The analysis of typological data

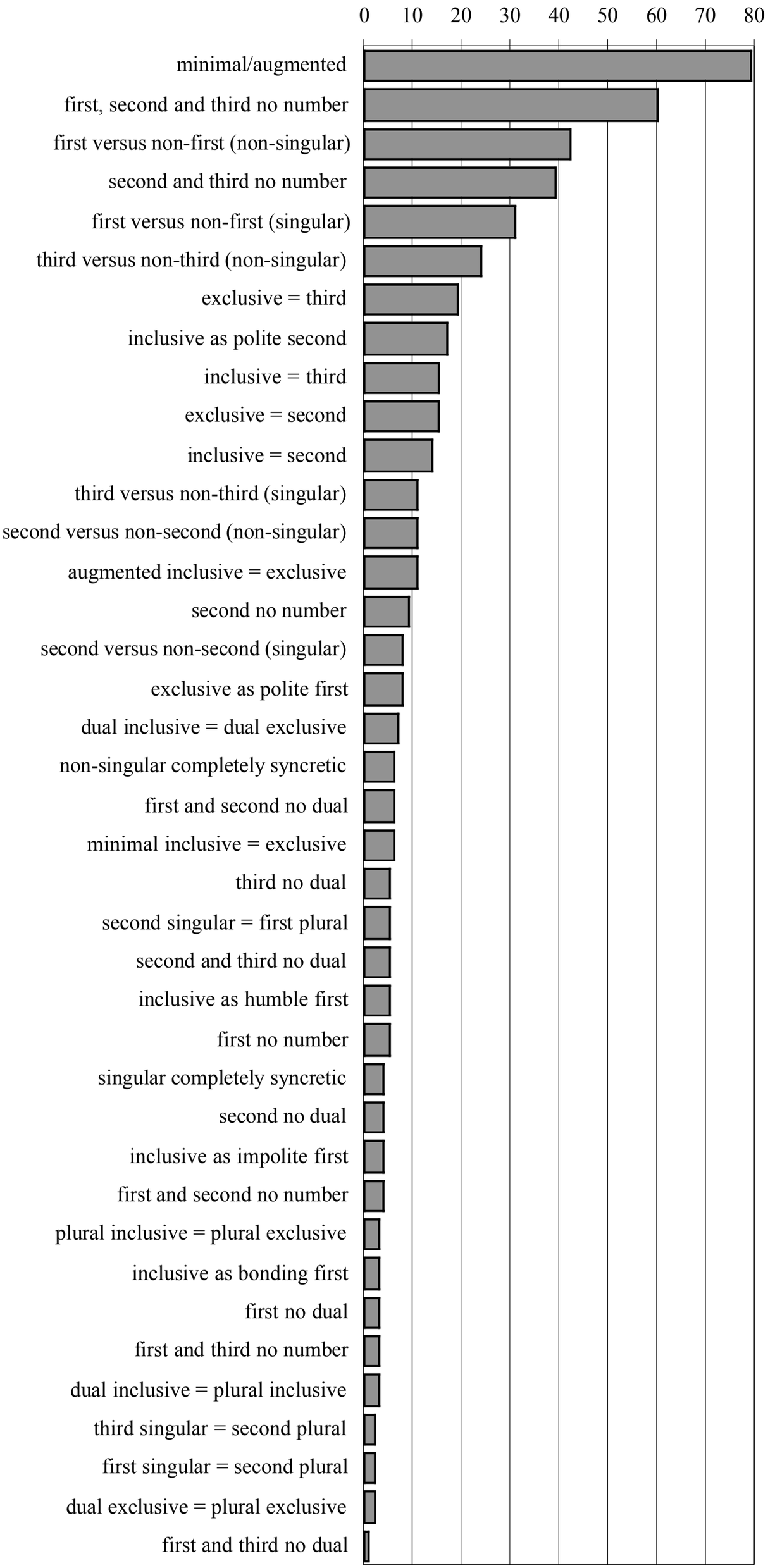
Michael Cysouw
Leipzig Spring School on Linguistics Diversity

