Quantifying models of linguistic diversity and change

Michael Cysouw
Philipps-University Marburg

- forget about trees,
  focus on reconstruction

- treat change in meaning/function
  just like change in form

Central Approach: Multialignment

- Alignments of sounds
- Alignments of words

In linguistics,
an alignment is a central result,
not an intermediate method

Multialignment of words

- Based on a sentence-by-sentence alignment, induce word-by-word alignment
- Translations can be (and often are!) quite different
- Bi-text alignment is widely researched problem
- Mulit-text alignment not so much

Kong Herodes blev skrækslagen , og Jerusalem begyndte at summe af rygter .

Als dies dem König Herodes zu Ohren kam , erschrak er , und mit ihm entsetzte sich auch ganz Jerusalem .

But Herodes the king heard , and was troubled , and all Urishlem with him .

Þegar Heródes heyrði þetta , varð hann skelkaður og öll Jerúsalem með honum .

Als des dr Kenig Herodes ghärt het , isch scha (er) arg vuschrocke un mit nem ganz Jerusalem ,

Kort voor lank het Herodes ook van die geleerdes uit die ooste se storie te hore gekom . Hy was baie omgekrap oor wat hulle oor die nuwe Joodse koning gesê het . So ook die res van Jerusalem .

Der König Herodes war total aufgebracht , als er das hörte , und nicht nur er , alle in Jerusalem waren das .

Kong Herodes blev skrækslagen , og Jerusalem begyndte at summe af rygter .

Als dies dem König Herodes zu Ohren kam , erschrak er , und mit ihm entsetzte sich auch ganz Jerusalem .

But Herodes the king heard , and was troubled , and all Urishlem with him .

Þegar Heródes heyrði þetta , varð hann skelkaður og öll Jerúsalem með honum .

Als des dr Kenig Herodes ghärt het , isch scha (er) arg vuschrocke un mit nem ganz Jerusalem ,

Kort voor lank het Herodes ook van die geleerdes uit die ooste se storie te hore gekom . Hy was baie omgekrap oor wat hulle oor die nuwe Joodse koning gesê het . So ook die res van Jerusalem .

Der König Herodes war total aufgebracht , als er das hörte , und nicht nur er , alle in Jerusalem waren das .

Kong Herodes blev skrækslagen , og Jerusalem begyndte at summe af rygter .

Als dies dem König Herodes zu Ohren kam , erschrak er , und mit ihm entsetzte sich auch ganz Jerusalem .

But Herodes the king heard , and was troubled , and all Urishlem with him .

Þegar Heródes heyrði þetta , varð hann skelkaður og öll Jerúsalem með honum .

Als des dr Kenig Herodes ghärt het , isch scha (er) arg vuschrocke un mit nem ganz Jerusalem ,

Kort voor lank het Herodes ook van die geleerdes uit die ooste se storie te hore gekom . Hy was baie omgekrap oor wat hulle oor die nuwe Joodse koning gesê het . So ook die res van Jerusalem .

Der König Herodes war total aufgebracht , als er das hörte , und nicht nur er , alle in Jerusalem waren das .

Kong Herodes blev skrækslagen , og Jerusalem begyndte at summe af rygter .

Als dies dem König Herodes zu Ohren kam , erschrak er , und mit ihm entsetzte sich auch ganz Jerusalem .

But Herodes the king heard , and was troubled , and all Urishlem with him .

Þegar Heródes heyrði þetta , varð hann skelkaður og öll Jerúsalem með honum .

Als des dr Kenig Herodes ghärt het , isch scha (er) arg vuschrocke un mit nem ganz Jerusalem ,

Kort voor lank het Herodes ook van die geleerdes uit die ooste se storie te hore gekom . Hy was baie omgekrap oor wat hulle oor die nuwe Joodse koning gesê het . So ook die res van Jerusalem .

Der König Herodes war total aufgebracht , als er das hörte , und nicht nur er , alle in Jerusalem waren das .

Universal Declaration of Human Rights

in, im	German
in	Dutch
in	English
yn	Frisian
in	Afrikaans
in	Scots
אין	Yiddish
í	Icelandic
i	Swedisch
i	Nynorsk
i	Bokmål
í	Faroese
i	Danish

Article 10

English: "everyone is entitled in full equality to a fair and public hearing by an independent and impartial tribunal , **in the determination of his rights** and obligations and of any criminal charge against him ."

Afrikaans: "elkeen het , in volle gelykheid , die reg tot 'n regverdige en openbare verhoor deur 'n onafhanklike en objektiewe tribunaal , **in die bepaling van sy regte** en verpligtinge en die ondersoek van enige kriminele saak teen hom ."

German: "jeder hat **bei der feststellung seiner rechte** und pflichten sowie bei einer gegen ihn erhobenen strafrechtlichen beschuldigung in voller gleichheit anspruch auf ein gerechtes und öffentliches verfahren vor einem unabhängigen und unparteiischen gericht ."

Preamble, sentence 5

English:

“whereas the peoples of the united nations have in the charter reaffirmed their faith in fundamental human rights , in the dignity and worth of the human person and in the equal rights of men and women and have determined to promote social progress and better standards of life in larger freedom ,”

German:

“da die völker der vereinten nationen in der charta ihren glauben an die grundlegenden menschenrechte , an die würde und den wert der menschlichen person und an die gleichberechtigung von mann und frau erneut bekräftigt und beschlossen haben , den sozialen fortschritt und bessere lebensbedingungen in größerer freiheit zu fördern ,”

Preamble, sentence 5

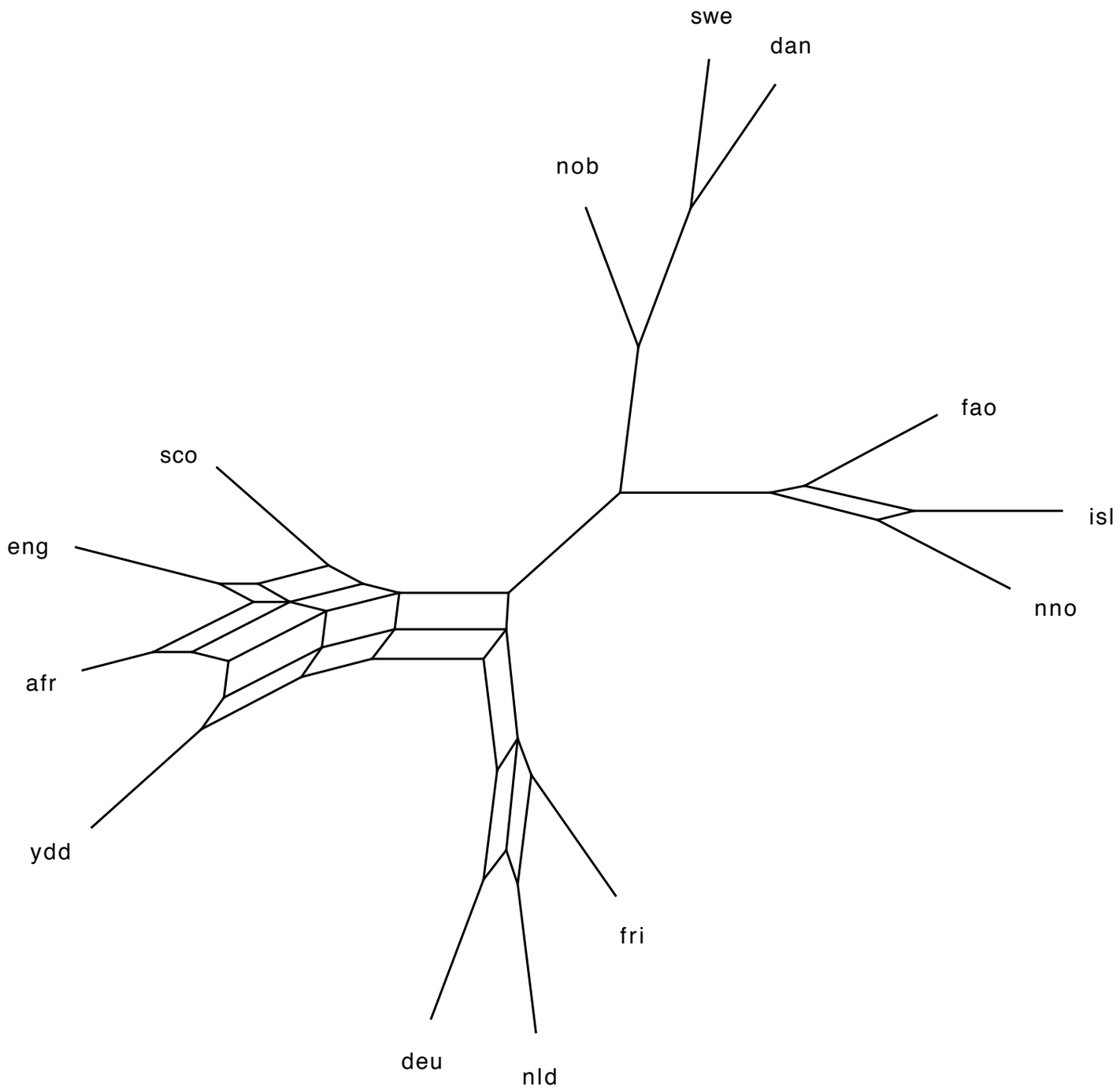
English:

