Using Parallel text
to quantify models of
linguistic diversity and change

Michael Cysouw
Philipps-University Marburg

Central Approach: Multialignment

- Alignments of words
- Alignments of sounds

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

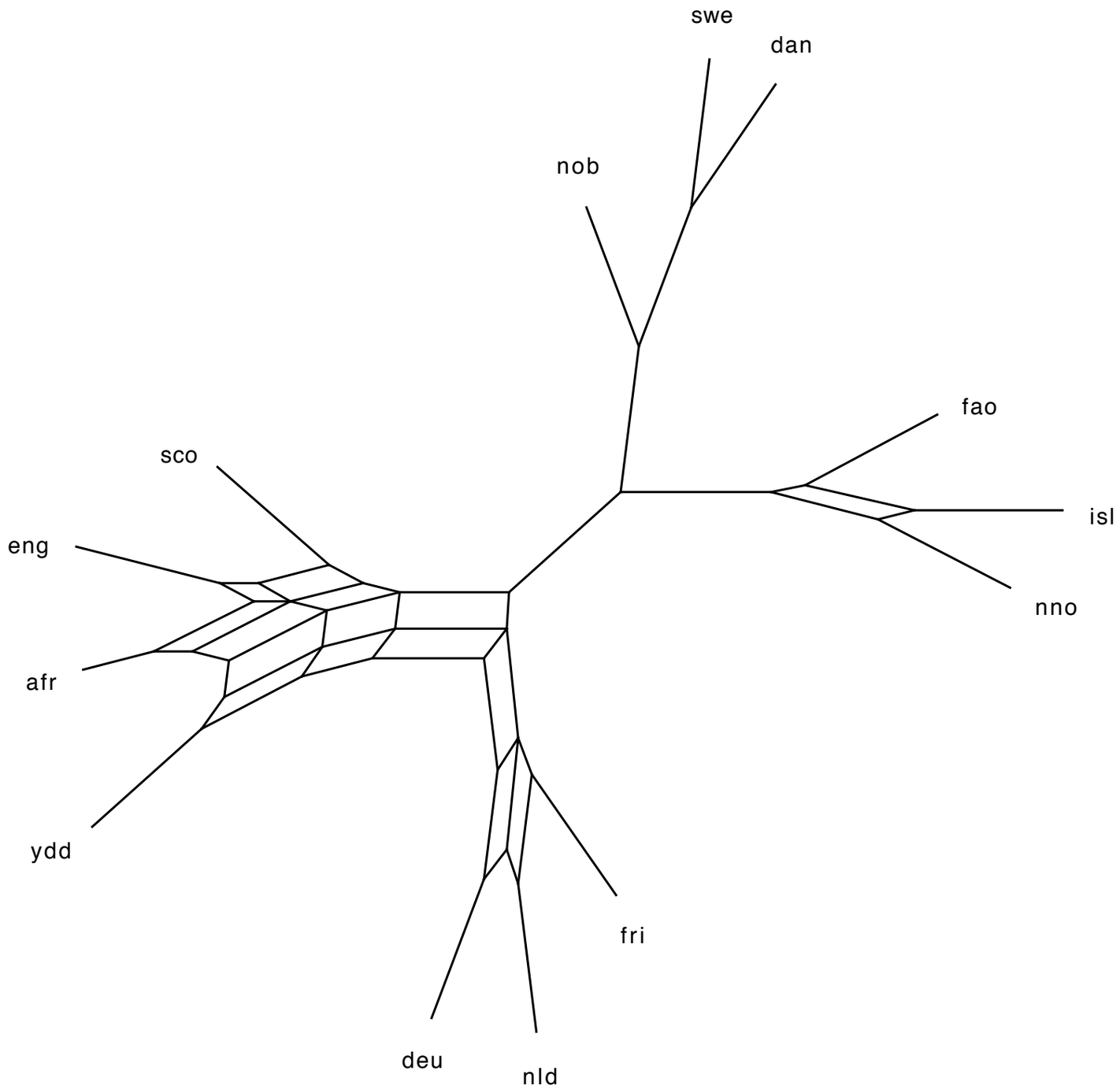
Article 10

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 ."

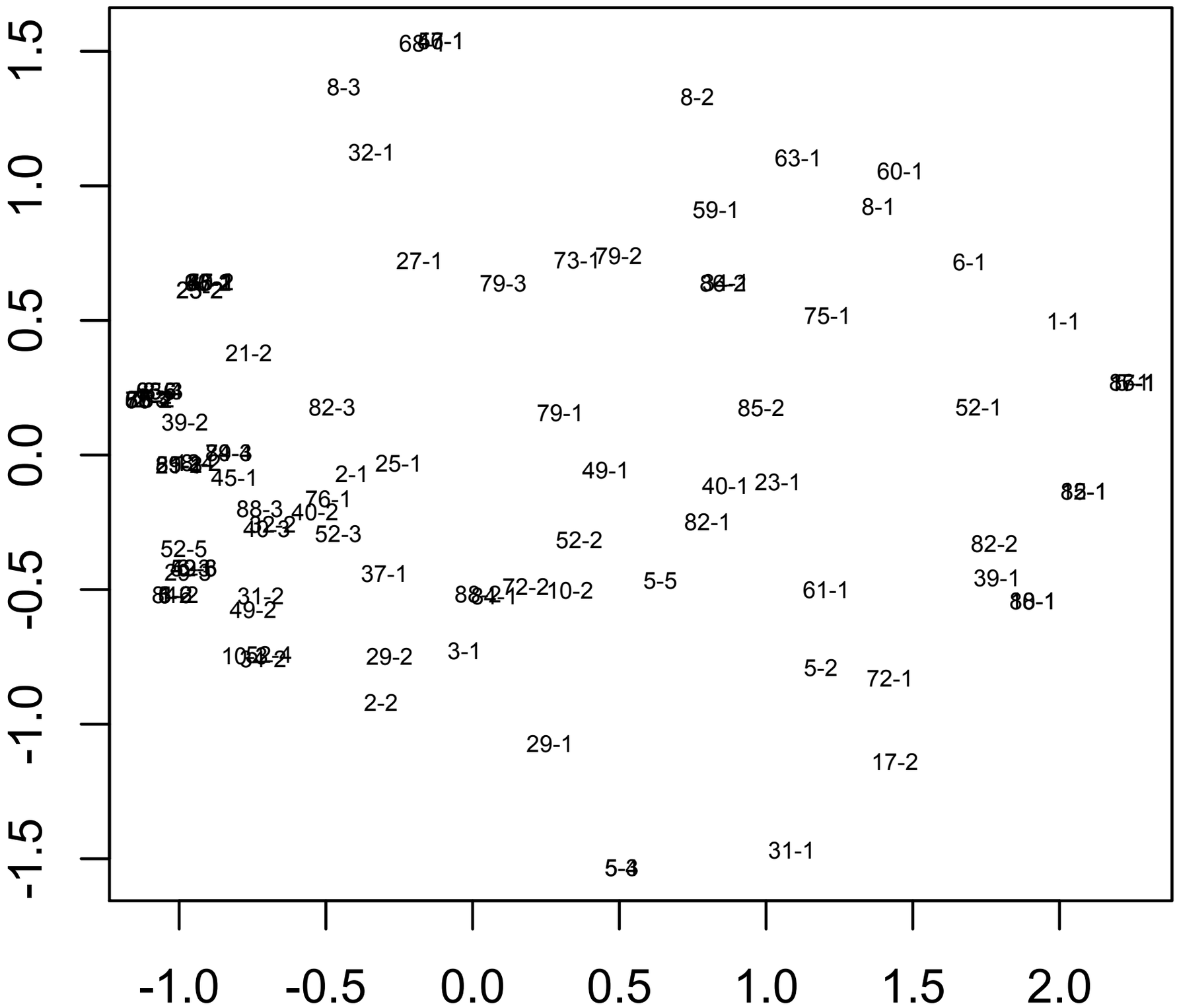
Afrikaans: "elkeen het , in volle gelykheid , die reg tot 'n regverdige en openbare verhoor deur 'n onafhanklike en objektiewe tribunaal , **in die bepaling van sy regte** en verpligtinge en die ondersoek van enige kriminele saak teen hom ."

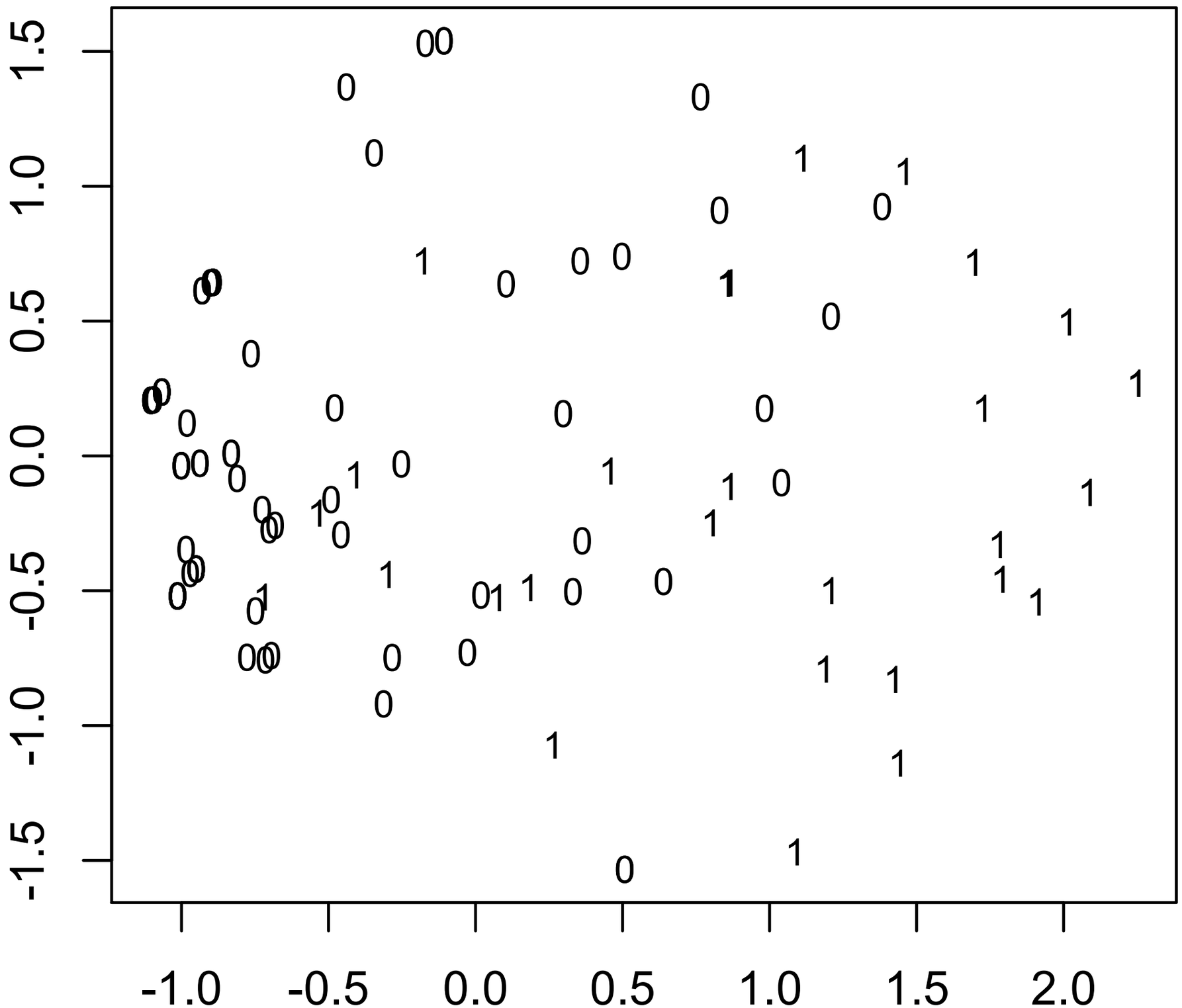
German: "jeder hat **bei der feststellung seiner rechte** und pflichten sowie bei einer gegen ihn erhobenen strafrechtlichen beschuldigung in voller gleichheit anspruch auf ein gerechtes und öffentliches verfahren vor einem unabhängigen und unparteiischen gericht ."

