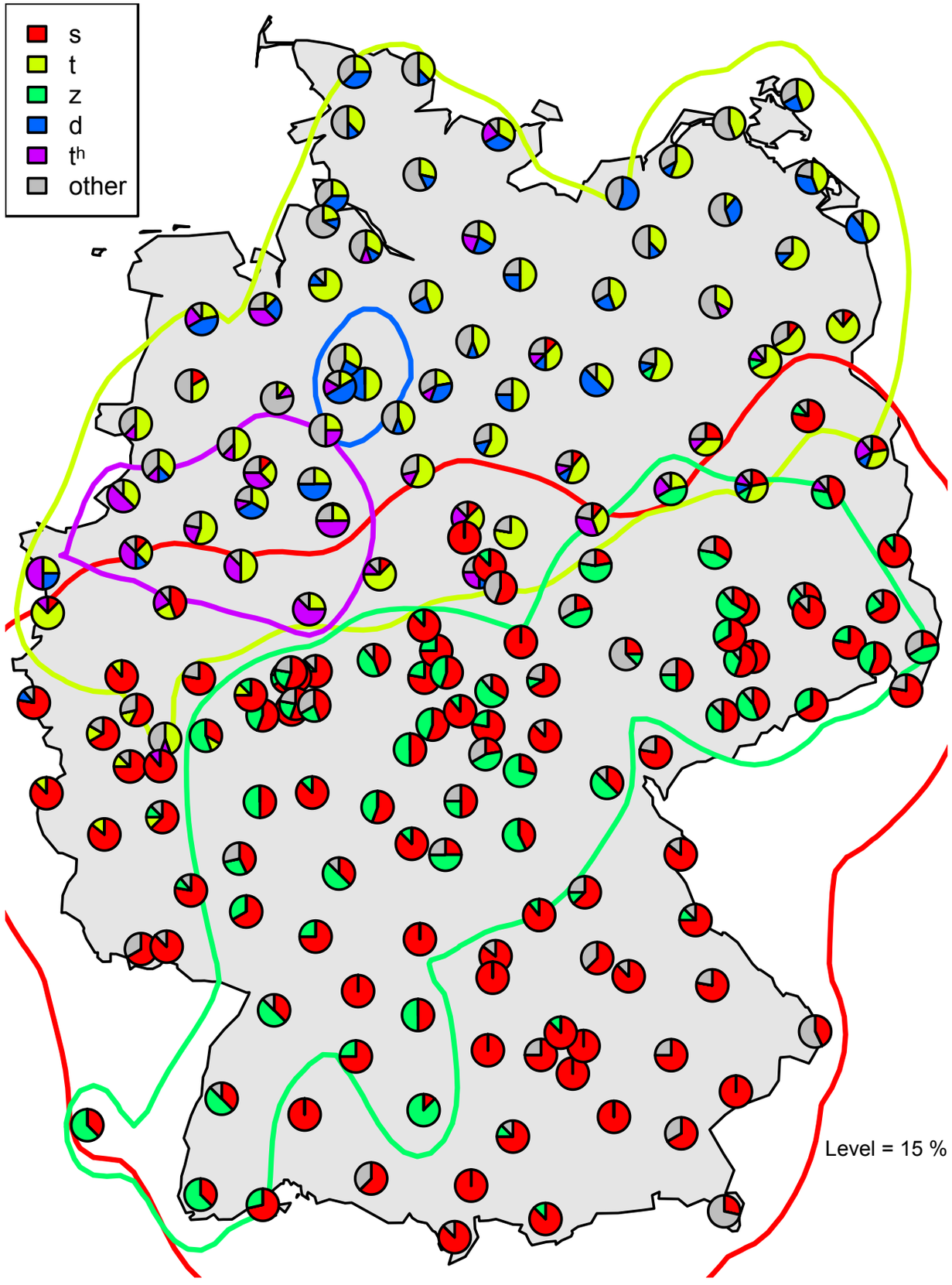


Correspondences “f/p”