"whereas the peoples of the united nations have ***in the charter*** reaffirmed their faith ***in fundamental human rights*** , ***in the dignity*** and worth of the human person and ***in the equal rights*** of men and women and have determined to promote social progress and better standards of life ***in larger freedom*** ,"

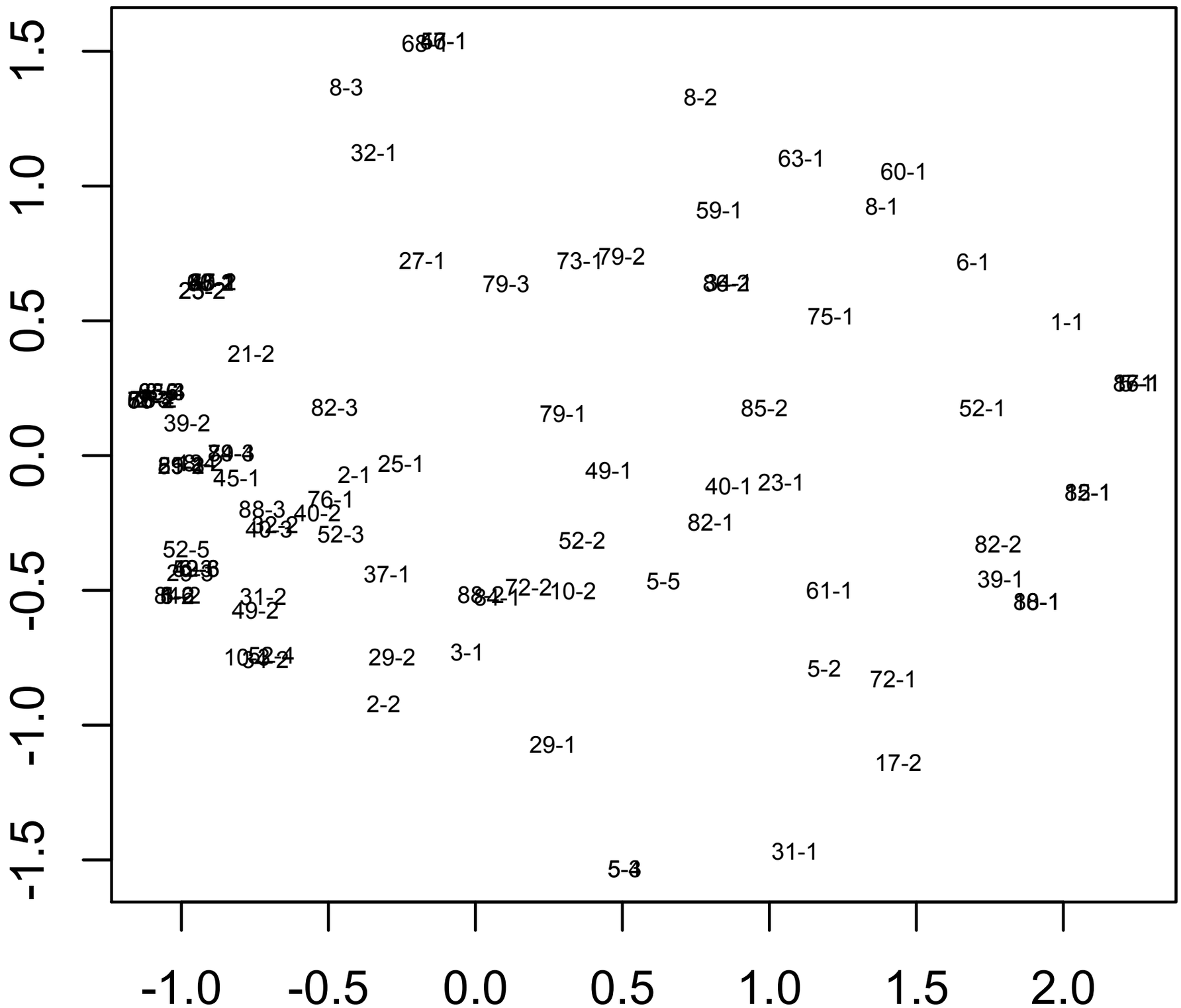
German:

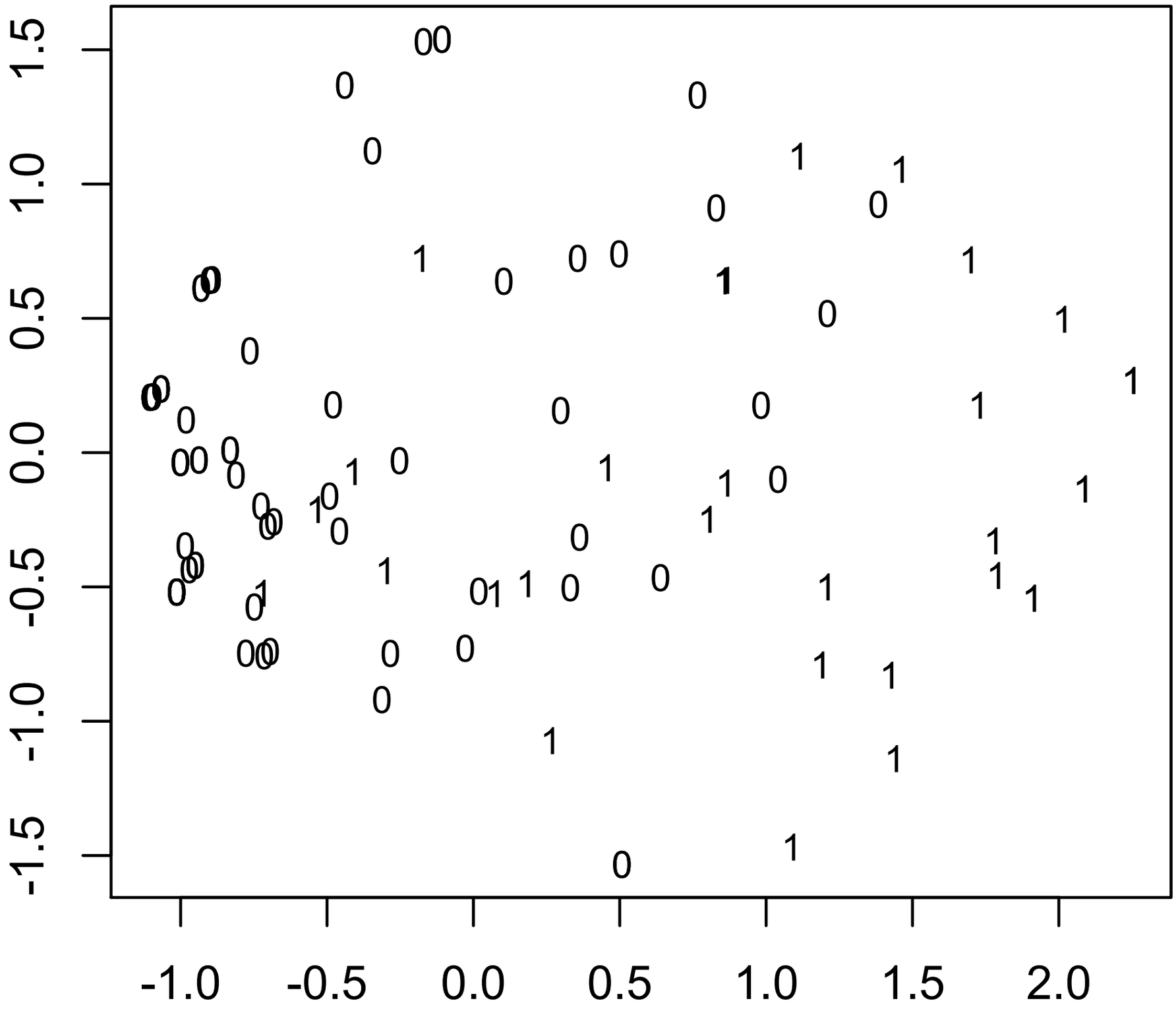
"da die völker der vereinten nationen ***in der charta*** ihren glauben ***an die grundlegenden menschenrechte , an die würde*** und den wert der menschlichen person und ***an die gleichberechtigung*** von mann und frau erneut bekräftigt und beschlossen haben , den sozialen fortschritt und bessere lebensbedingungen ***in größerer freiheit*** zu fördern ,"

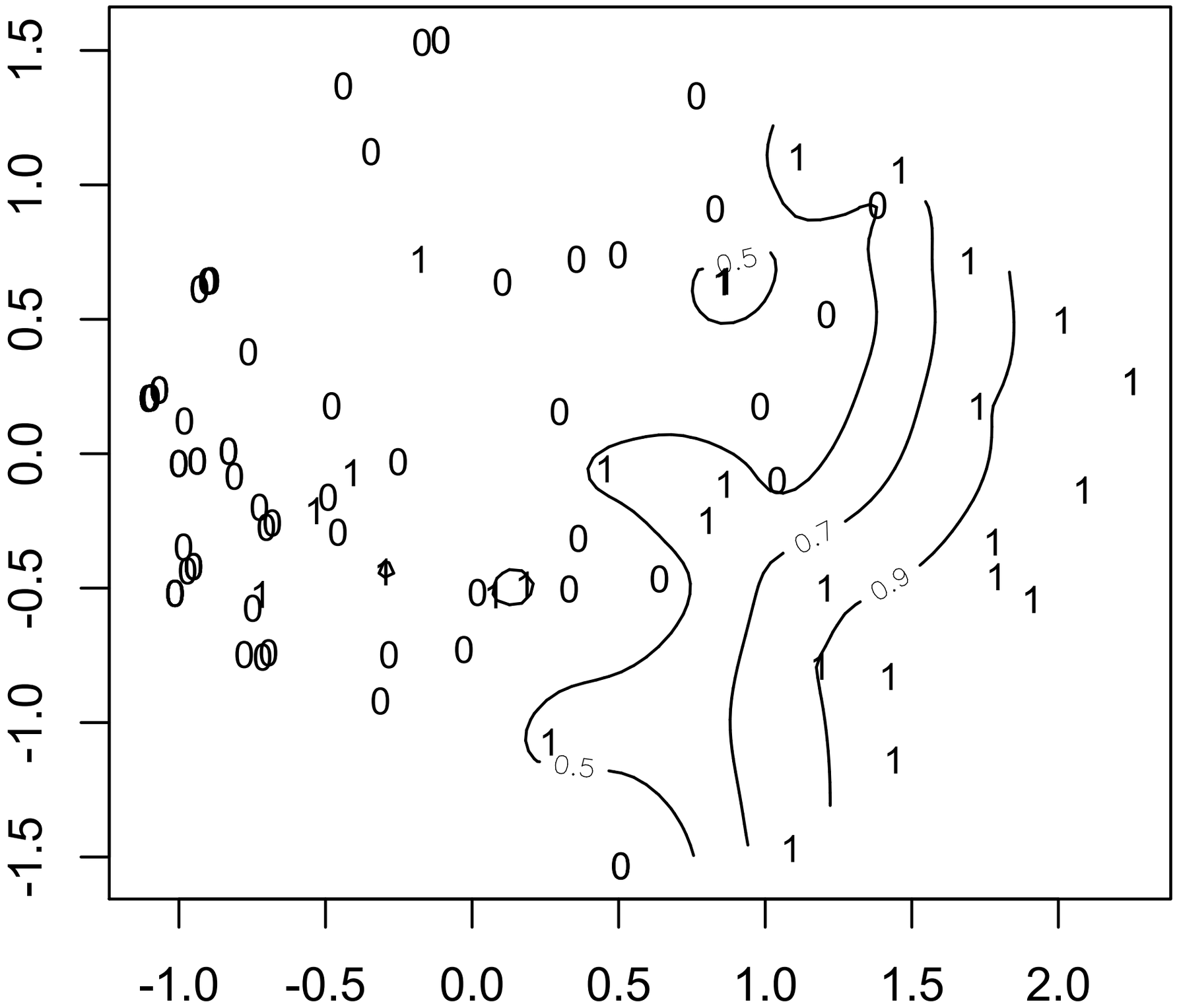
deu	2		1				1
eng	5		1	1	1	1	1
nld	5		1	1	1	1	1
fri	6	1	1	1	1	1	1
afr	5		1	1	1	1	1
ydd	4		1	1	1		1
sco	5		1	1	1	1	1
nno	3		1			1	1
nob	1		1				
swe	1		1				
fao	3		1			1	1
dan	1		1				
isl	1						1