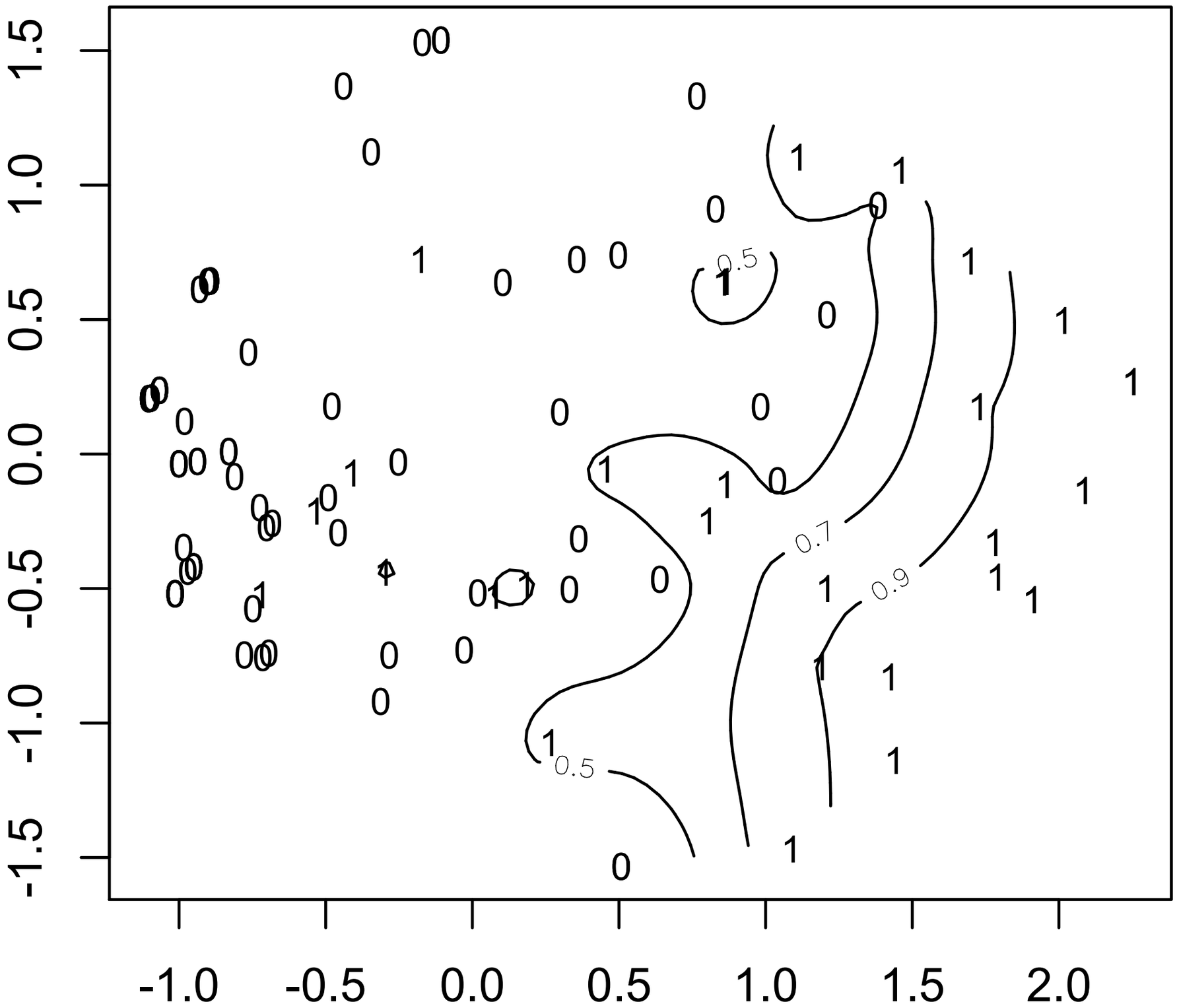
deu	2		1				1
eng	5		1	1	1	1	1
nld	5		1	1	1	1	1
fri	6	1	1	1	1	1	1
afr	5		1	1	1	1	1
ydd	4		1	1	1		1
sco	5		1	1	1	1	1
nno	3		1			1	1
nob	1		1				
swe	1		1				
fao	3		1			1	1
dan	1		1				
isl	1						1

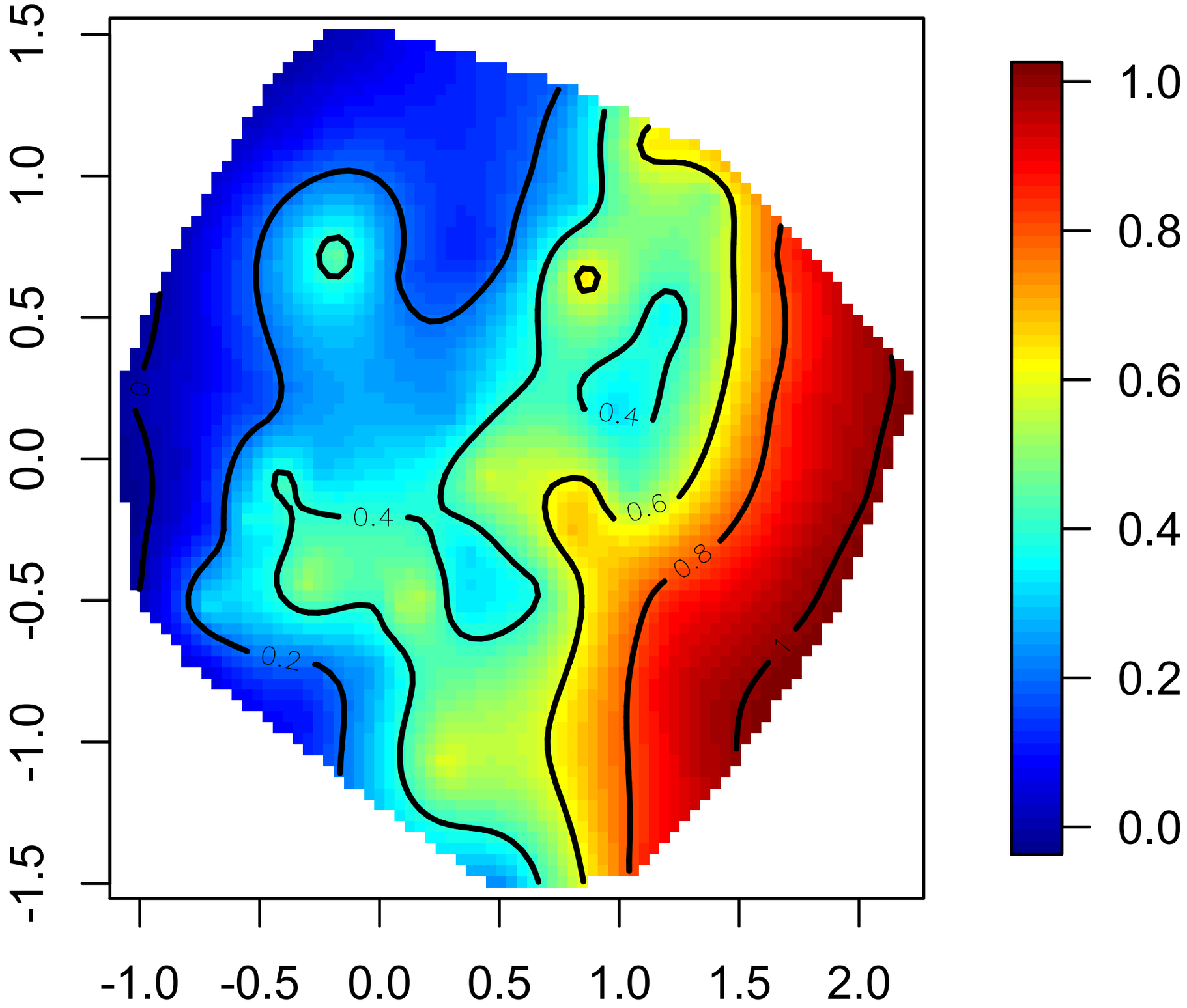


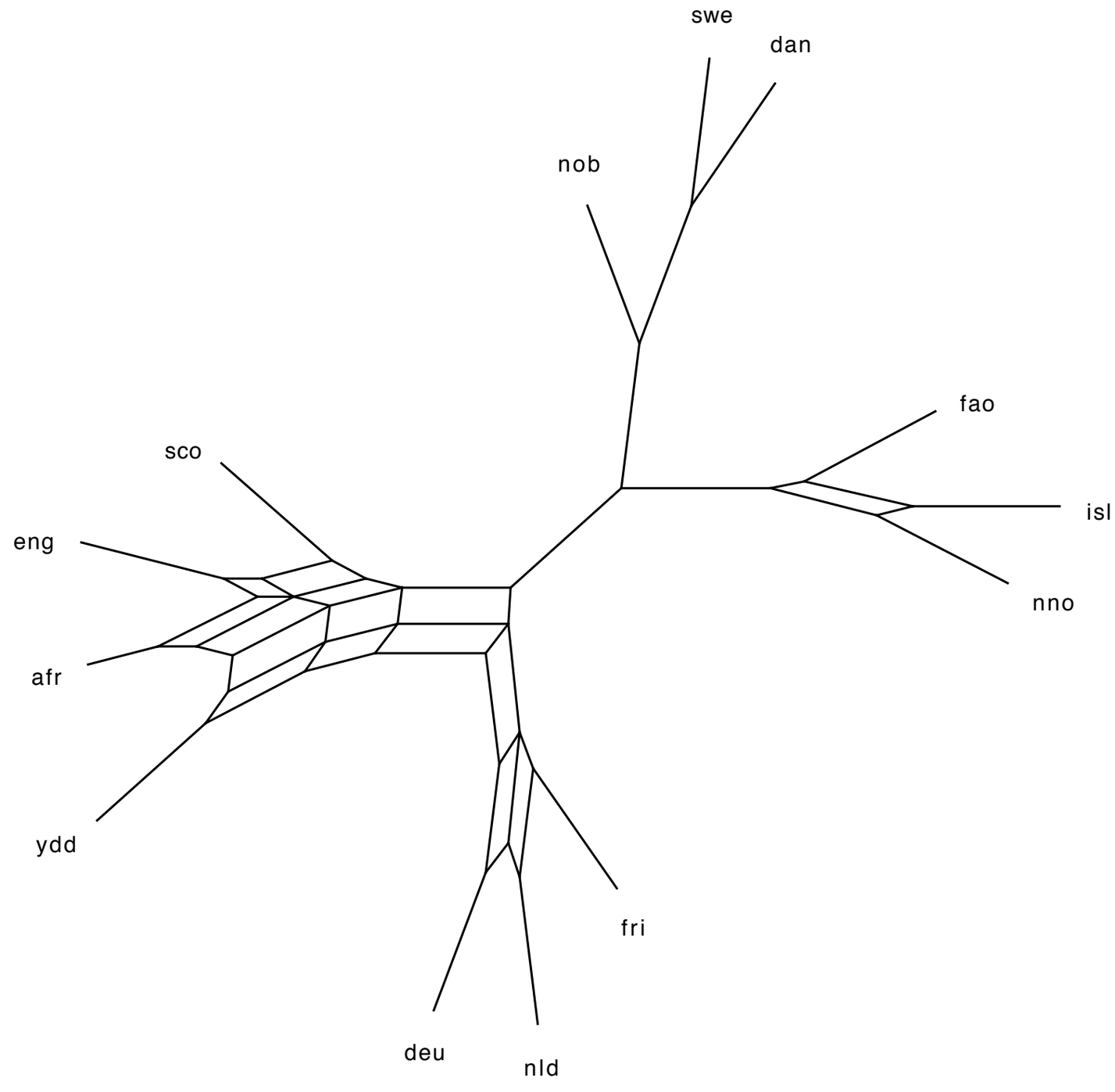
Consensus network of 4 optimal trees according to dollo Maximum Parsimony

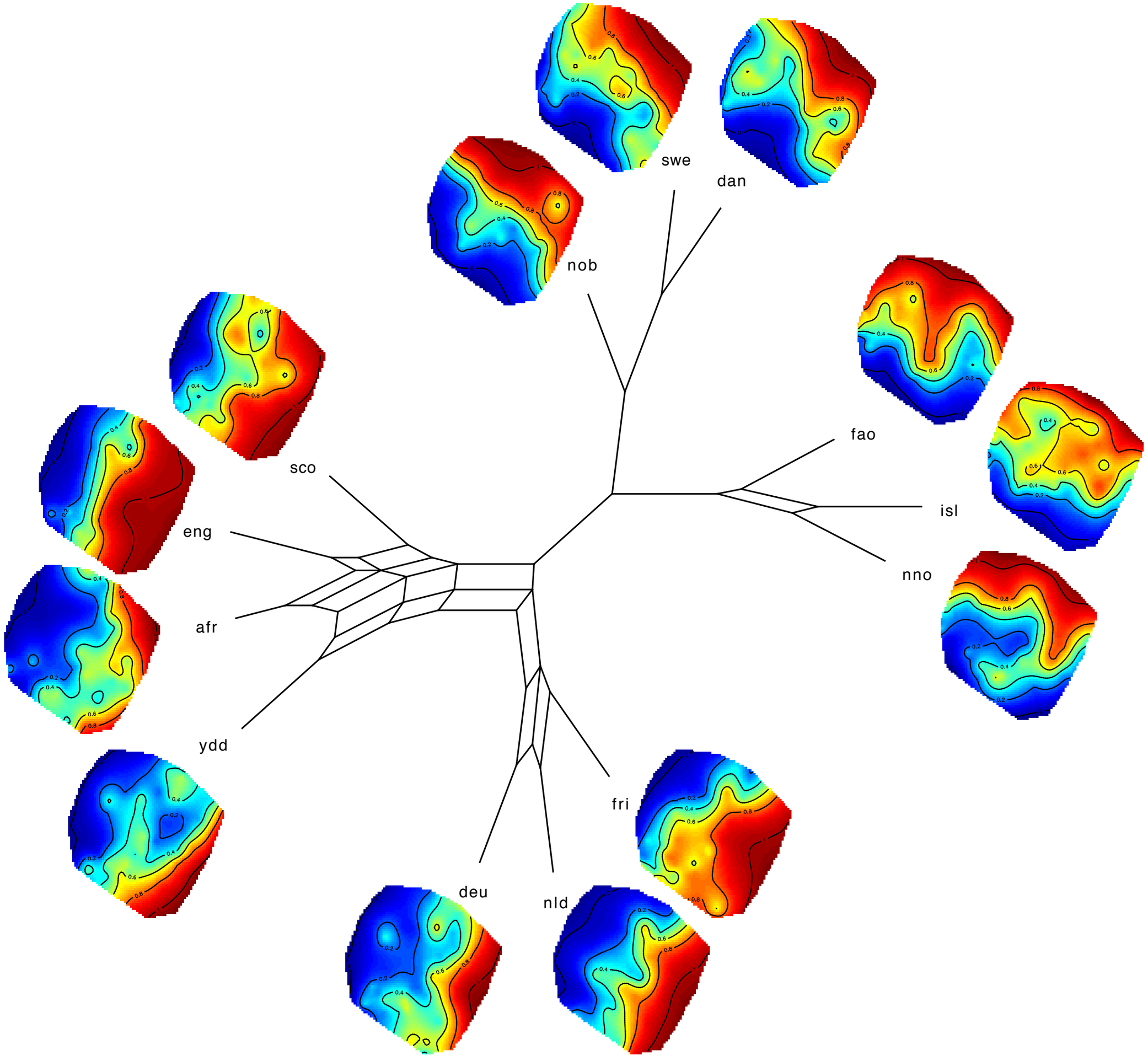

1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
8-3
8-2
32-1
63-1
60-1
59-1
8-1
27-1
73-1
79-2
79-3
6-1
75-1
1-1
21-2
82-3
79-1
85-2
52-1
39-2
25-1
2-1
45-1
49-1
40-1
23-1
76-1
88-3
40-2
82-1
52-3
52-2
82-2
52-5
37-1
5-5
39-1
72-2
10-2
61-1
31-2
49-2
3-1
29-2
5-2
72-1
2-2
29-1
17-2
31-1