Survey of this course

1. Collecting data
   - Choosing Languages
   - Establishing Types
2. Implicational Universals (and the like)
   - The typological tradition
   - Statistical view of things
   - Dryer's test
   - Summing up parameters (don't !)
3. Semantic Maps (and other graphs)
   - The typological tradition
   - Taking frequencies into account
4. Relationships between languages

1. Collecting data

- Investigate worldwide linguistic diversity
- Sample the world's languages
- Classify languages into types
- Any results are statements about **actual** and not **possible** human language !
- By sampling, only major types are captured

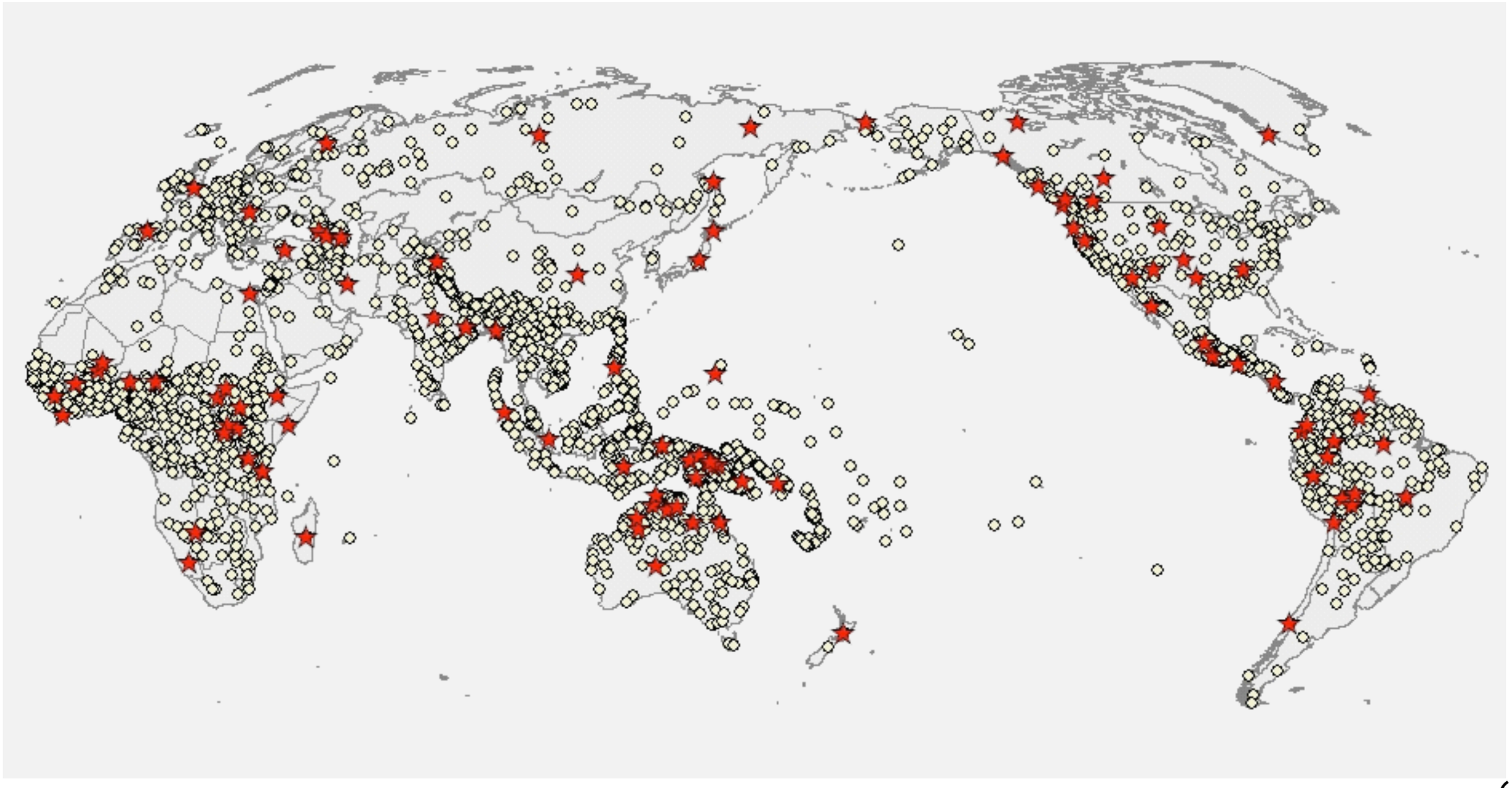

0
10
20
30
40
50
60
70
80
minimal/augmented
first, second and third no number
first versus non-first (non-singular)
second and third no number
first versus non-first (singular)
third versus non-third (non-singular)
exclusive = third
inclusive as polite second
inclusive = third
exclusive = second
inclusive = second
third versus non-third (singular)
second versus non-second (non-singular)
augmented inclusive = exclusive
second no number
second versus non-second (singular)
exclusive as polite first
dual inclusive = dual exclusive
non-singular completely syncretic
first and second no dual
minimal inclusive = exclusive
third no dual
second singular = first plural
second and third no dual
inclusive as humble first
first no number
singular completely syncretic
second no dual
inclusive as impolite first
first and second no number
plural inclusive = plural exclusive
inclusive as bonding first
first no dual
first and third no number
dual inclusive = plural inclusive
third singular = second plural
first singular = second plural
dual exclusive = plural exclusive
first and third no dual

Choosing languages

- Tradition: sample from linguistic families
- Indeed: don’t take 20 Indo-European languages and 5 other
- Watch out for large areal consistencies !

Diversity Sample

(from Ö. Dahl, forthcoming)



Choosing languages

- Tradition: sample from linguistic families
- Indeed: don’t take 20 Indo-European languages and 5 other
- Watch out for large areal consistencies !