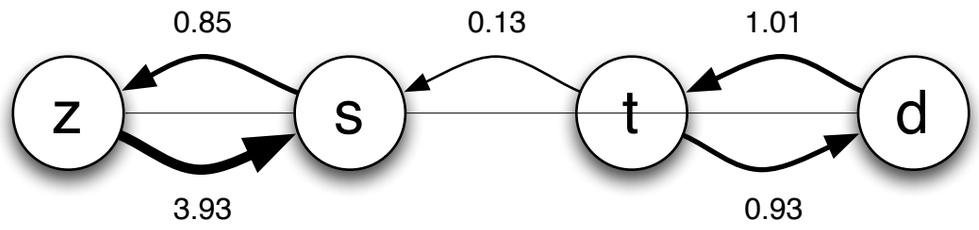


Step 4: Sound Change Modelling

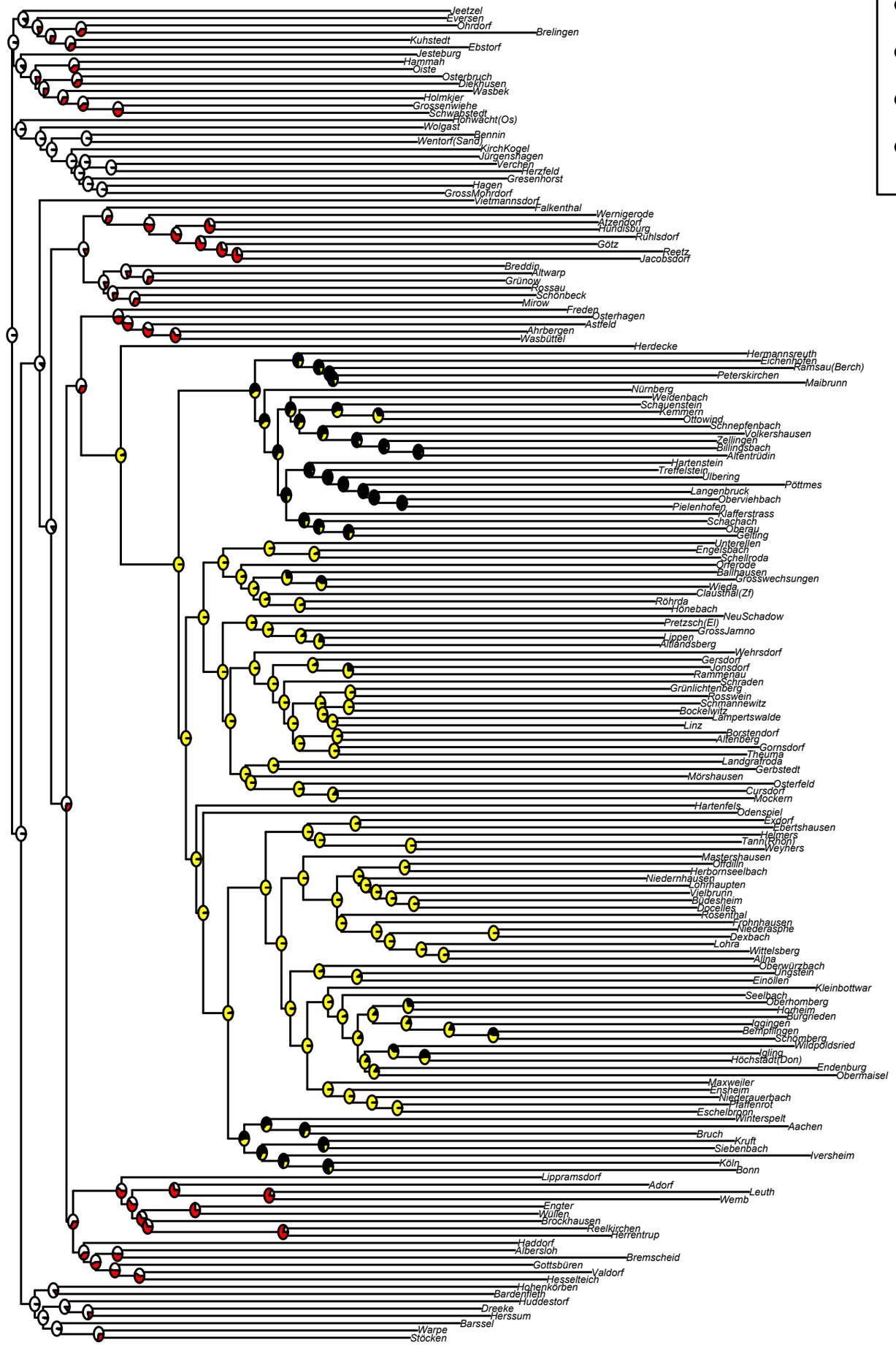
Correspondences “s/t”