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

Introducing the *Parallel Text*

Parallel Bible Corpus

- 1169 translations online (soon 1600+)
- 906 different ISO-639/3 codes (soon 1300+)
- In total more than 350 Million tokens
- More than 17 Million different wordforms
- http://paralleltext.info/data

Software

- Contact me personally for access
- R-package “qlcMatrix”

  http://cran.r-project.org/web/packages/qlcMatrix/index.html
  https://github.com/cysouw/qlcMatrix

- Python library

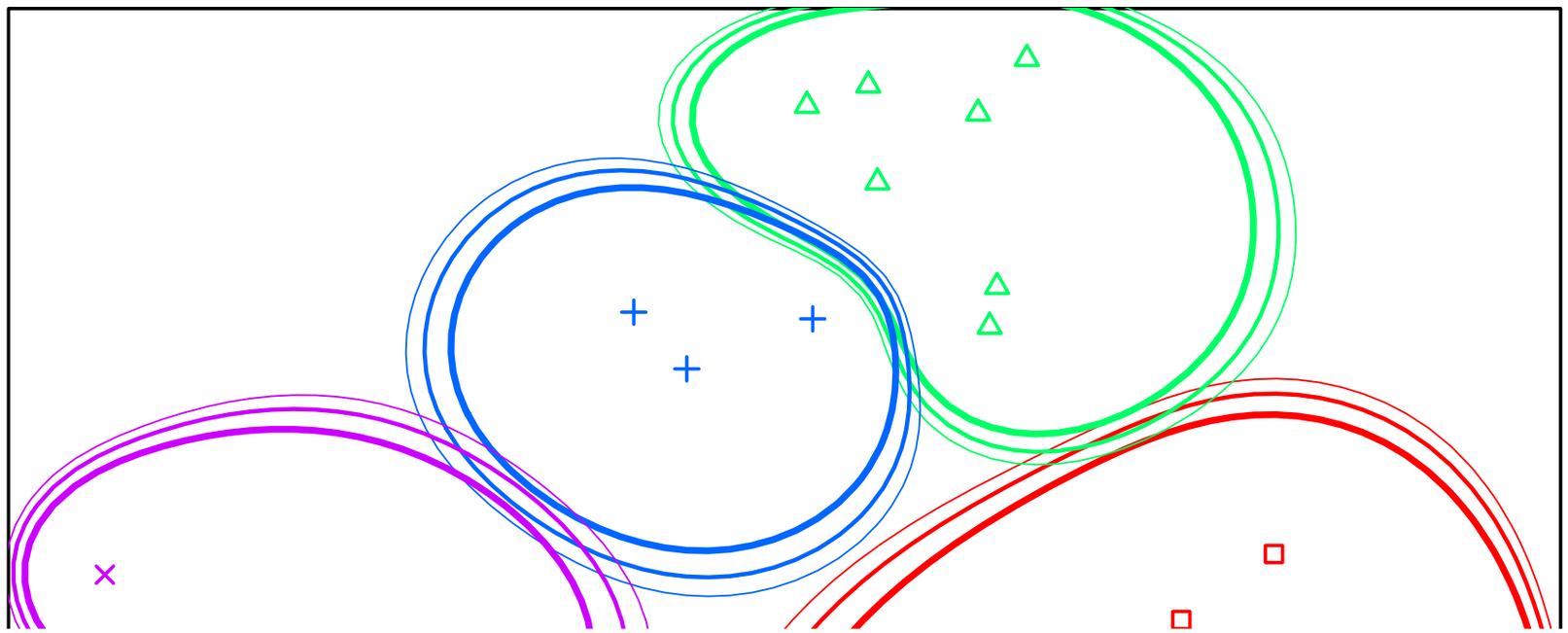
  https://github.com/tmayer/paralleltextprocessing

Multiple Alignment

- Small-scale experiment
  - use *fastalign* for bitext-alignment on all pairs
  - build multi-text-alignment using graph clustering
- Only for 77 Germanic translations
- New Testament produced almost 100.000 Germanic alignments, which are directly comparable ‘words’

trees and wood

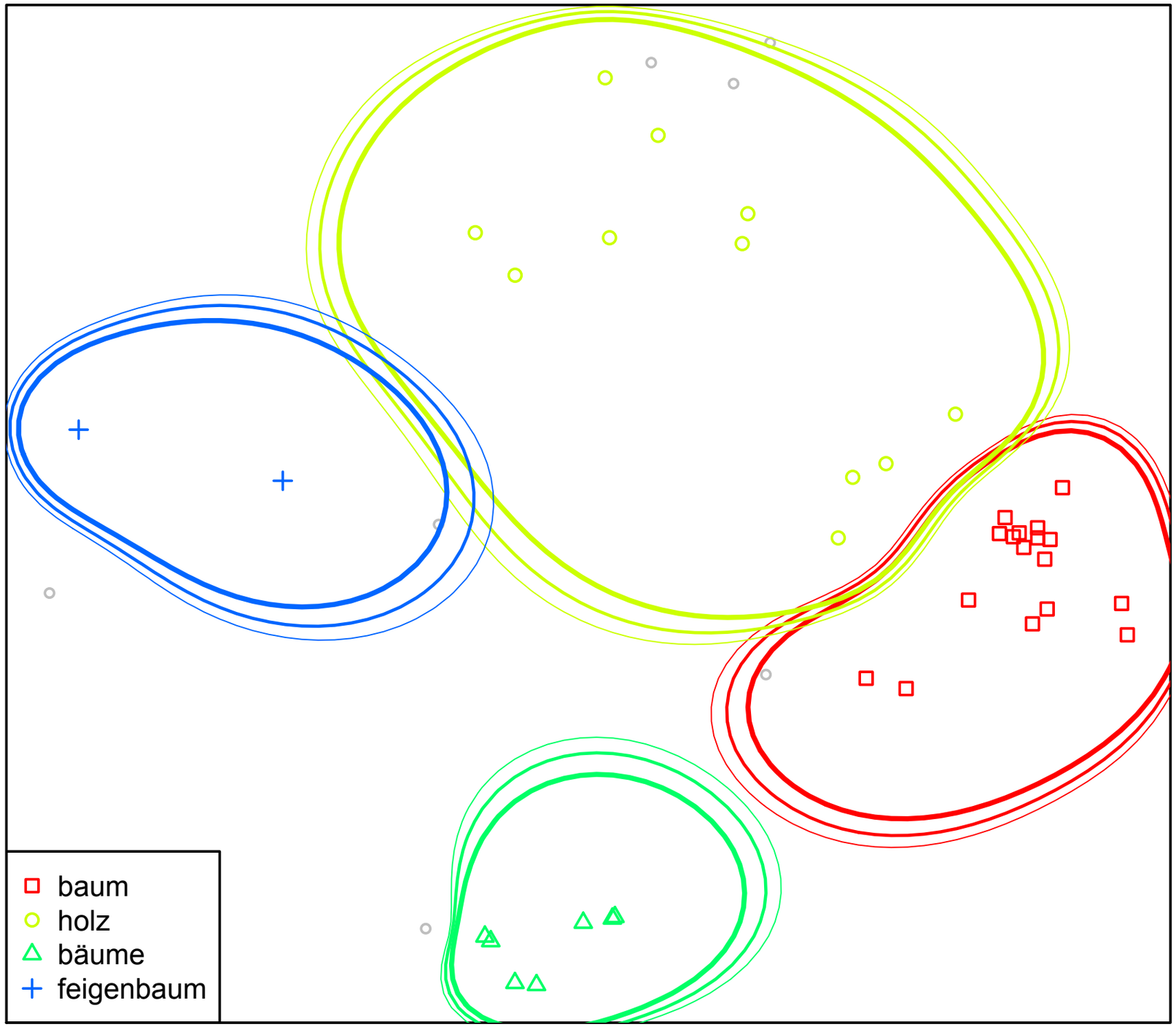
afr-x-bible-1953.txt



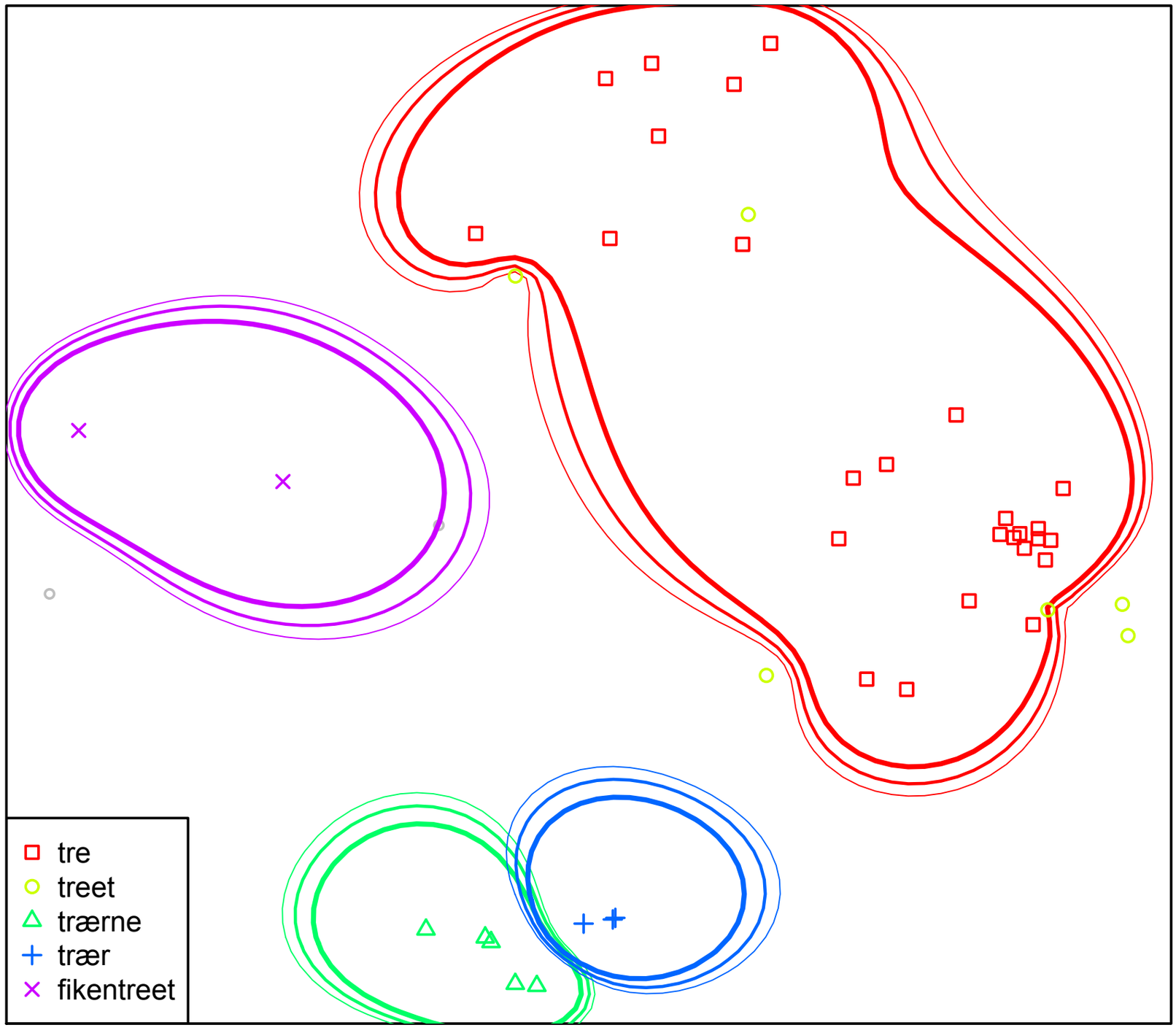
	tree	wood (stuff)	firewood	small forest	large forest
German	*Baum*	*Holz*		*Wald*	
Danish	*træ*			*skov*	
French	*arbre*	*bois*		*forêt*	
Spanish	*árbol*	*madera*	*leña*	*bosque*	*selva*