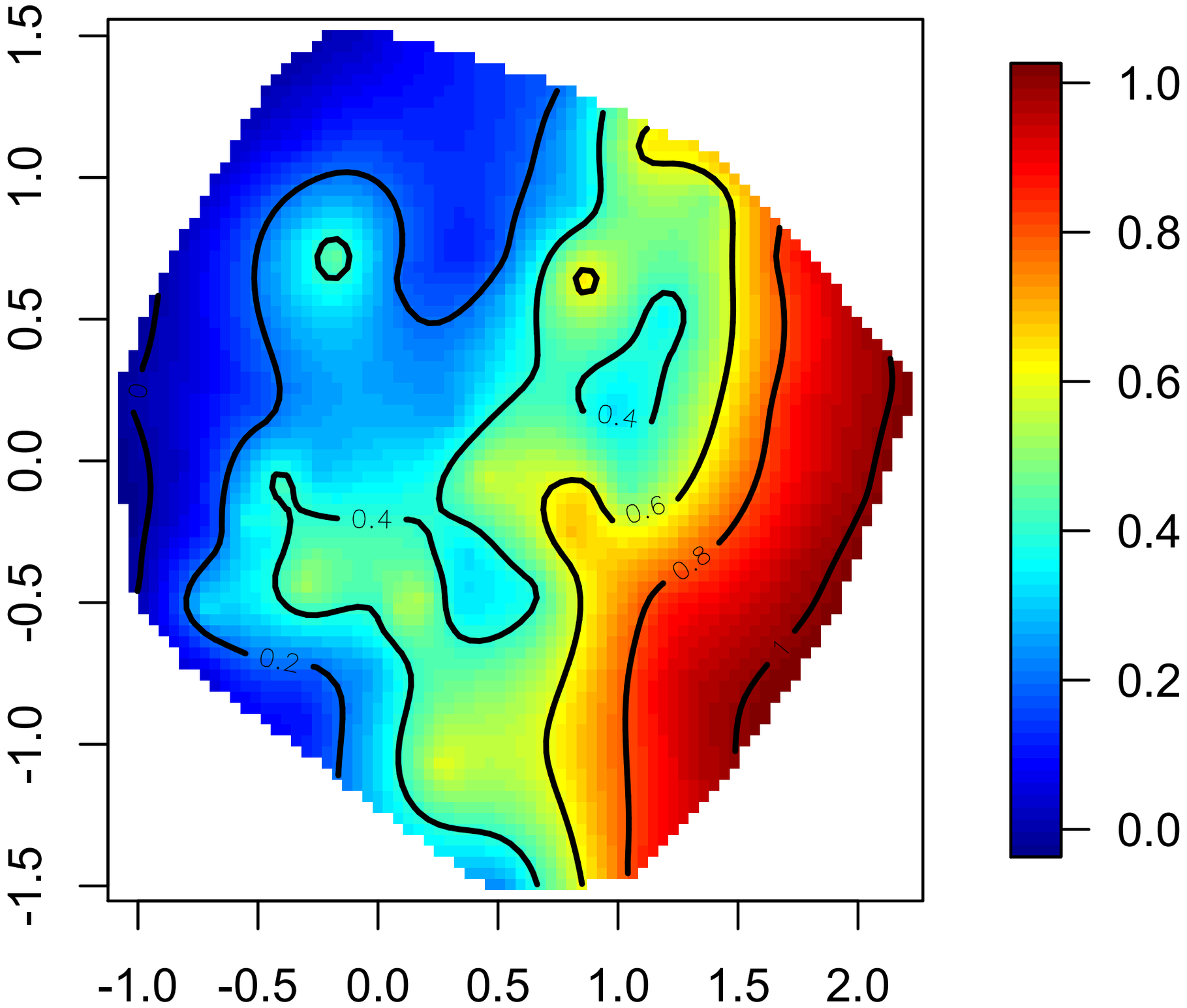


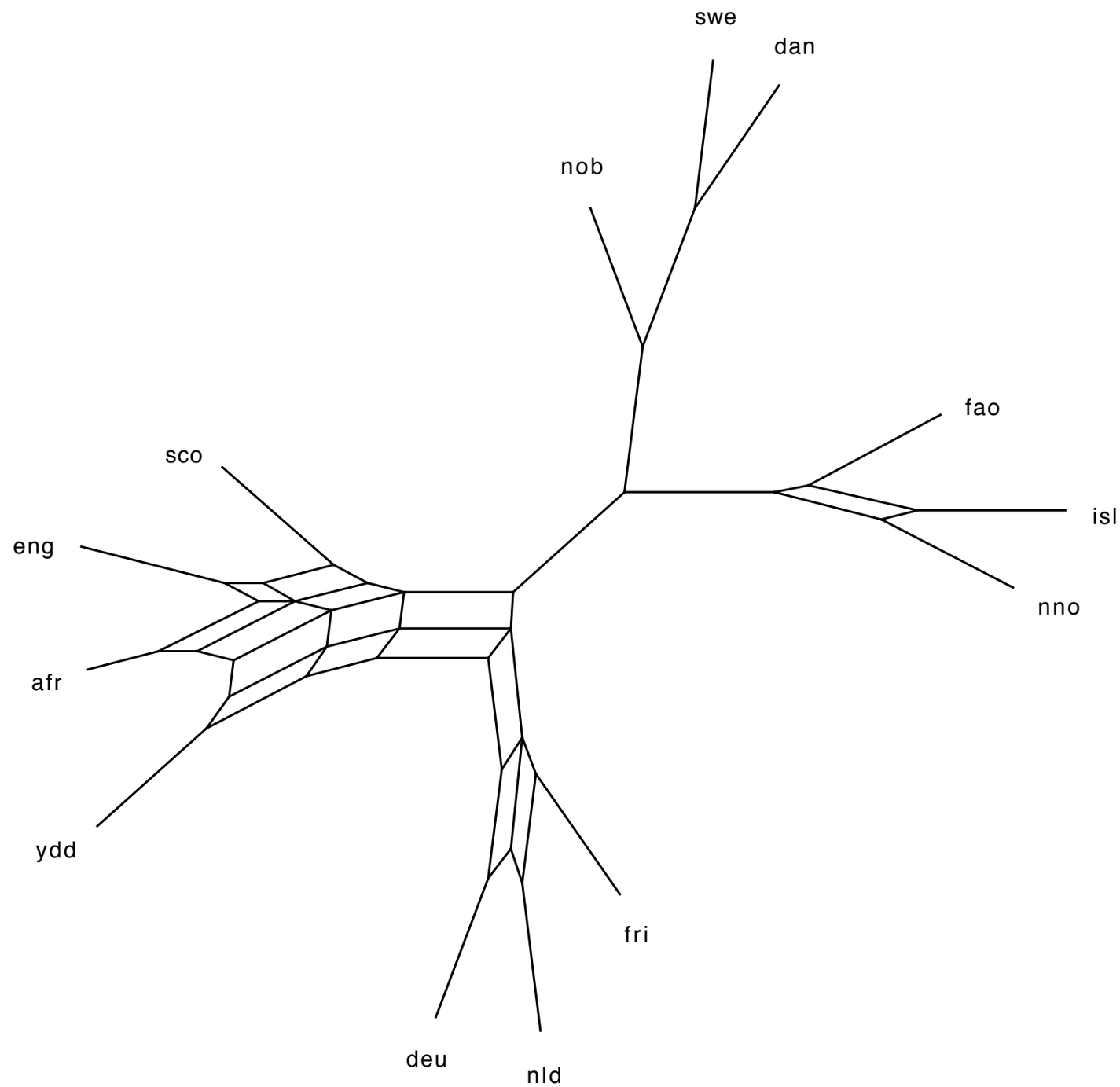
Consensus network of 4 optimal trees according to dollo Maximum Parsimony