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



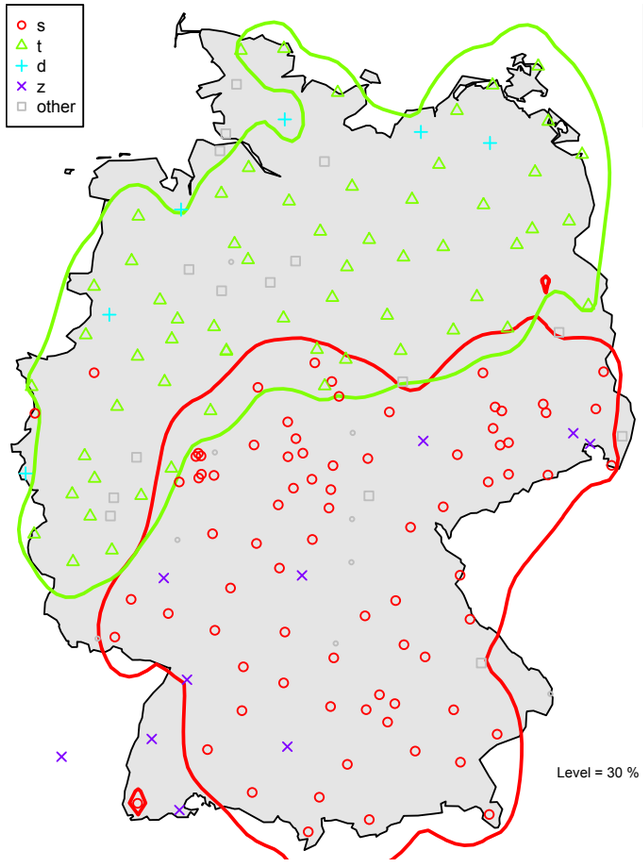
Based on /s/ in German words:
beißen, besser, das, größer, groß, heiß, muss, Wasser, weiße


d
s
t
z
Jeetzel
Eversen
Ohrdorf
Brelingen
Kuhstedt
Ebstorf
Jesteburg
Hammah
Oiste
Osterbruch
Diekhusen
Wasbek
Holmkjer
Grosseniwehe
Schwabstedt
Hohwacht(Os)
Wolgast
Bennin
Wentorf(Sand)
KirchKogel
Jürgenshagen
Verchen
Herzfeld
Gresenhorst
Hagen
GrossMohrdorf
Vietmannsdorf
Falkenthal
Wernigerode
Atzendorf
Hundisburg
Ruhlsdorf
Götz
Reetz
Jacobsdorf
Breddin
Altwarp
Grünow
Rossau
Schönbeck
Mirow
Freden
Osterhagen
Astfeld
Ahrbergen
Wasbüttel
Herdecke
Hermannsreuth
Eichenhofen
Ramsau(Berch)
Peterskirchen
Maibrunn
Nürnberg
Weidenbach
Schauenstein
Kemmern
Ottowind
Schnepfenbach
Volkershausen
Zellingen
Billingsbach
Altentrüdin
Hartenstein
Treffelstein
Ulbering
Pöttmes
Langenbruck
Oberviehbach
Pielenhofen
Klafferstrass
Schachach
Oberau
Celting
Unterellen
Engelsbach
Schellroda
Orferode
Ballhausen
Grosswechsungen
Wieda
Clausthal(Zf)
Röhrda
Hönebach
NeuSchadow
Pretzsch(El)
GrossJamno
Lippen
Altlandsberg
Wehrsdorf
Gersdorf
Jonsdorf
Rammenau
Schraden
Grünlichtenberg
Rosswein
Schmannewitz
Bockelwitz
Lampertswalde
Linz
Borstendorf
Altenberg
Gornsdorf
Theuma
Landgrafroda
Gerbstedt
Mörshausen
Osterfeld
Cursdorf
Mockern
Hartenfels
Odenspiel
Exdorf
Ebertshausen
Helmers
Tann(Rhön)
Weyhers
Mastershausen
Offdilln
Herbornseelbach
Niedernhausen
Lohrhaupten
Vielbrunn
Büdesheim
Docelles
Rosenthal
Tonnhausen
Niederasphe
Dexbach
Lohra
Wittelsberg
Allna
Oberwürzbach
Ungstein
Einöllen
Kleinbottwar
Seelbach
Obernhomberg
Horheim
Burgrieden
Iggingen
Bempflingen
Schömberg
Wildpoldsried
Igling
Höchstadt(Don)
Endenburg
Obermaisel
Maxweiler
Ensheim
Niederauerbach
Hatzenrot
Eschelbronn
Winterspelt
Aachen
Bruch
Kruft
Siebenbach
Iversheim
Köln
Bonn
Lippramsdorf
Adorf
Leuth
Wemb
Engter
Wüllen
Brockhausen
Reelkirchen
Herrentrup
Haddorf
Albersloh
Bremscheid
Gottsbüren
Valdorf
Hesselteich
Hohenkörben
Bardenfleth
Huddestorf
Dreeke
Herssum
Barssel
Warpe
Stöcken