Louis Hjelmslev
Prolegomena to a Theory of Language (1963)

deu-x-bible-erben.txt

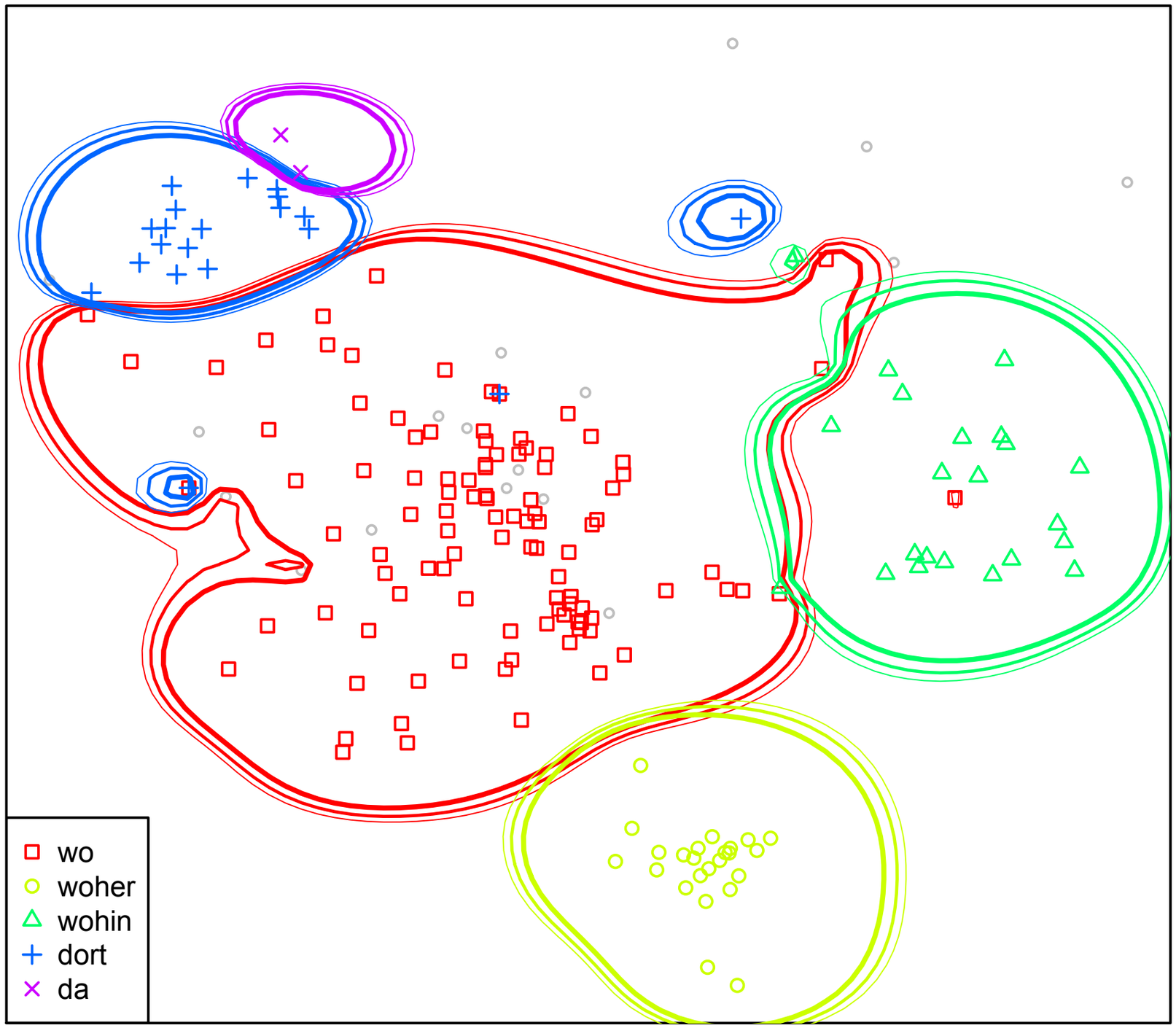


nob-x-bible-2007.txt

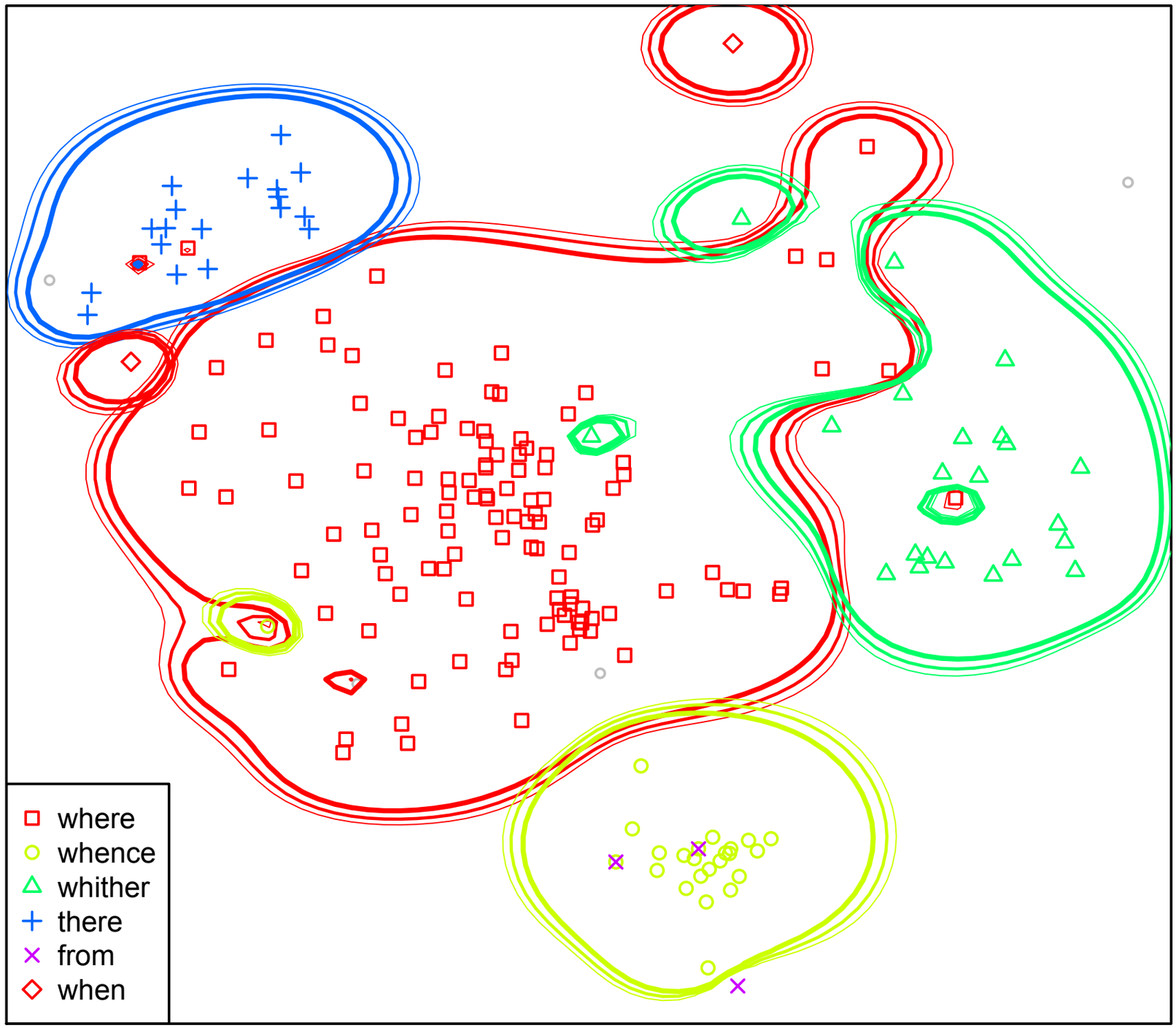


where

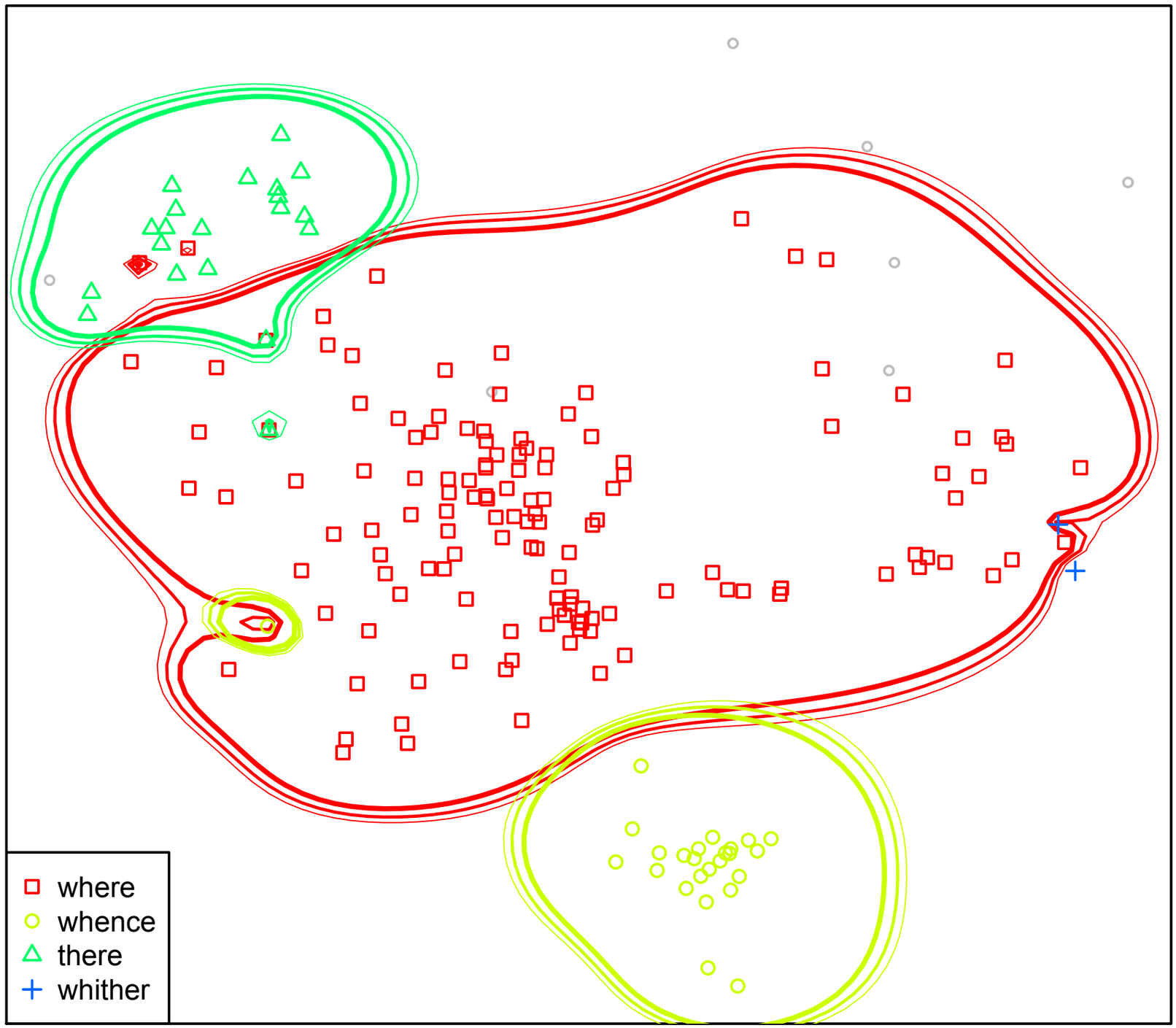
deu-x-bible-pattloch.txt



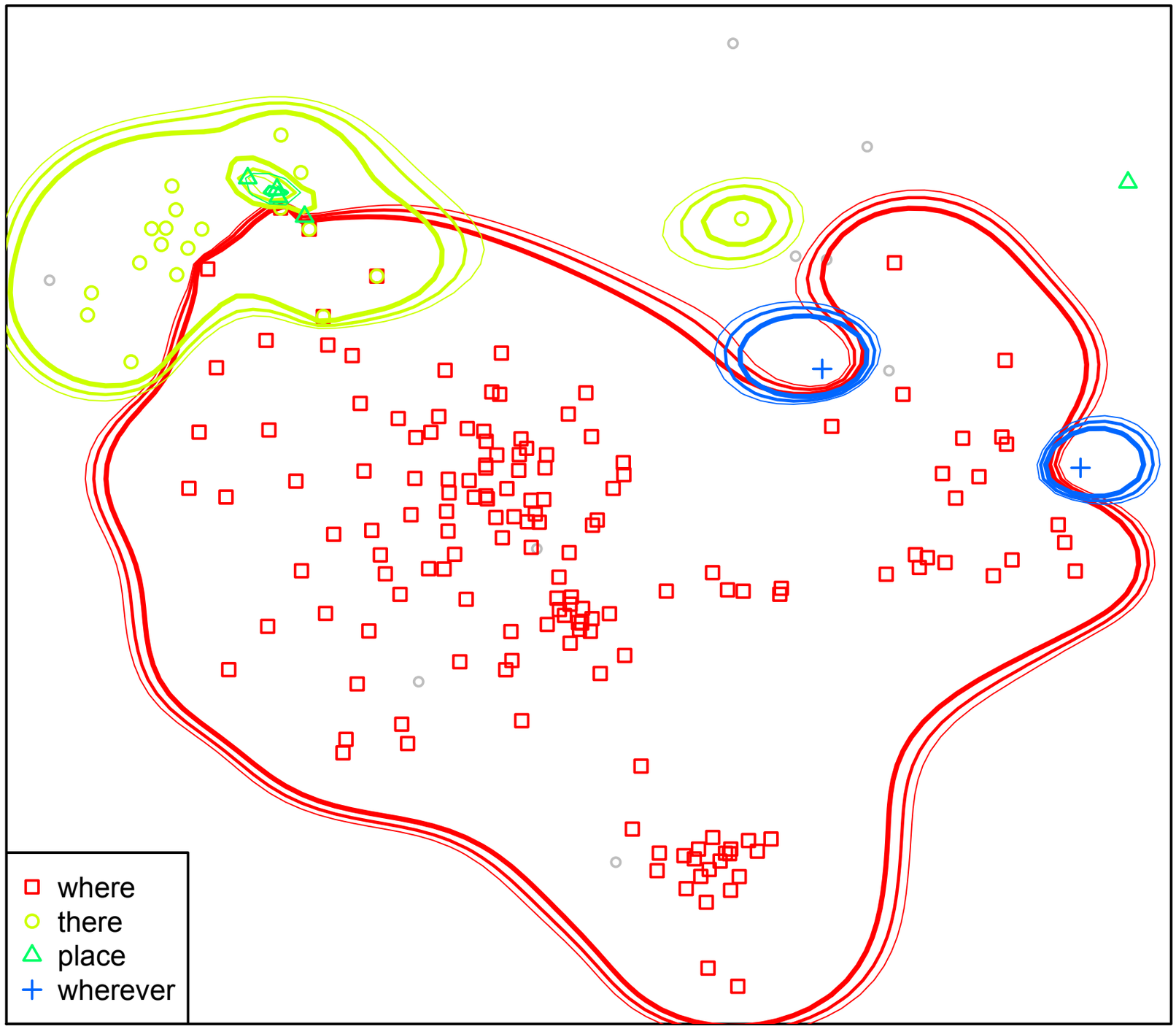