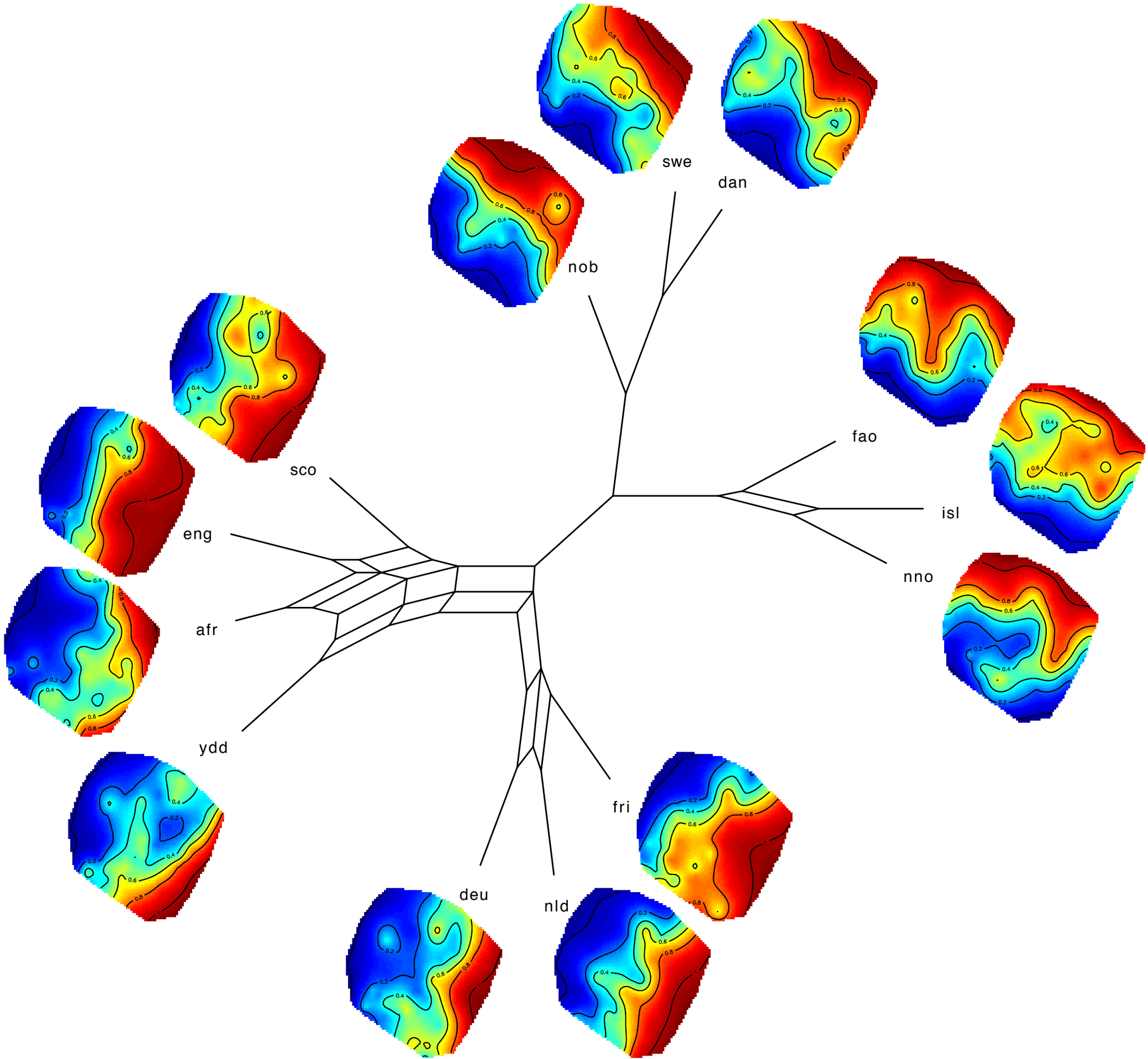

1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
8-3
8-2
32-1
63-1
60-1
59-1
8-1
27-1
73-1
79-2
79-3
6-1
75-1
1-1
21-2
82-3
79-1
85-2
52-1
39-2
49-1
40-1
23-1
45-1
2-1
25-1
76-1
88-3
40-2
82-1
52-3
52-2
82-2
52-5
37-1
5-5
39-1
61-1
31-2
49-2
72-2
10-2
3-1
29-2
5-2
72-1
2-2
29-1
17-2
31-1


1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0


1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
0.5
0.7
0.9
0.5


1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
0
0.2
0.4
0.4
0.6
0.8
1
1.0
0.8
0.6
0.4
0.2
0.0


swe
dan
nob
fao
sco
isl
eng
nno
afr
ydd
fri
deu
nld


swe
dan
nob
sco
fao
eng
isl
nno
afr
ydd
fri
deu
nld

Multialignment of Sounds

Lfd. Nr.	Stichwort		Transkription	Bemerkungen
1	Winter	178	ˈvɪn̥tʰə˞	˞= kontinuant ə bereits velarisiert
2	fliegen	56	ˈflaɪ̯ə̯tʰə	'fliegen die', Sequenzierung unklar - kein gemeinsames [L]
3	Blätter	23	ˈblɛ̝ːdə̯˞̊	
4	Luft	103	lʊrˈftʰ‿	˞= kontinuant
5	hört	89	hɪ̯˞̥t‿	˞= kontinuant
6	gleich	78	‿klɛ̝ːɪ̯k‿	folgt P
7	schneien	130	ˈʃnɛɪ̯ən	
8	Wetter	174	/	statt dessen ˈvɪtəʀʊŋə
9	tu	151	dɛʊ	

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	‿klaɪ̯ç‿	folgt P
7	schneien	130	ˈʃnaɪ̯ən	
8	Wetter	174	/	statt dessen ˈvɪtəʀʊŋə
9	tu	151	dəʊ	

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

LOCATION	WORD
Aachen	a̠ːph
Adorf	ɑ̟ːbʱə
Ahrbergen	o̜→ɔ̟phə̣
Albersloh	ɑːpʱ̮ə̤
Allna	ɑɸh
Altenberg	ʌfɛ̤
Altentrüdin	a̠f
Altlandsberg	ɑˀfəˀ
Altwarp	oːph
Astfeld	ɒˤːpʰə
Atzendorf	afɛ̠
Ballhausen	ʌˀfə
Bardenfleth	ɔ̣ːp͡ɸ
Barssel	ɒ̣ːpʱə
Bempflingen	afː
Bennin	ɔ̜pʰ
Billingsbach	a̠f
Bockelwitz	ʌ̟və̤
Bonn	ɑːp’
Borstendorf	ɤ̜fː
Breddin	ɒːph
Brelingen	ɑ̟f͡βə
Bremscheid	ɒˤːpḫə̟
…	…

A	FF	E
a̠ː	ph	-
ɑ̟ː	bʱ	ə
o̜→ɔ̟	ph	ə̣
ɑː	pʱ̮	ə̤
ɑ	ɸh	-
ʌ	f	ɛ̤
a̠	f	-
ɑˀ	f	əˀ
oː	ph	-
ɒˤː	pʰ	ə
a	f	ɛ̠
ʌˀ	f	ə
ɔ̣ː	p͡ɸ	-
ɒ̣ː	pʱ	ə
a	fː	-
ɔ̜	pʰ	-
a̠	f	-
ʌ̟	v	ə̤
ɑː	p’	-
ɤ̜	fː	-
ɒː	ph	-
ɑ̟	f͡β	ə
ɒˤː	pḫ	ə̟
…	…	…

- **Workflow:**
  - Tokenisation of segments (github.com/cysouw/qlcData)
  - 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	œˁ	x	-
Altenberg	ɤɣ̩	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	-