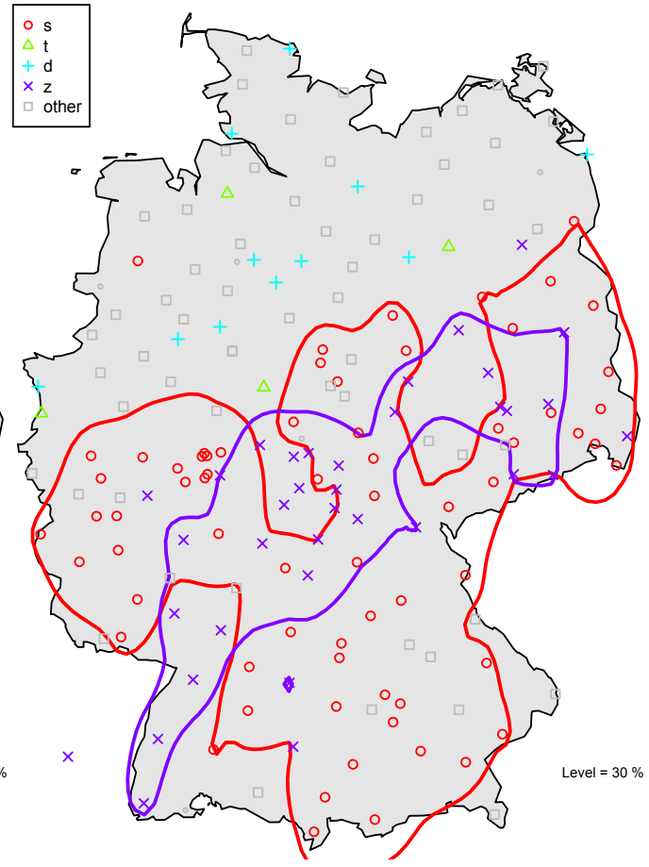
Conclusions

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

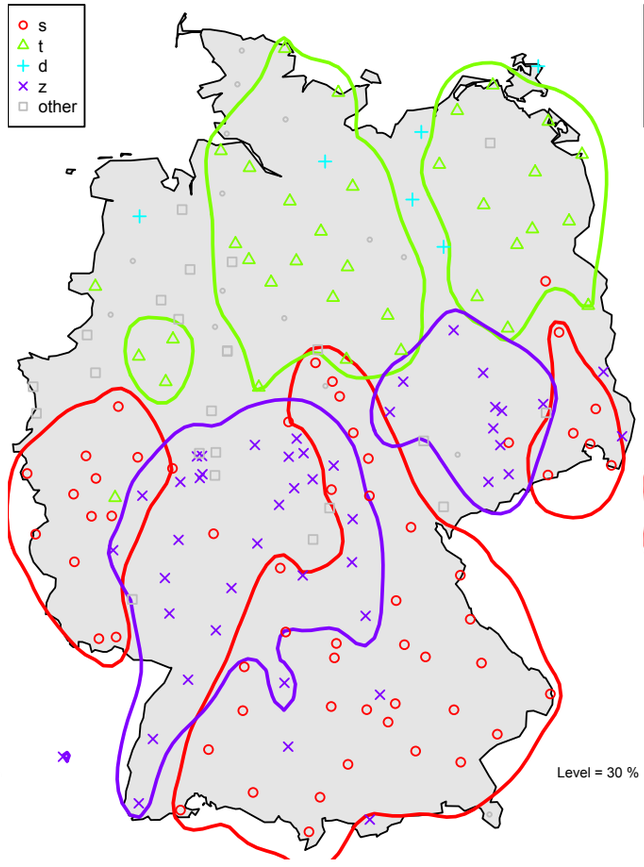