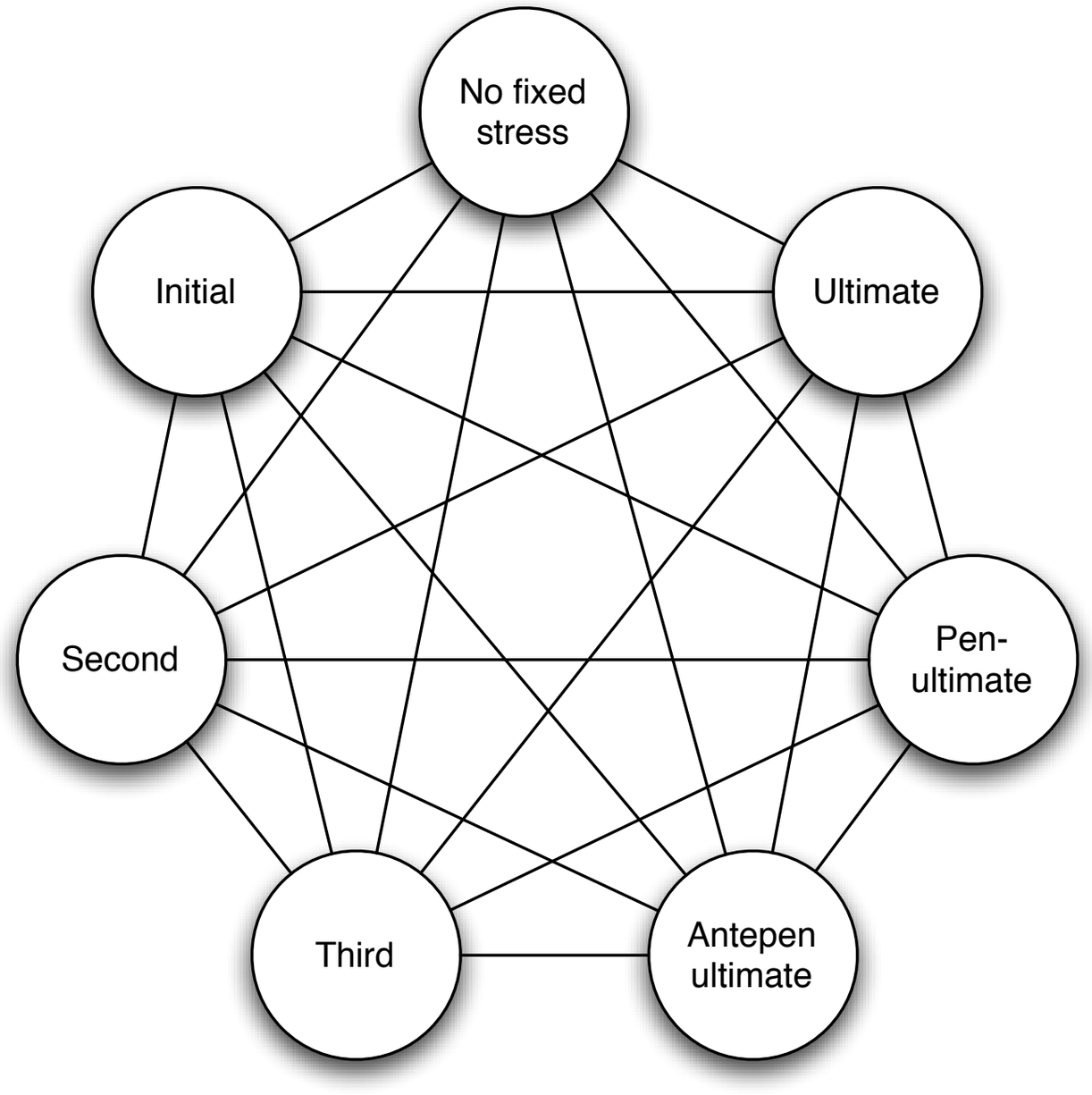
Choosing languages

- Tradition: sample from linguistic families
- Indeed: don’t take 20 Indo-European languages and 5 other
- Watch out for large areal consistencies !
- Watch out for internal variation in families !