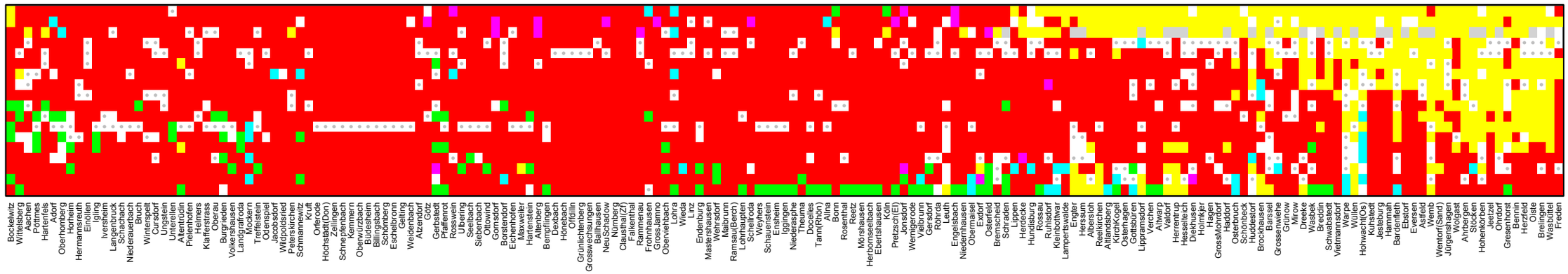
github.com/digitallinguist/msa-editor

Correspondence Sets

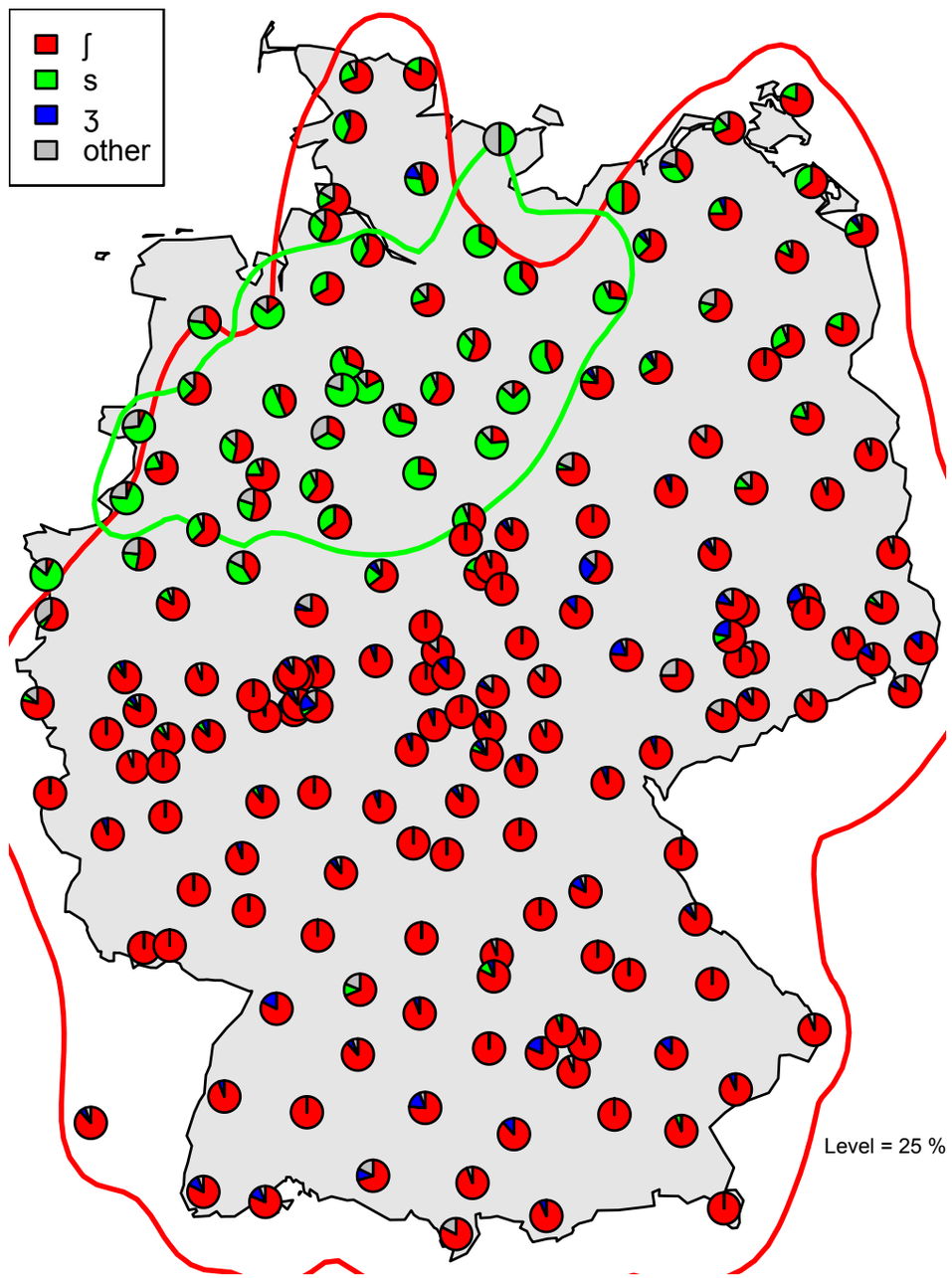
- more than 700 columns of aligned segments (“correspondences”)
- Comparative-historical linguistics uses clusters of correspondences (“correspondence sets”)
- Automatic clustering of columns is a good start, but needs correction
- Visualisations in R
  github.com/cysouw/qlcVisualize

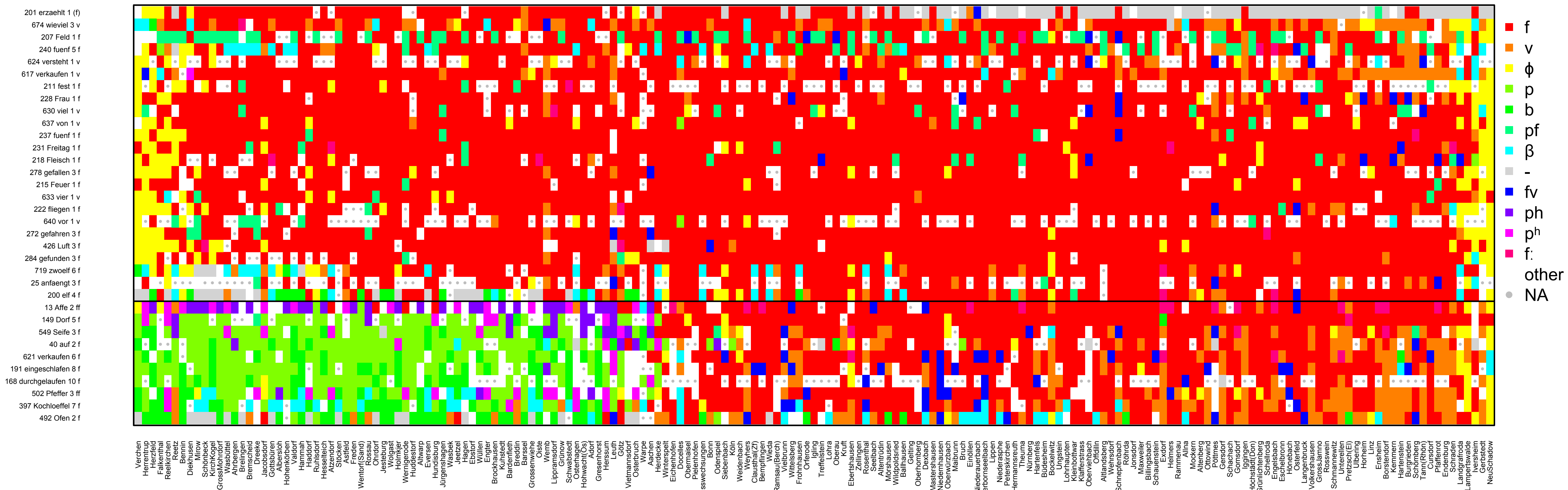
limage ("level-image")
from R-package qlcVisualize
(Cysouw 2015)