eng-x-bible-kingjames.txt



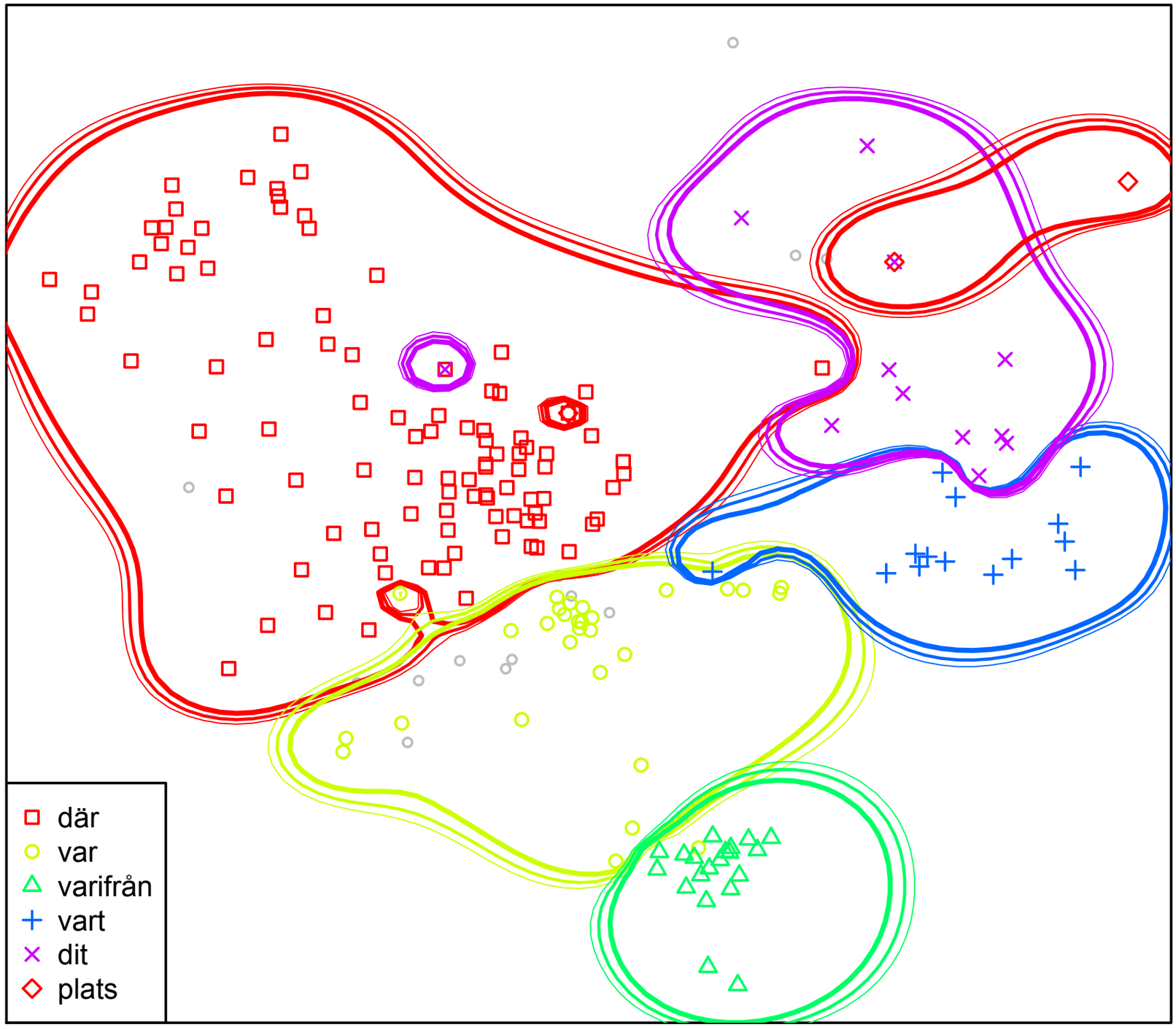
eng-x-bible-darby.txt



eng-x-bible-treeoflife.txt

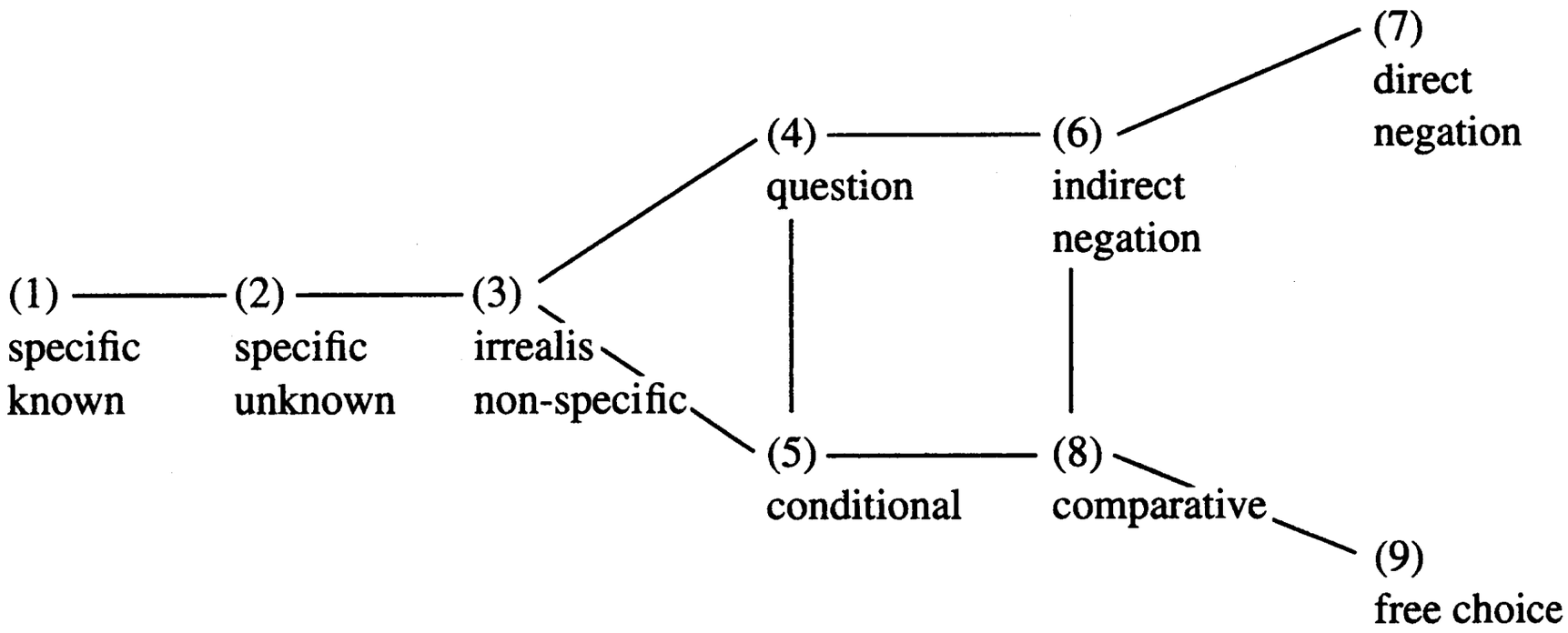


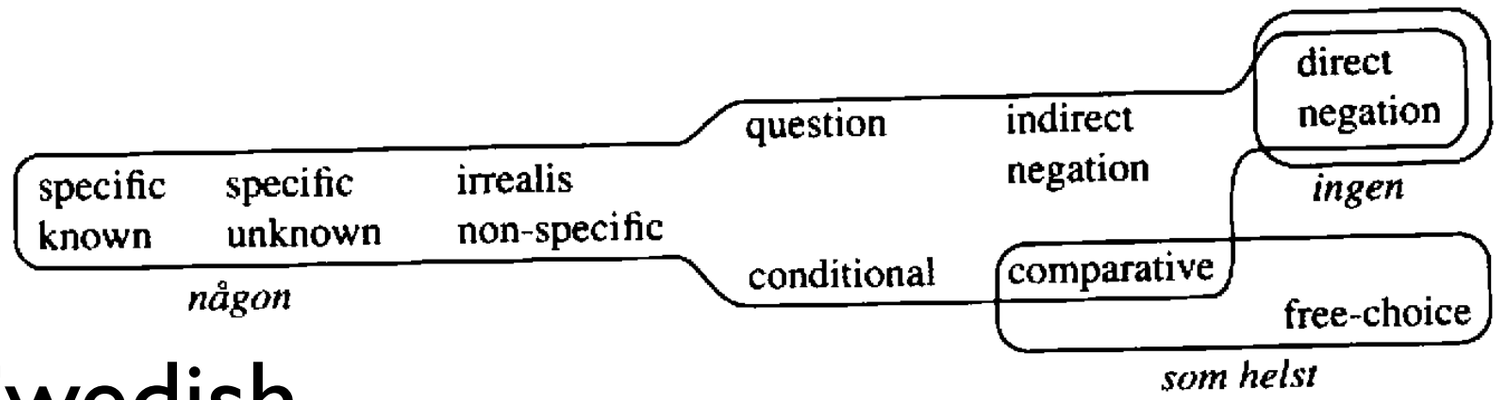
swe-x-bible-folk1998.txt



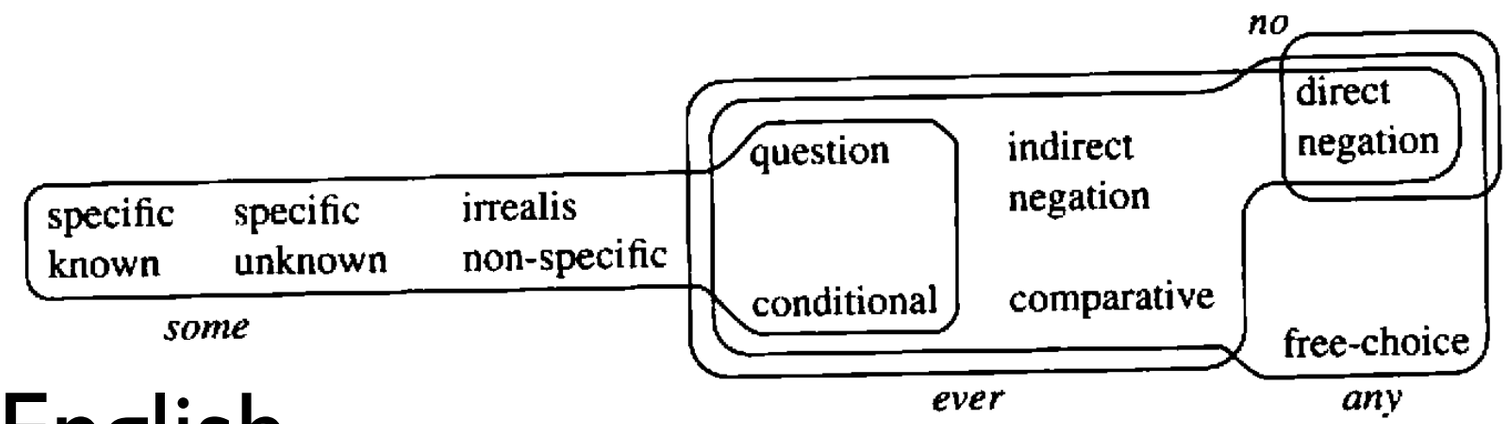
Indefinite person

(someone, anyone)