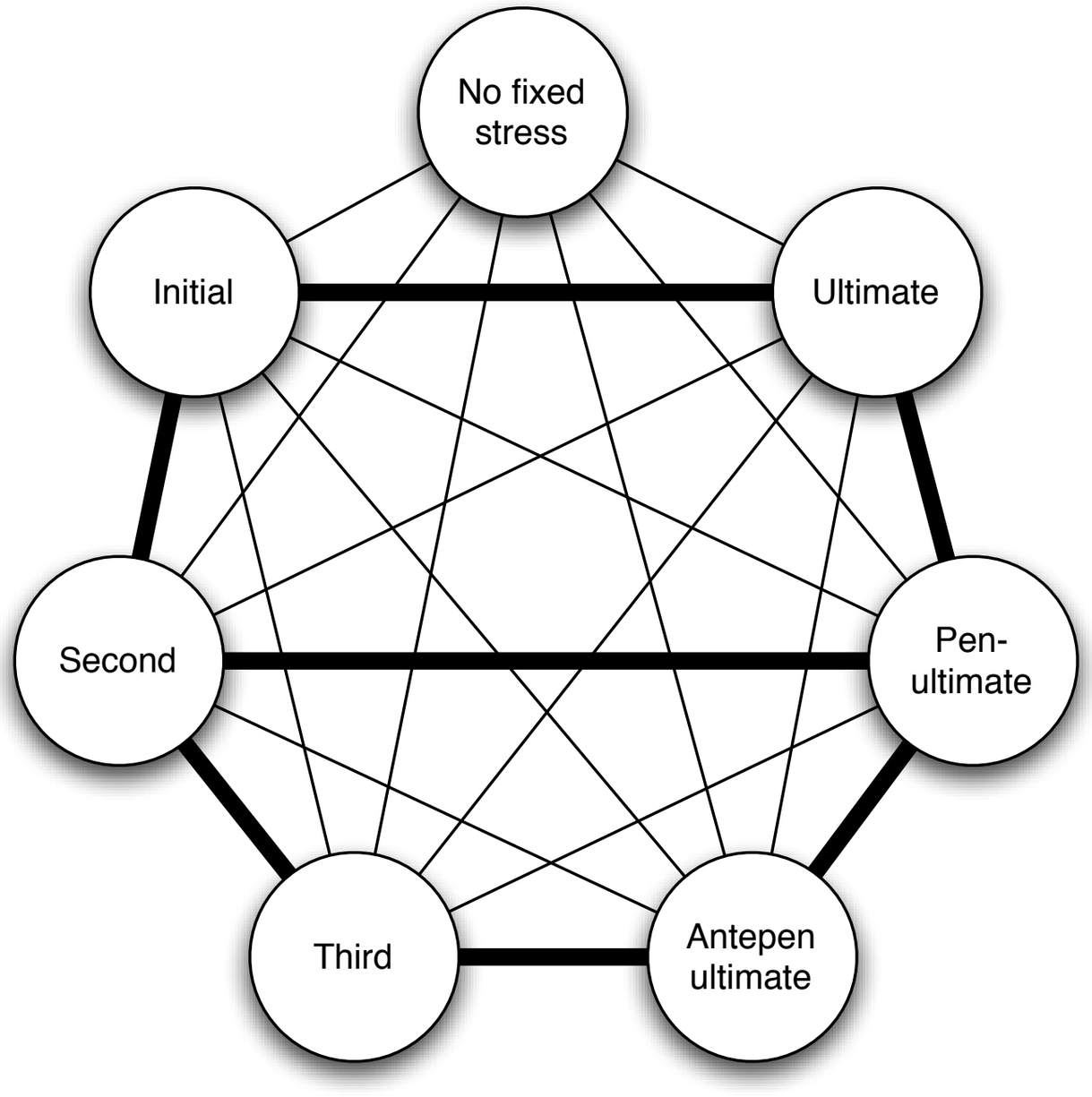
Establishing types

- Don’t group the dissimilar !
- Specify internal structure of types
  - Based on definitional structure of types
  - Based on empirical measures of similarity

Undifferentiated Typology



Including Similarities



Undifferentiated Typology

	T_1	T_2	T_3	T_4	T_5	T_6	T_7