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

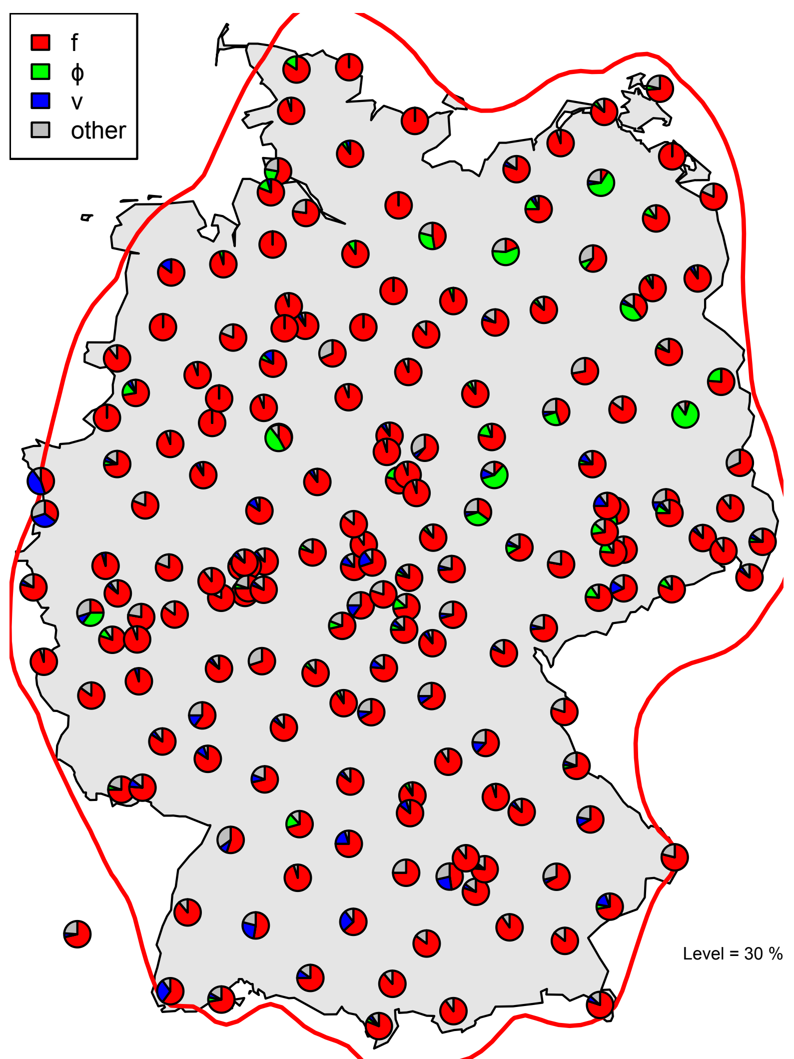


Correspondences “sch”

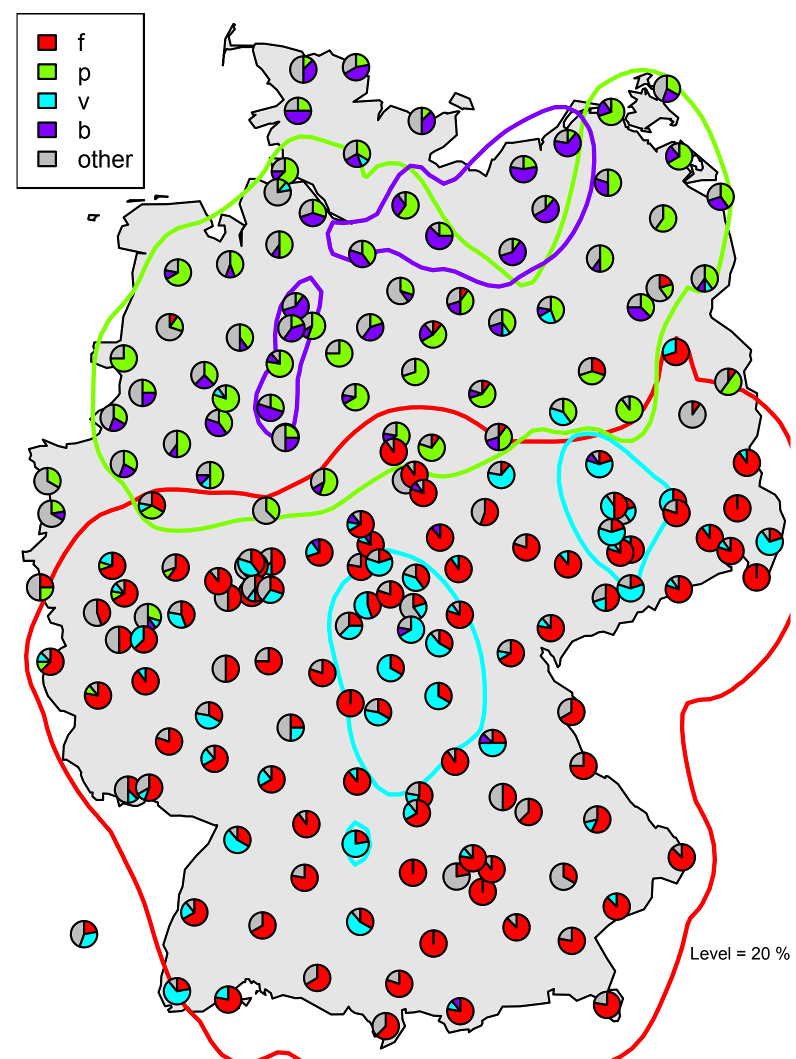




Correspondences “f”



Correspondences “f/p”