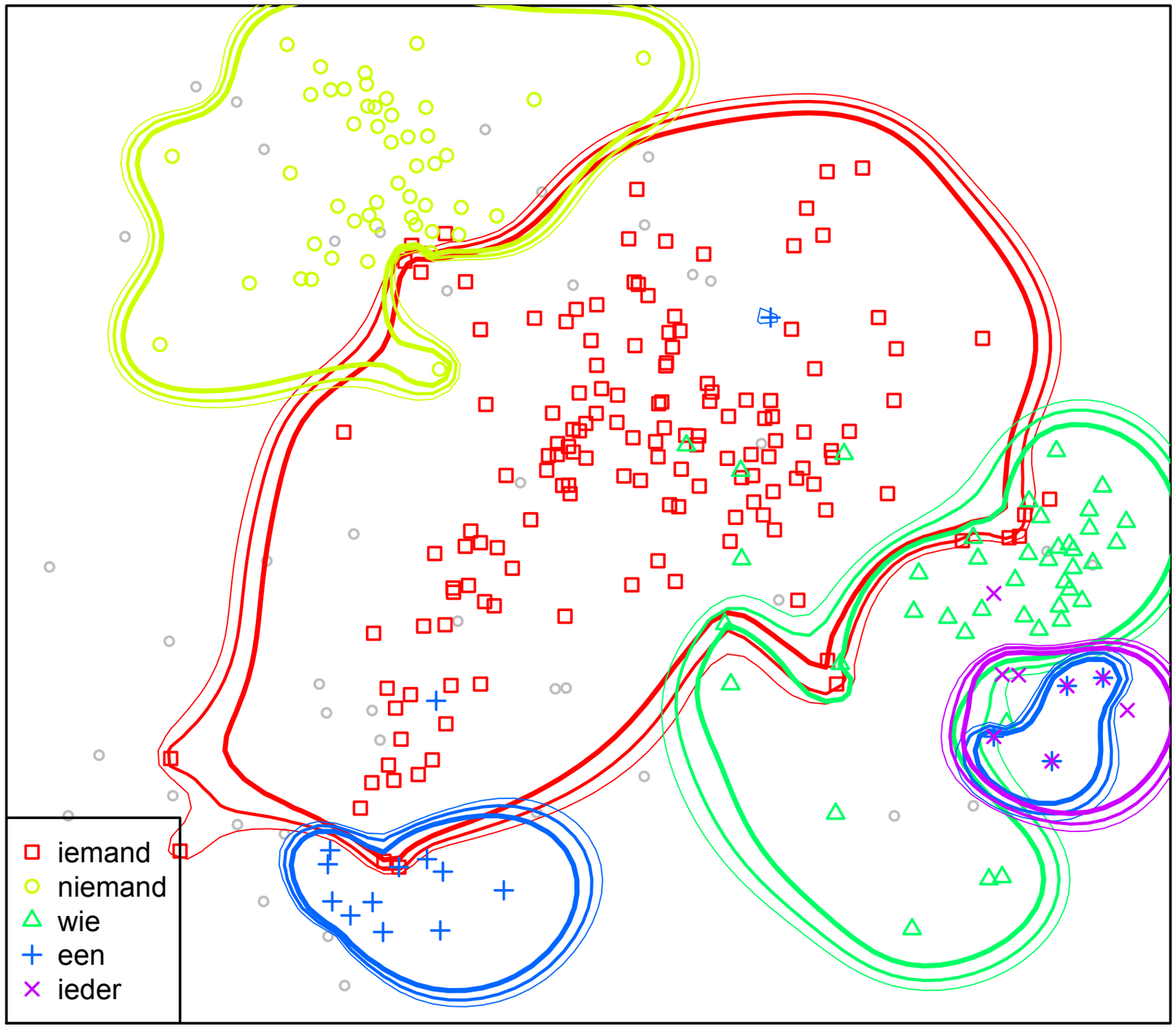
Swedish



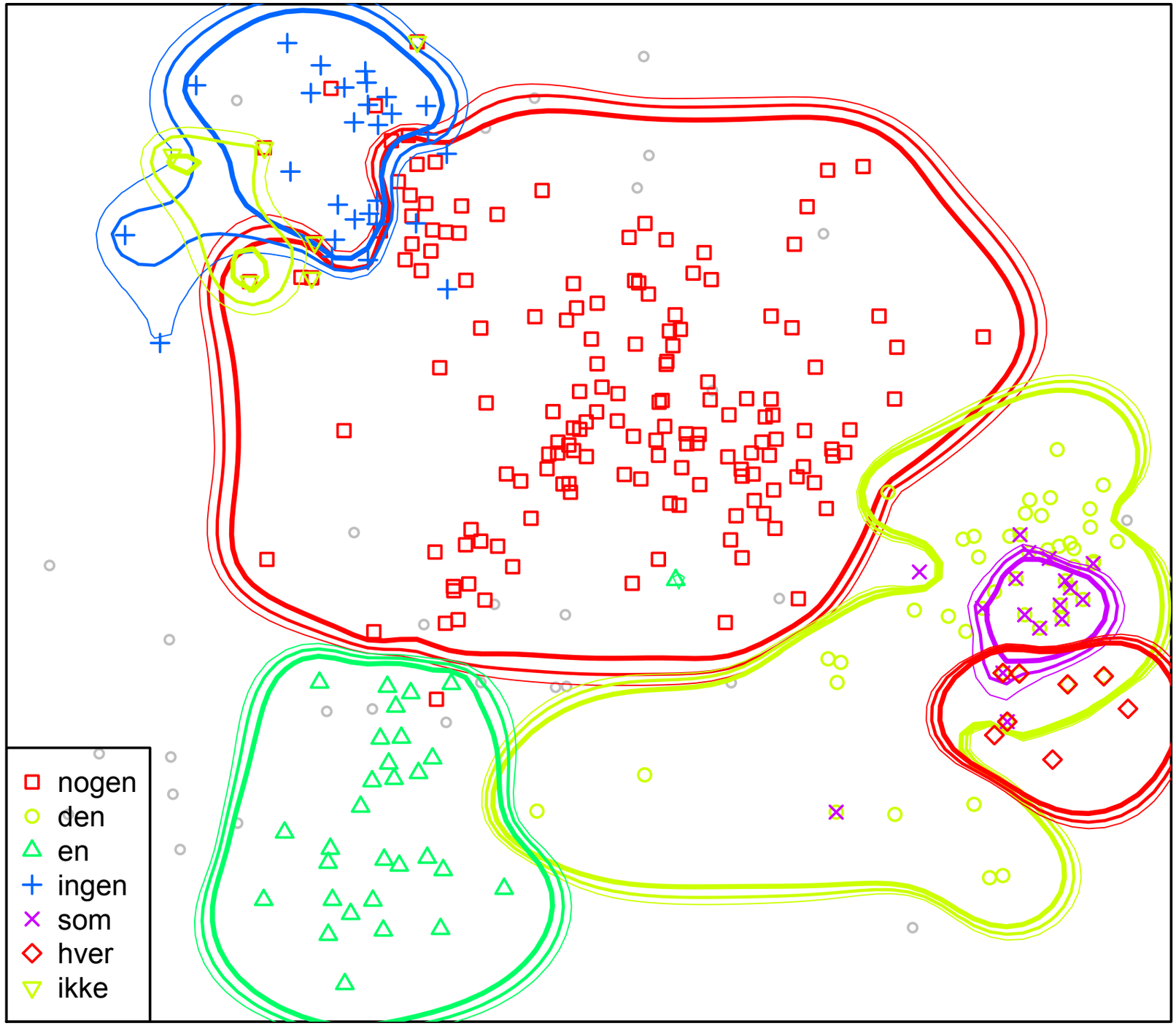
English

Haspelmath, Martin. *Indefinite Pronouns*. Oxford Studies in Typology and Linguistic Theory. Oxford: Clarendon, 1997.