T_1	I						
T_2		I					
T_3			I				
T_4				I			
T_5					I		
T_6						I	
T_7							I

Undifferentiated Typology

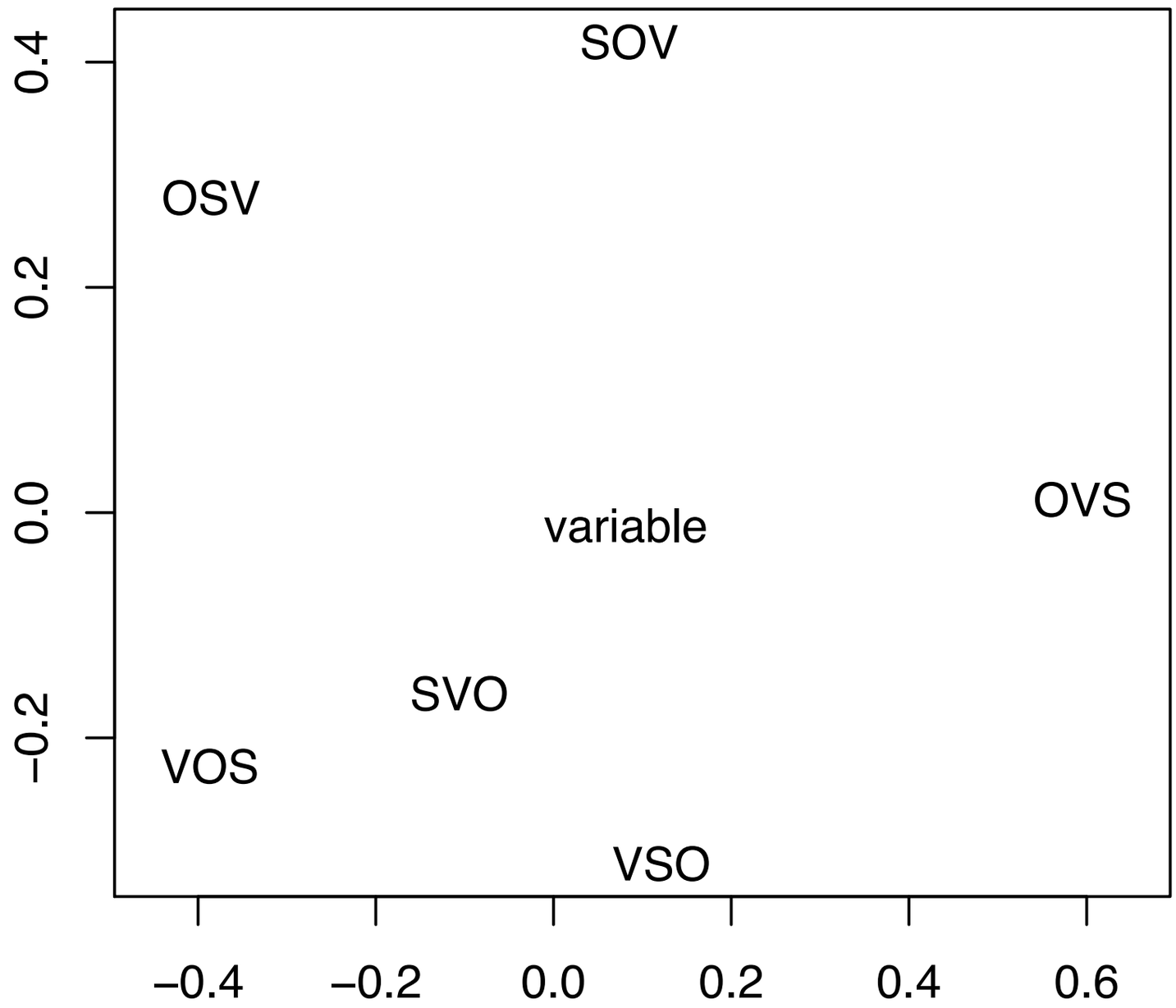
	T_1	T_2	T_3	T_4	T_5	T_6	T_7
T_1	1	0	0	0	0	0	0
T_2	0	1	0	0	0	0	0
T_3	0	0	1	0	0	0	0
T_4	0	0	0	1	0	0	0
T_5	0	0	0	0	1	0	0
T_6	0	0	0	0	0	1	0
T_7	0	0	0	0	0	0	1