121 das 3 s



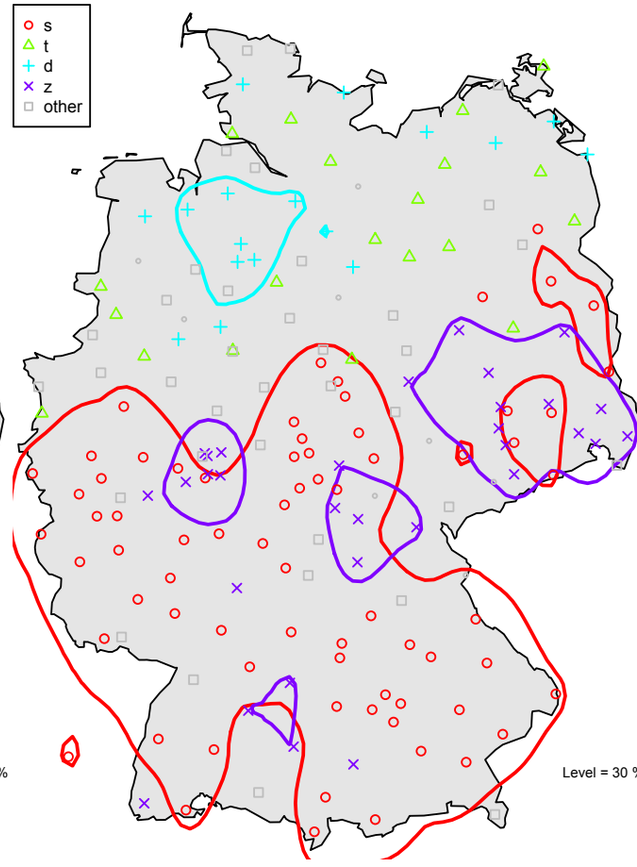
81 besser 3 ss



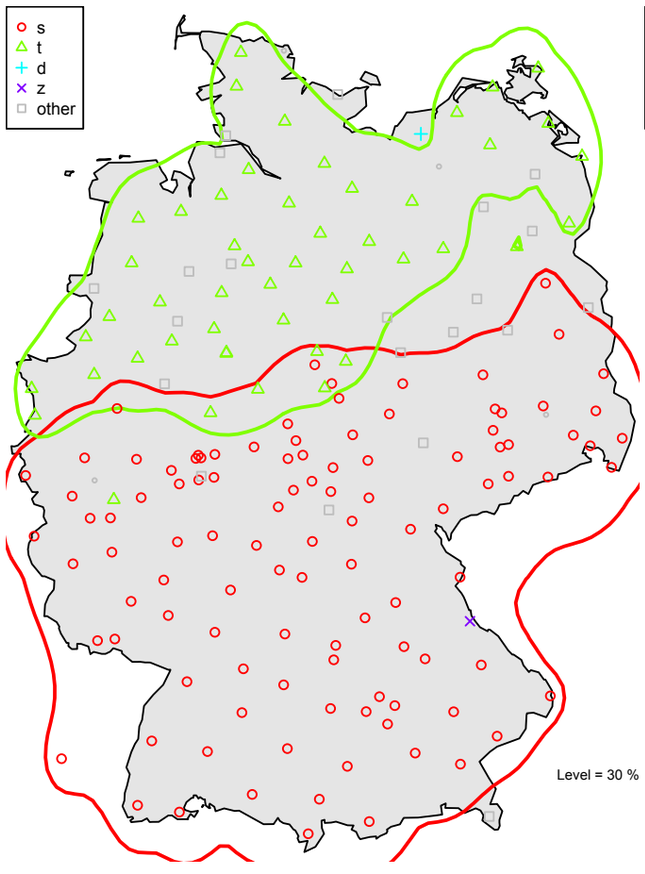
72 beissen 3 ß