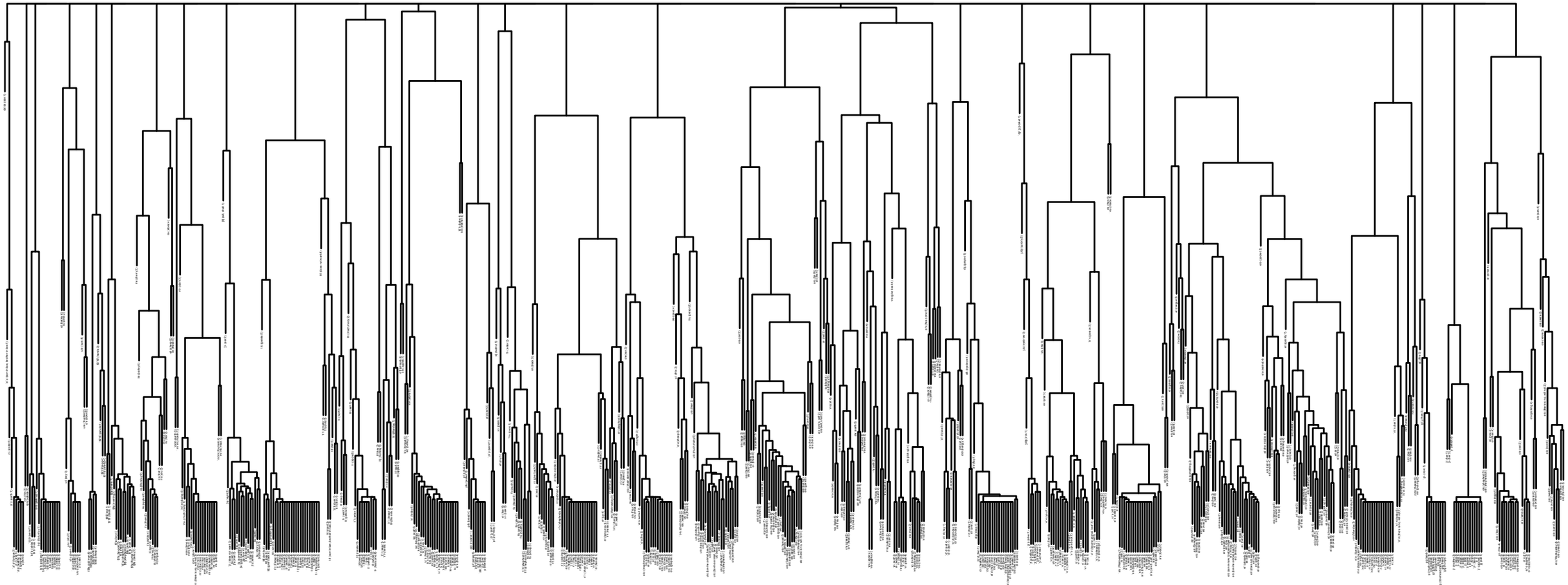
Bulgarian dialect data

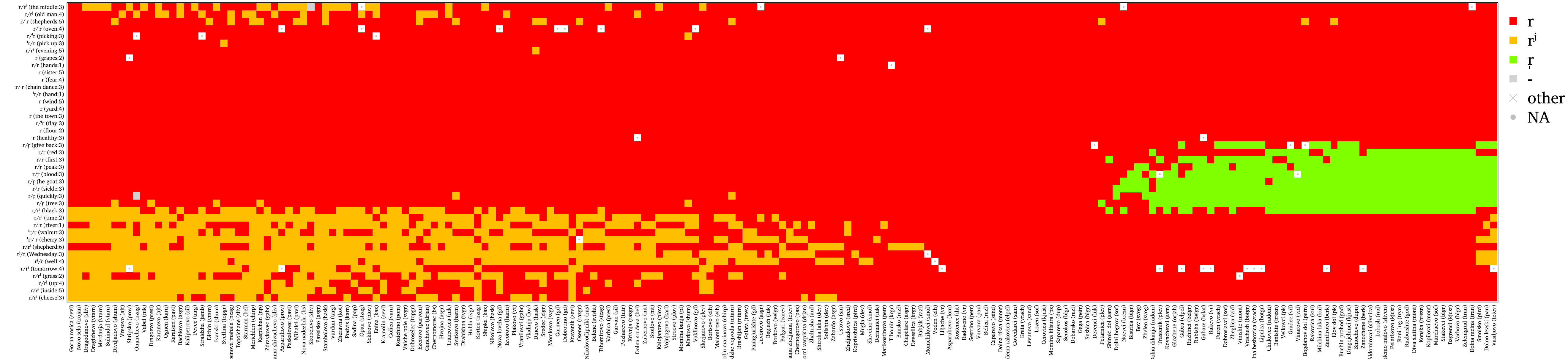
- Data from Jelena Prokić
- Taken from the ***benchmark database of phonetic alignments*** (BDPA)

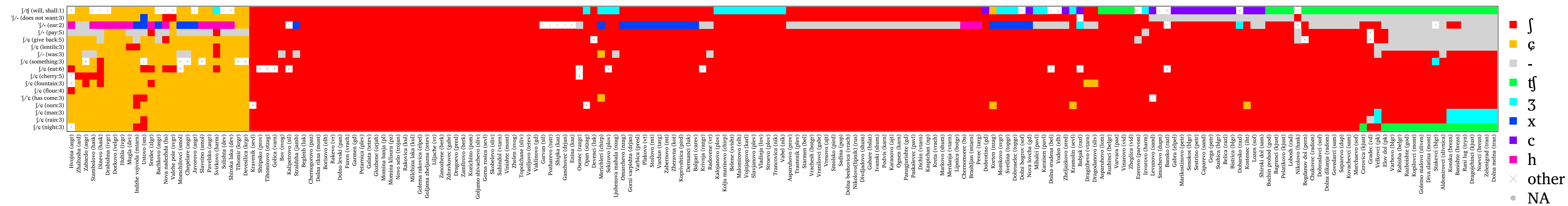
List, Johann-Mattis and Jelena Prokić. (2014). A benchmark database of phonetic alignments in historical linguistics and dialectology. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), May 2014, Reykjavik. 288-294.

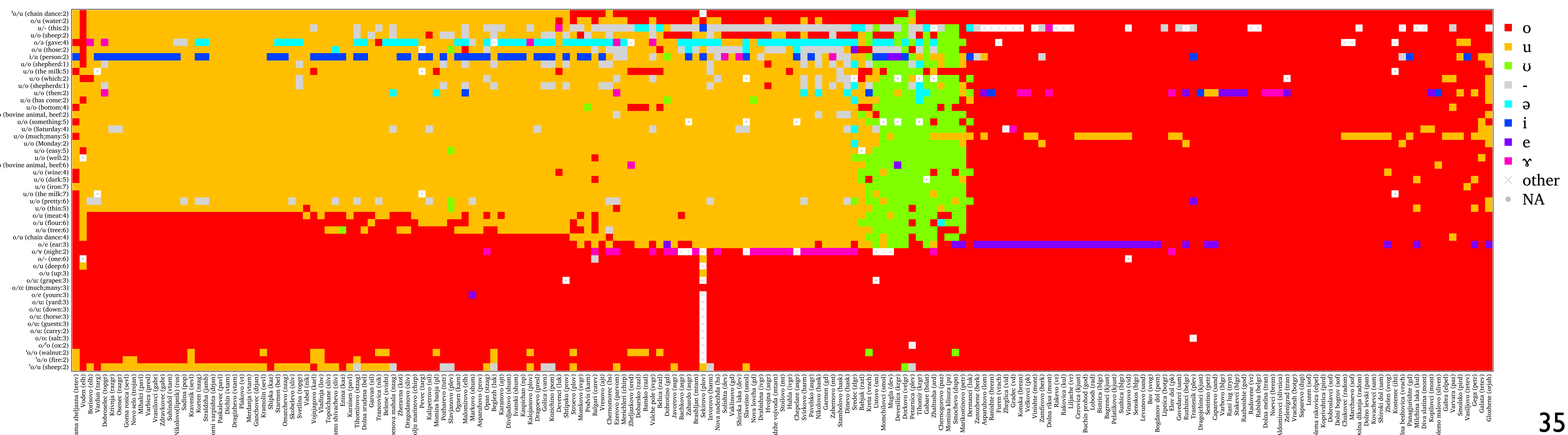
Bulgarian dialect data

- Data from 197 villages for 152 words, resulting in 807 aligned columns
- There are many possible methods to cluster these into correspondence sets
- And they give different results!
  (though mostly smaller or larger sets)
- Here: a simple greedy clustering into 27 sets




r
rʲ
r̥
-
other
NA


ʃ
ɕ
-
ʧ
ʒ
x
c
h
other
NA

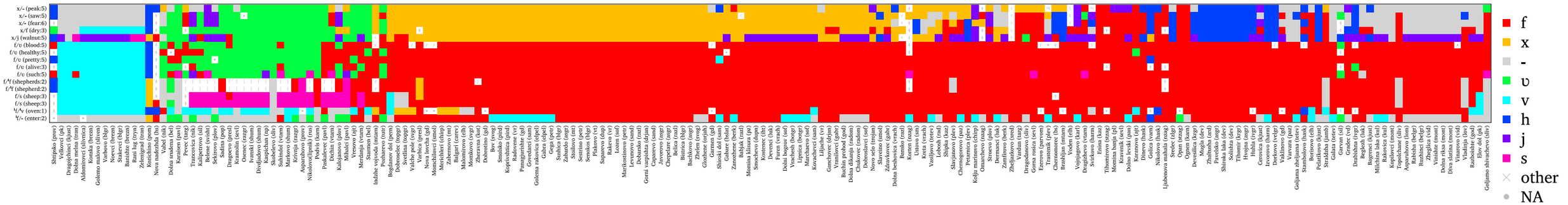

o
u
ʊ
-
ə
i
e
ɤ
other
NA

Proto-Bulgarian

	Labial	Dental	Palatal	Velar
Plosive	p b	t d		k g
Affricate		t͡s	t͡ʃ	
Fricative	f v	s z	ʃ ʒ	
Nasal	m	n	n^j	
Lateral		l		
Trill		r		
Approximant			j	

Old Church Slavonic

	Labial	Dental	Palatal	Velar
Plosive	p b	t d		k g
Affricate		t͡s d͡z	t͡ʃ	
Fricative		s z	ʃ ʒ	x
Nasal	m	n	nʲ	
Lateral		l	lʲ	
Trill		r	rʲ	
Approximant	v		j	



Reconstructed /f/ or /x/ or both ?

Conclusion

- Value the actual decisions that you make in the analysis, and share them
- Multialignments and reconstructions are central decisions, that can be profitably discussed and collectively improved