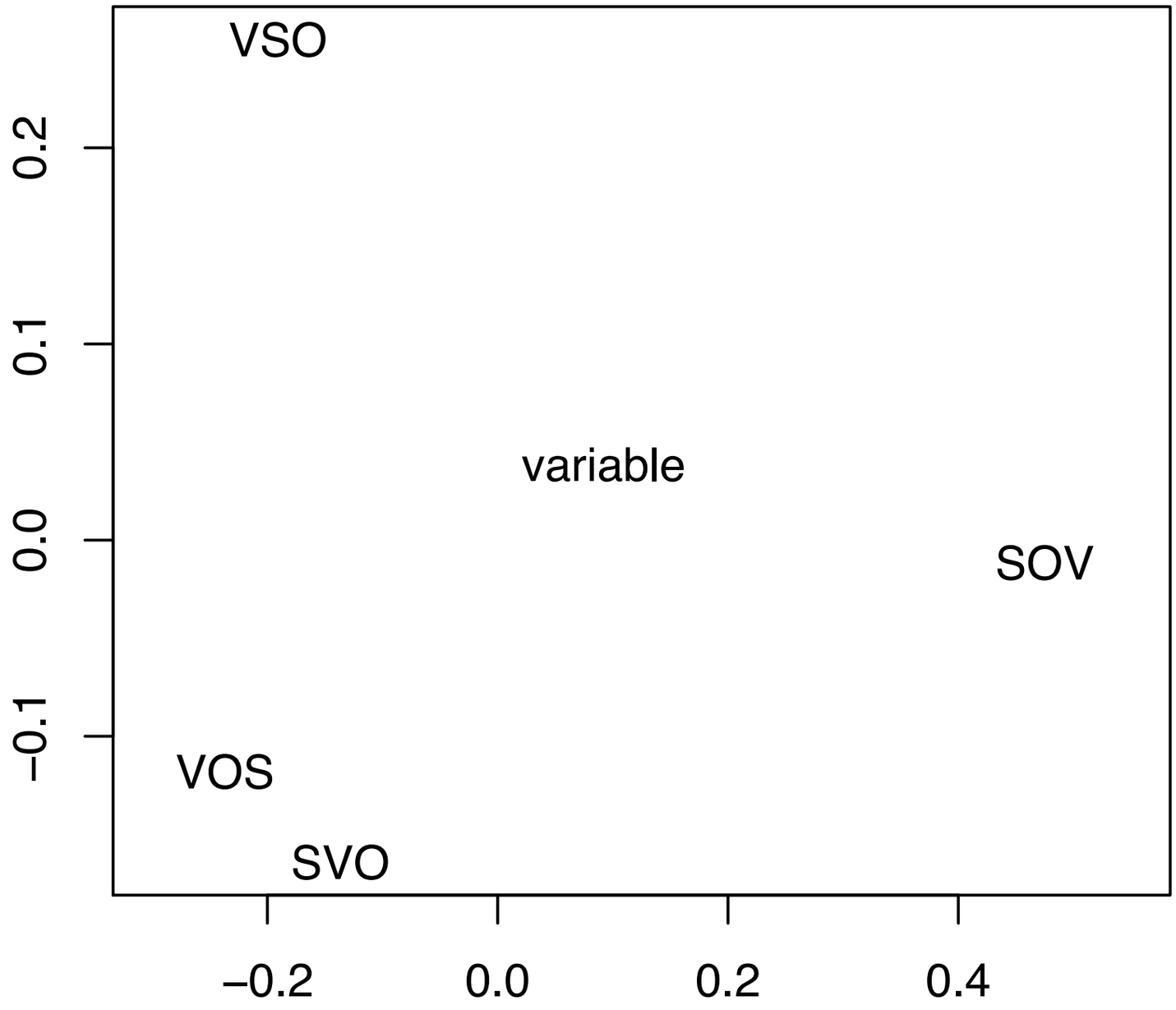
Specifying similarities

	T_1	T_2	T_3	T_4	T_5	T_6	T_7
T_1	1						
T_2		1					
T_3			1				
T_4				1			
T_5					1		
T_6						1	
T_7							1

Specifying similarities

	T_1	T_2	T_3	T_4	T_5	T_6	T_7
T_1	1	0.37	0.28	0.27	0.51	0.76	0.66
T_2	0.37	1	0.58	0.78	0.35	0.51	0.65
T_3	0.28	0.58	1	0.6	0.55	0.28	0.67
T_4	0.27	0.78	0.6	1	0	0.58	0.70
T_5	0.51	0.35	0.55	0	1	0	0.68
T_6	0.76	0.51	0.28	0.58	0	1	0.51
T_7	0.66	0.65	0.67	0.70	0.68	0.51	1


SOV
OSV
variable
OVS
SVO
VOS
VSO
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6


VSO
variable
SOV
VOS
SVO
0.2
0.1
0.0
−0.1
−0.2
0.0
0.2
0.4

‘Deconstructing’ Typology

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1									
L_2									
L_3									
L_4									
L_5									
L_6									
L_7									
L_8									
…									

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	I								
L_2		I							
L_3			I						
L_4				I					
L_5					I				
L_6						I			
L_7							I		
L_8								I	