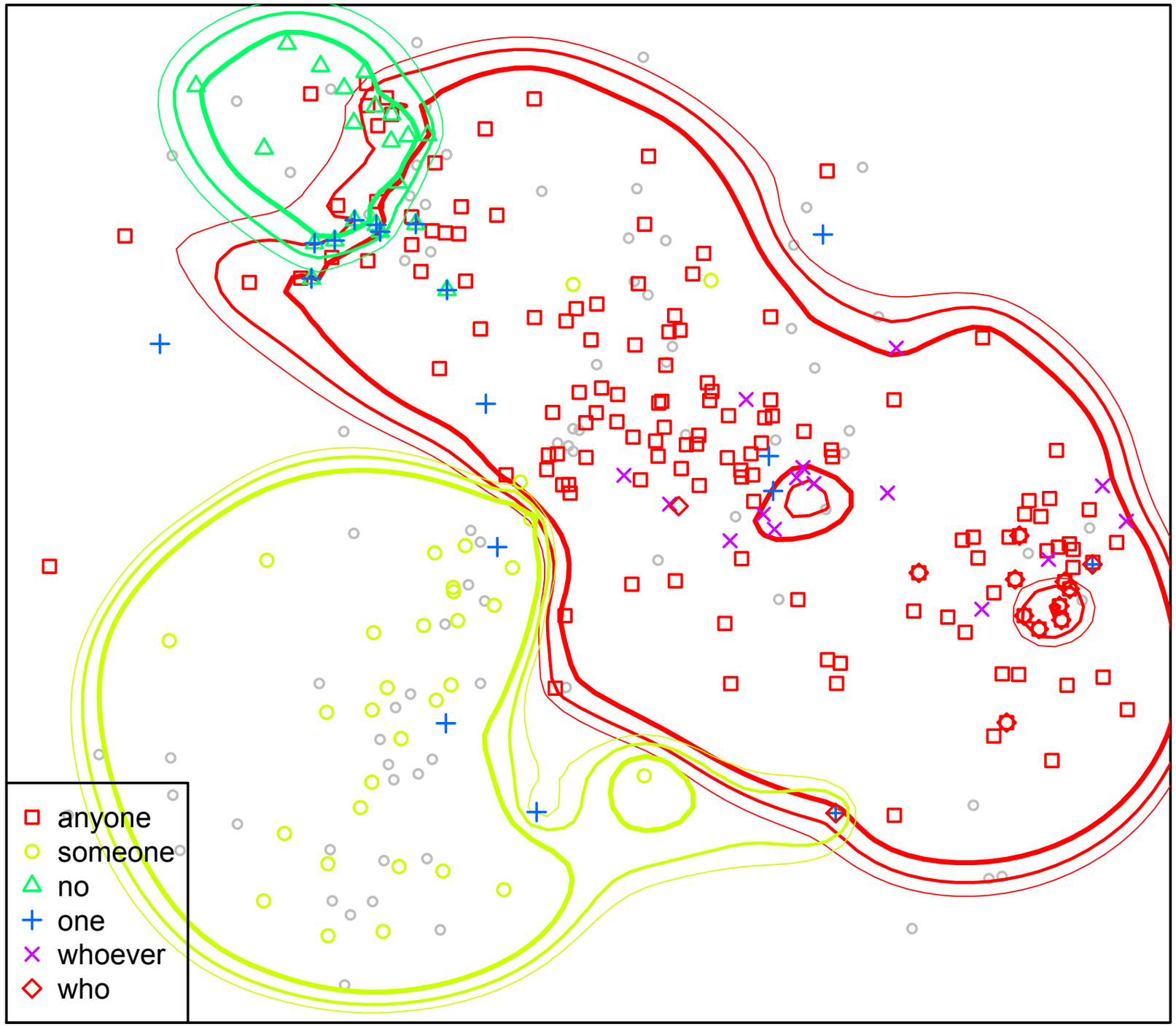
nld-x-bible-1951.txt



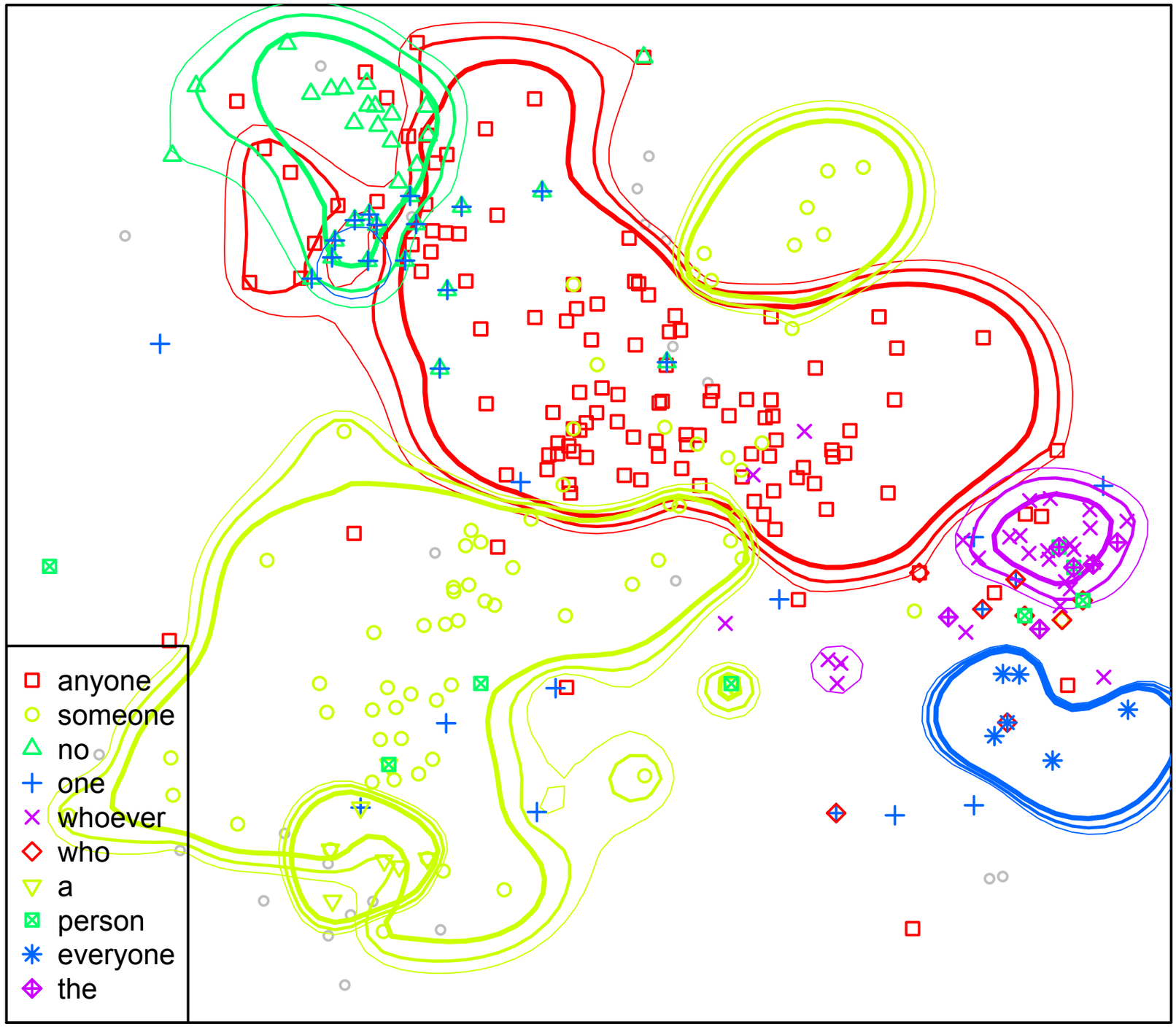
dan-x-bible-1931.txt



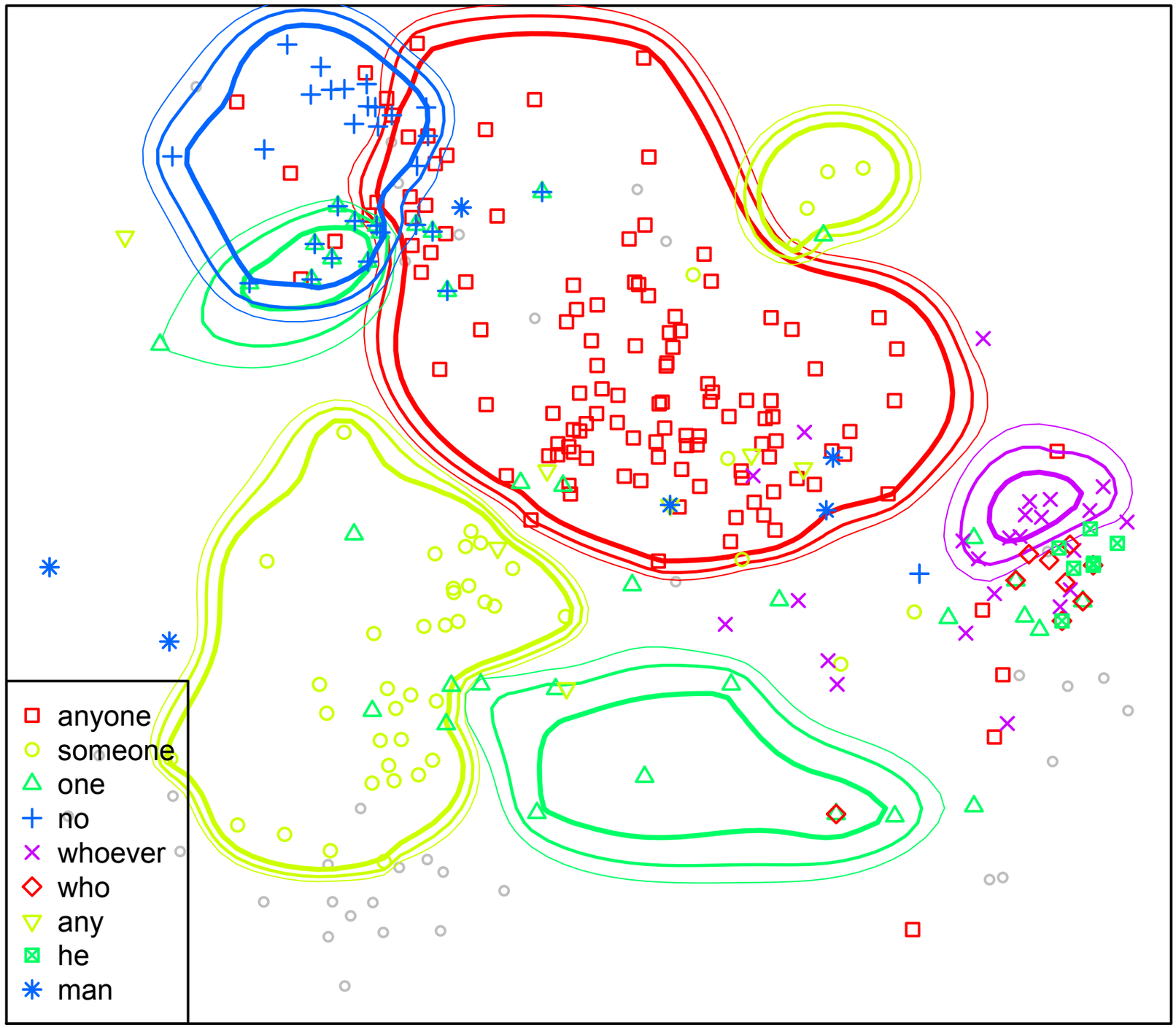
eng-x-bible-books.txt



eng-x-bible-new2007.txt



eng-x-bible-treeoflife.txt



Multialignment of Sounds

Lfd. Nr.	Stichwort	Transkription	Bemerkungen
1	178 Winter	ˈvɪntʰəʁ	ʁ = kontinuant ə bereits velarisiert
* 2	56 fliegen	ˈflaɪ->ətʰə	'fliegen die', Segmentierung unklar - kein geminiertes [t]
3	23 Blätter	ˈblɛːdəʁ	
4	103 Luft	lʊ°ftʰ	ʁ = kontinuant
5	89 hört	hɪʁt	ʁ = kontinuant
6	78 gleich	klɛ->ɪk	folgt ?
7	130 schneien	ˈʃnɛ->ɪən	
8	174 Wetter	/	statt dessen ˈvɪtəʀʊŋə
9	151 tu	dɛ->ʊ	

LOCATION	WORD
Aachen	a̟ːph
Adorf	ɑ̟ːbʱə
Ahrbergen	o̟→ɔ̟phə̟
Albersloh	ɑːpᵝ̈ə̟
Allna	ɑɸh
Altenberg	ʌfɛ̟
Altentrüdin	a̟f
Altlandsberg	ɑʾfəʾ
Altwarp	oːph
Astfeld	ɒˤːpʰə
Atzendorf	afɛ̟
Ballhausen	ʌʾfə
Bardenfleth	ɔːp͡ɸ
Barssel	ɒːpʱə
Bempflingen	afː
Bennin	ɔ̟pʰ
Billingsbach	a̟f
Bockelwitz	ʌ̟və̟
Bonn	ɑːpʾ
Borstendorf	ɤ̟fː
Breddin	ɒːph
Brelingen	ɑ̟f͡βə
Bremscheid	ɒˤːpḫə̟
…	…

A	FF	E
a̟ː	ph	-
ɑ̟ː	bʱ	ə
o̟→ɔ̟	ph	ə̟
ɑː	pᵝ̈	ə̟
ɑ	ɸh	-
ʌ	f	ɛ̟
a̟	f	-
ɑʾ	f	əʾ
oː	ph	-
ɒˤː	pʰ	ə
a	f	ɛ̟
ʌʾ	f	ə
ɔ̟ː	p͡ɸ	-
ɒː	pʱ	ə
a	fː	-
ɔ̟	pʰ	-
a̟	f	-
ʌ̟	v	ə̟
ɑː	pʾ	-
ɤ̟	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

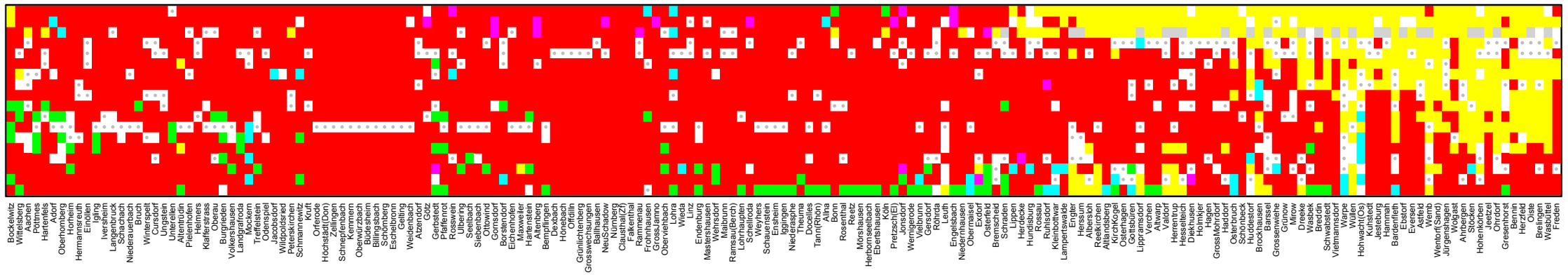
COLUMNID	1	2	3	4	5	6	7	8	9
STANDARD	Au	g	e	n	b	l	i	ck	(e)
Adorf	ɐ→ʊ	ʁ	-	ŋ̣	b	l	e̯	kʰ	-
Ahrbergen	ə̟→o͡o̟	ɣ	ə	m	b	l	ɪ̈	k’	-
Albersloh	ɐ→ʊ	-	-	m	β̃	l	ɪ	k	-
Allna	ɔː	-	-	-	p	l	œˤ	x	-
Altenberg	ɤɣ̣	ʁ	ə̃	-	b	l	ɪ	k	-
Altentrüdin	ɐ→ʊ	ʁ	ə	-	p	l	ɪ̣	g	-
Altlandsberg	ɑ’→ʊ	ɣ	-	ŋ̣	b	l	ɪ̣	k͡x	-
Altwarp	o͡ʊ	-	-	ŋ̣	b	l	ɪ	k	-
Astfeld	ʊ̝ː	ɣ	ə	m	b	l	ɪ	kʰ	ə
Ballhausen	a̠→ʊ	ɣ	-	ŋ̣	p	l	ɪ̈	k	-
Bardenfleth	oː	g	-	ŋ̣	b	l	ɛ	k͡x̟	-
Barssel	oː	g̃	-	ŋ̣	p	l	ɪ̝	k̟	-
Bempflingen	ɐ→ʊ	g	ə	-	b	l	ɪ̣	cʰ	-
Bennin	oː	-	-	ŋ̣	b	l	ɪ̝	x	-
Billingsbach	ɑ̟ː	x	ə	-	p	l	ɪ̣	k̃x	-
Bockelwitz	a̠→ɯ	ɣ	-	ŋ̣	b	l	ɪ	k	-
Borstendorf	ɔ̝ː	ɣ͡x	-	ŋ̣	p	l	e̯	k	-

github.com/digitallinguist/msa-editor

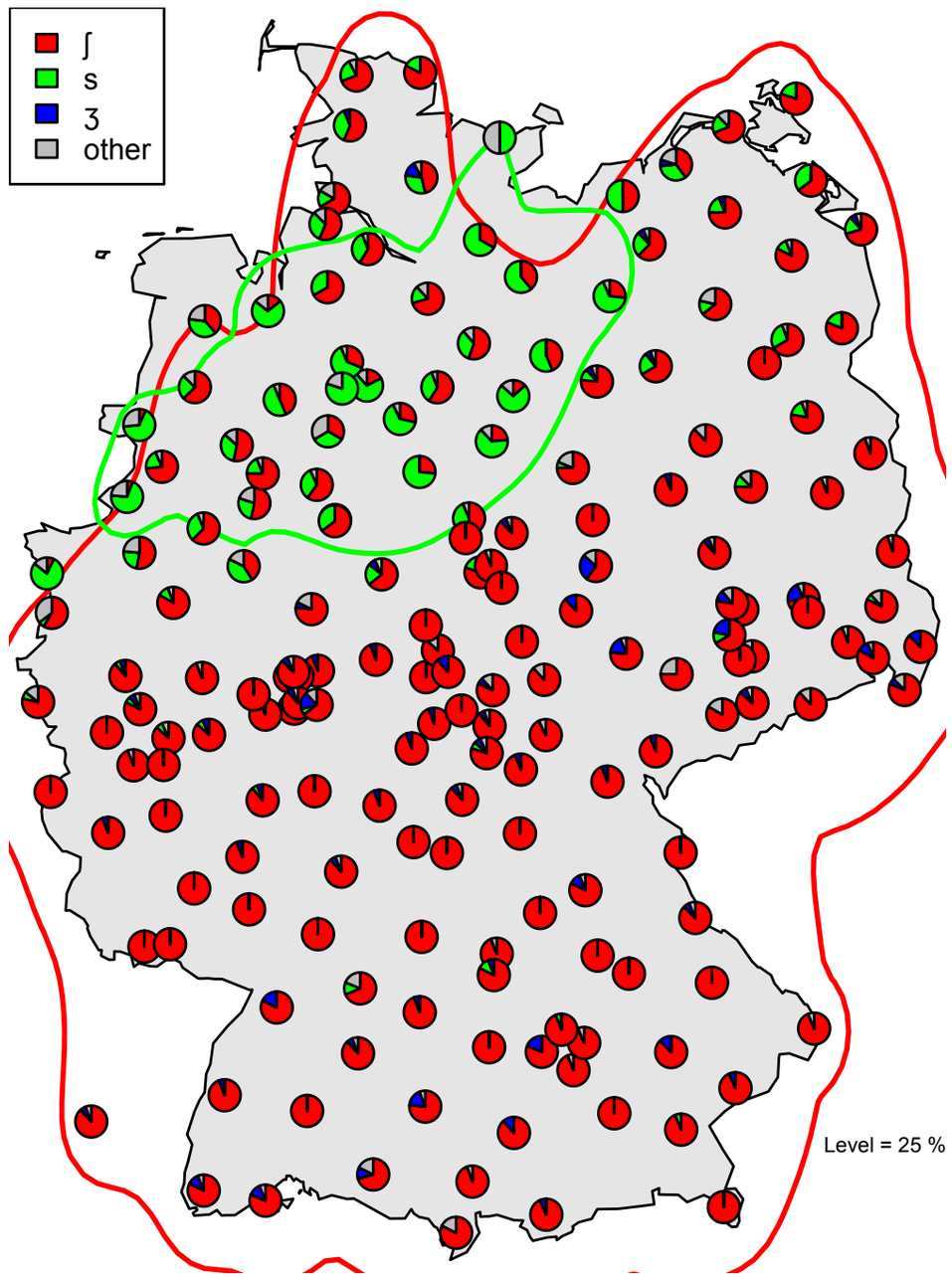
Correspondence Sets

- 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

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”