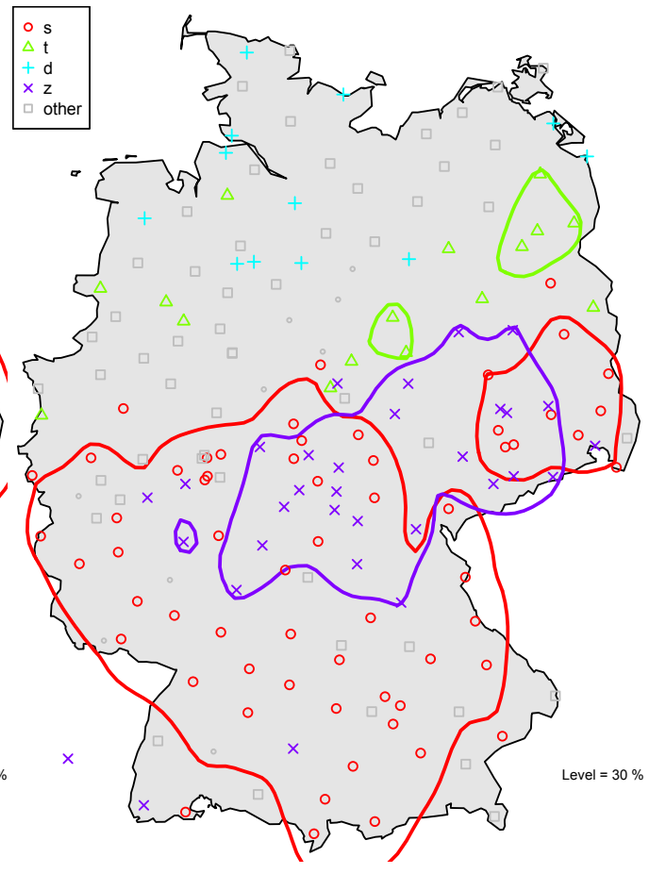
660 weisse 3 ß



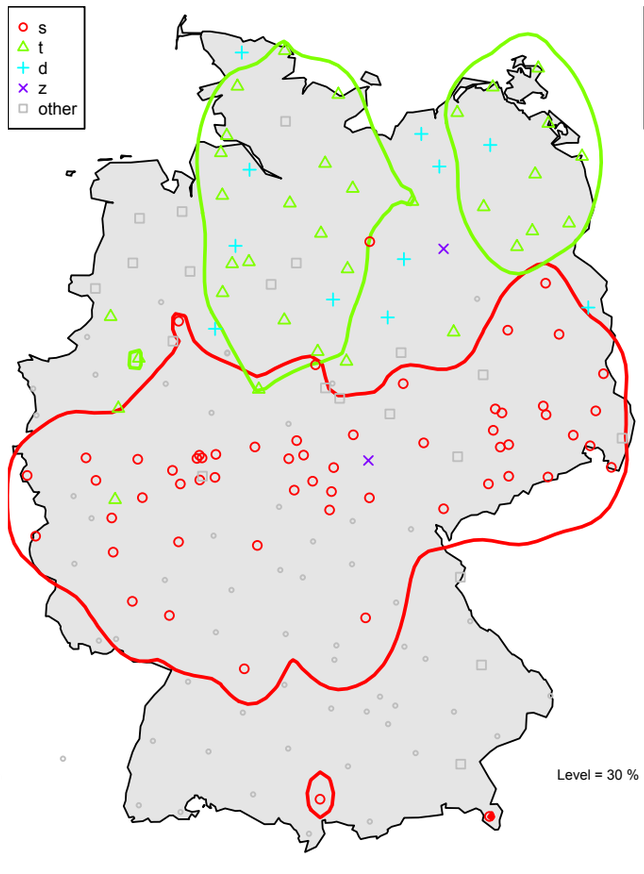
341 gross 4 ß



651 Wasser 3 ss



473 muss 3 ss



357 heiss 3 ss