…									

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	I								
L_2		I							
L_3			I						
L_4				I					
L_5					I				
L_6						I			
L_7							I		
L_8								I	
…									

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	I	I	I						
L_2	I	I	I						
L_3	I	I	I						
L_4				I	I				
L_5				I	I				
L_6						I	I	I	
L_7						I	I	I	
L_8						I	I	I	
…									

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	1	1	1	0	0	0	0	0	
L_2	1	1	1	0	0	0	0	0	
L_3	1	1	1	0	0	0	0	0	
L_4	0	0	0	1	1	0	0	0	
L_5	0	0	0	1	1	0	0	0	
L_6	0	0	0	0	0	1	1	1	
L_7	0	0	0	0	0	1	1	1	
L_8	0	0	0	0	0	1	1	1	
…									

Undifferentiated Typology

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	1	1	1	0.37	0.37	0.28	0.28	0.28	
L_2	1	1	1	0.37	0.37	0.28	0.28	0.28	
L_3	1	1	1	0.37	0.37	0.28	0.28	0.28	
L_4	0.37	0.37	0.37	1	1	0.58	0.58	0.58	
L_5	0.37	0.37	0.37	1	1	0.58	0.58	0.58	
L_6	0.28	0.28	0.28	0.58	0.58	1	1	1	
L_7	0.28	0.28	0.28	0.58	0.58	1	1	1	
L_8	0.28	0.28	0.28	0.58	0.58	1	1	1	
…									