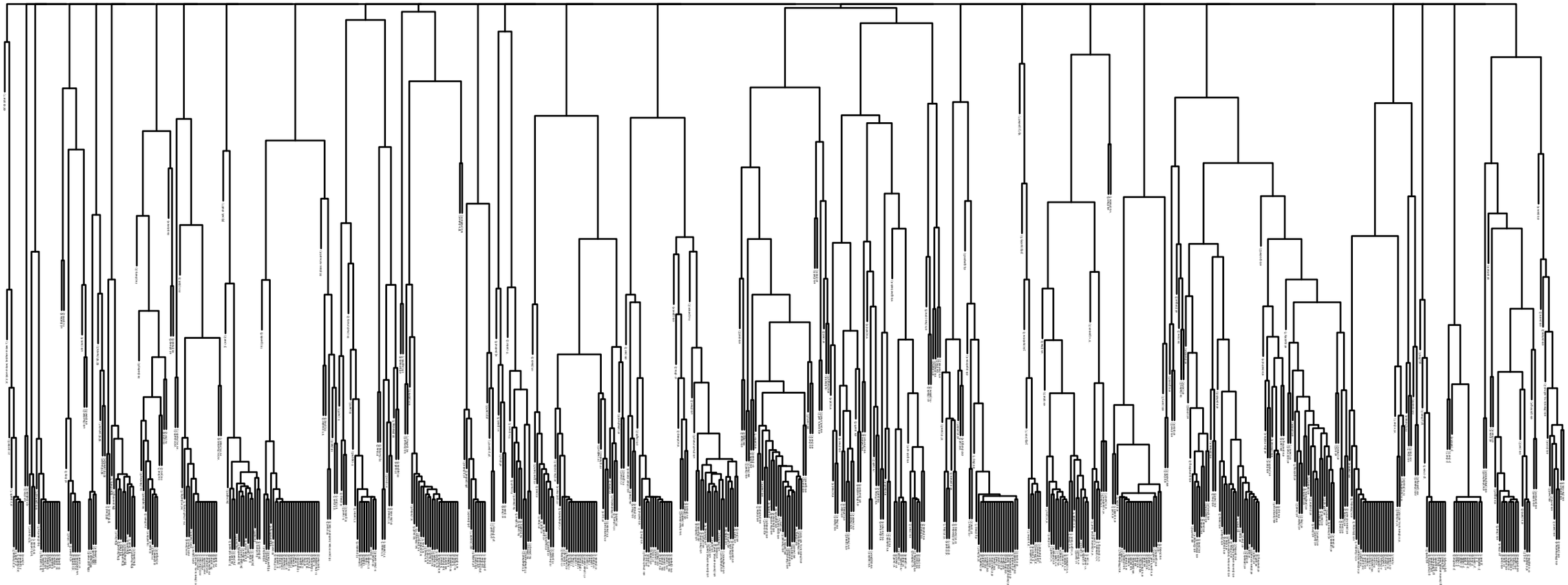
Bulgarian dialect data

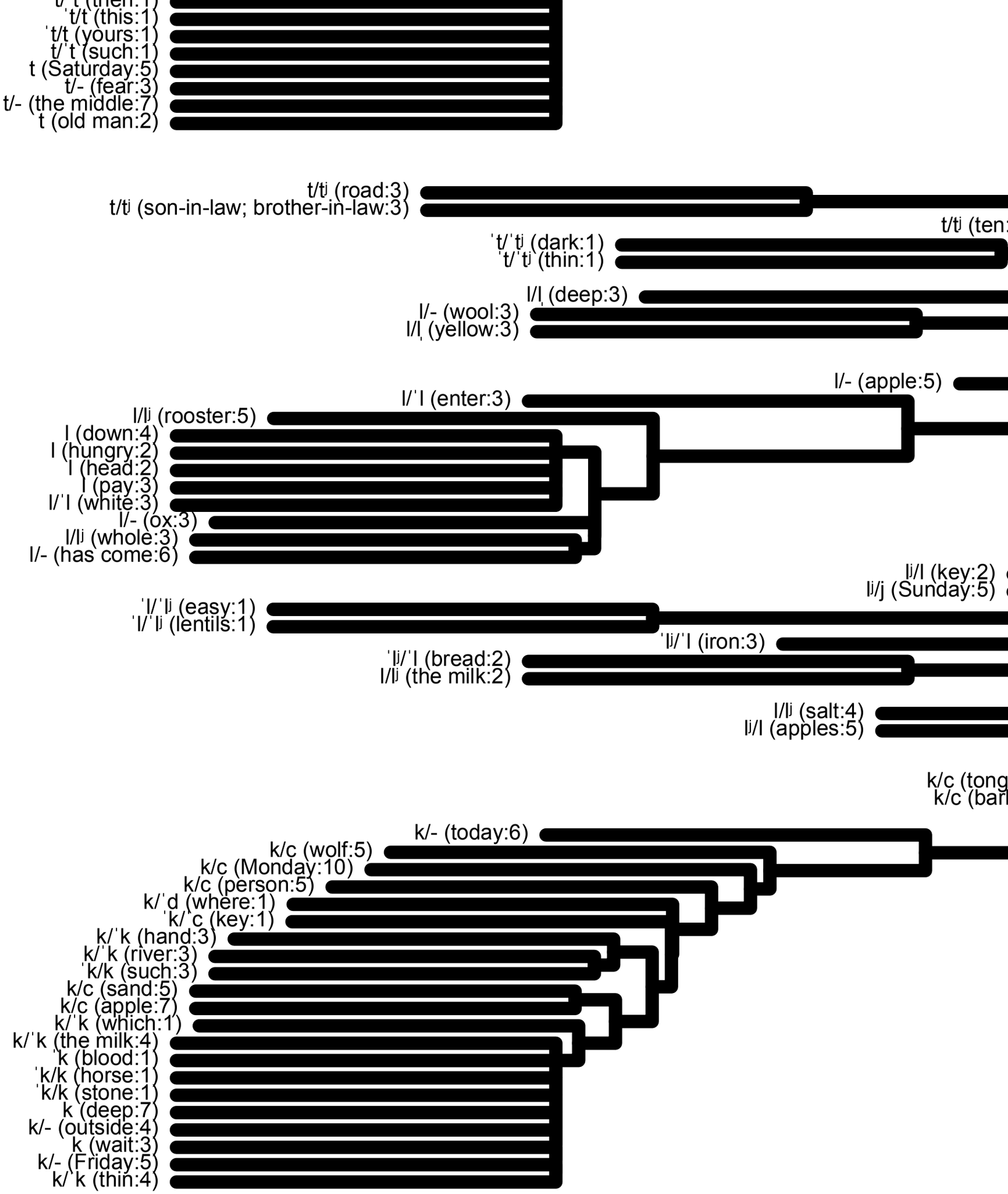
- Data from Jelena Prokić
- Taken from the ***benchmark database of phonetic alignments*** (BDPA)

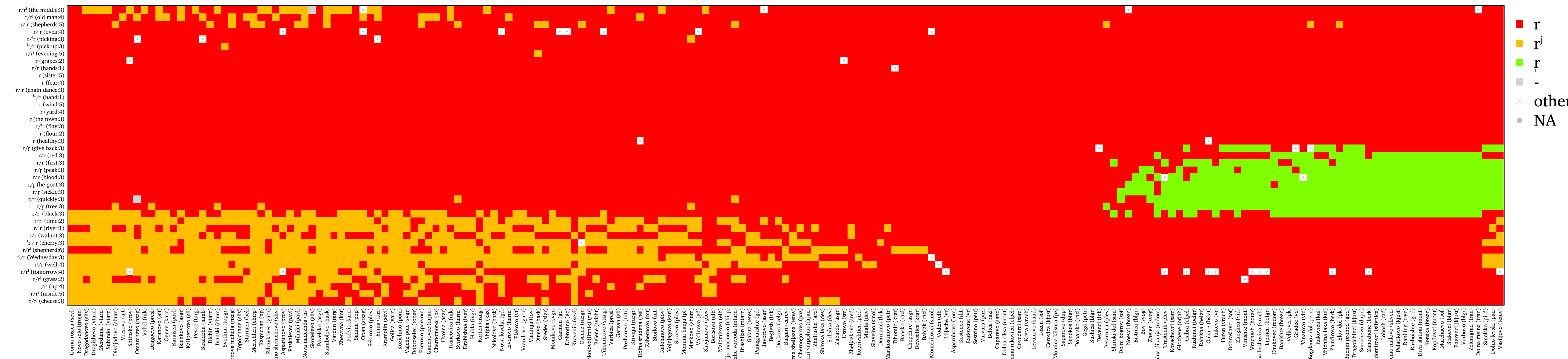
List, Johann-Mattis and Jelena Prokić. (2014). A benchmark database of phonetic alignments in historical linguistics and dialectology. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), May 2014, Reykjavik. 288-294.

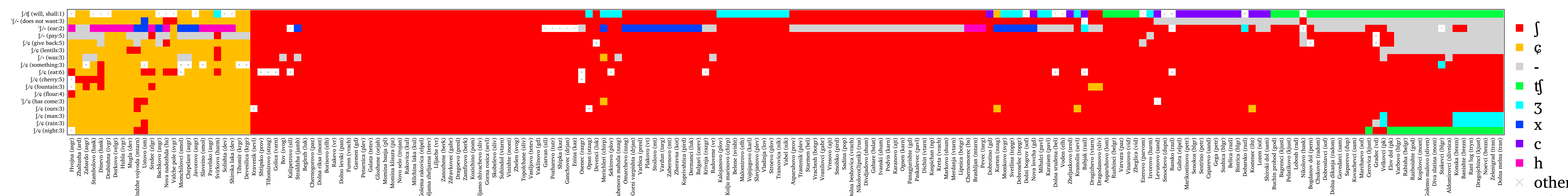
Bulgarian dialect data

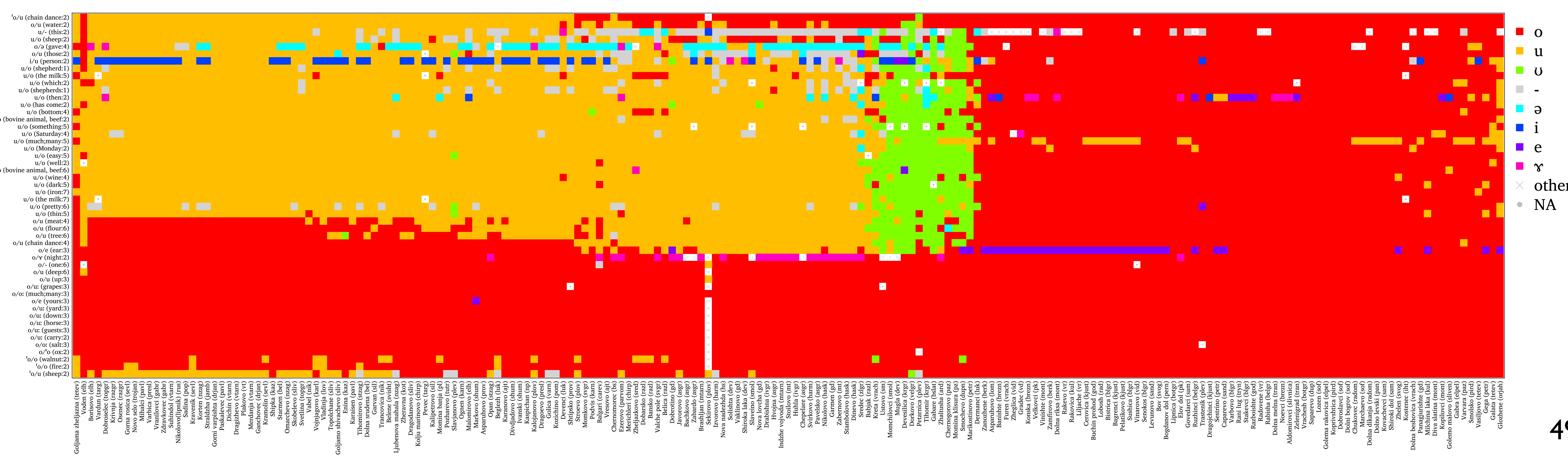
- Data from 197 villages for 152 words, resulting in 807 aligned columns
- There are many possible methods to cluster these into correspondence sets
- And they give different results!
  (though mostly smaller or larger sets)
- Here: a simple greedy clustering into 27 sets




ˈt/t (this:1)
ˈt/t (yours:1)
t/ˈt (such:1)
t (Saturday:5)
t/- (fear:3)
t/- (the middle:7)
t (old man:2)
t/tʲ (road:3)
t/tʲ (son-in-law; brother-in-law:3)
ˈt/ˈtʲ (dark:1)
ˈt/ˈtʲ (thin:1)
l/lʲ (deep:3)
l/- (wool:3)
l/lʲ (yellow:3)
l/- (apple:5)
l/ˈl (enter:3)
l/lʲ (rooster:5)
l (down:4)
l (hungry:2)
l (head:2)
l (pay:3)
l/ˈl (white:3)
l/- (ox:3)
l/lʲ (whole:3)
l/- (has come:6)
lʲ/l (key:2)
lʲ/j (Sunday:5)
ˈl/ˈlʲ (easy:1)
ˈl/ˈlʲ (lentils:1)
ˈlʲ/ˈl (iron:3)
ˈlʲ/ˈl (bread:2)
l/lʲ (the milk:2)
l/lʲ (salt:4)
lʲ/l (apples:5)
k/- (today:6)
k/c (wolf:5)
k/c (Monday:10)
k/c (person:5)
k/ˈd (where:1)
ˈk/ˈc (key:1)
k/ˈk (hand:3)
k/ˈk (river:3)
ˈk/k (such:3)
k/c (sand:5)
k/c (apple:7)
k/ˈk (which:1)
k/ˈk (the milk:4)
ˈk (blood:1)
ˈk/k (horse:1)
ˈk/k (stone:1)
k (deep:7)
k/- (outside:4)
k (wait:3)
k/- (Friday:5)
k/ˈk (thin:4)
p/ˈp (bake:2)
p/b (pocket:3)
p/b (bread:4)
p/pʲ (rooster:1)
p/f (sickle:5)
ˈp/p (road:1)
ˈp/p (first:1)
ˈp/p (he-goat:1)
ˈp/p (pay:1)
p (Monday:1)
ˈp/ˈpʲ (ash:1)
ˈp/ˈpʲ (Friday:1)
x/f (dry:3)
x/- (fear:6)
x/- (saw:5)


r
rʲ
r̥
-
other
NA


ʃ
ɕ
-
ʧ
ʒ
x
c
h
other
NA

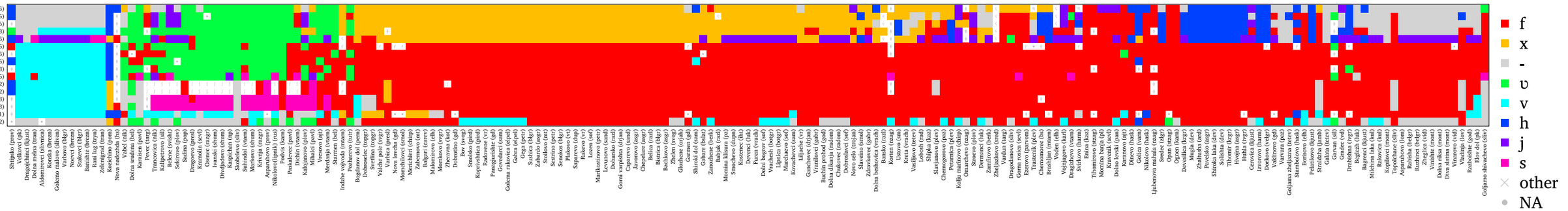

o
u
ʊ
-
ə
i
e
ɤ
other
NA

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

- Multialignments are an ideal data structure to compare language variation and change
- Extracting Multialignments from large data sets is not trivial, and bound to produce errors
- Multialignments can be profitably discussed and collectively improved