Inter-type similarities

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	I								
L_2		I							
L_3			I						
L_4				I					
L_5					I				
L_6						I			
L_7							I		
L_8								I	
…									

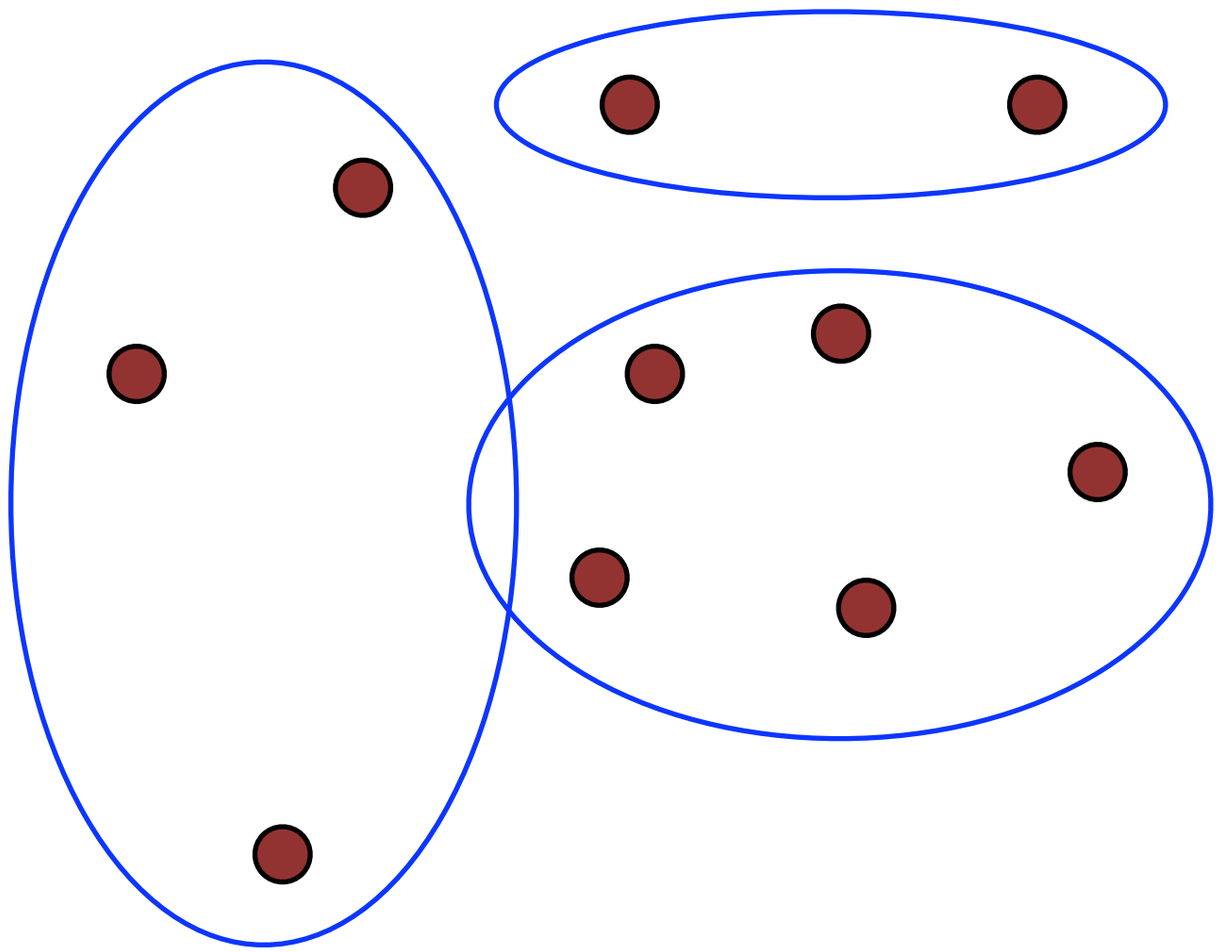
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	1	0.55	0.72	0.31	0.70	0.61	0.50	0.58	
L_2	0.55	1	0.55	0.31	0.40	0.44	0.31	0.48	
L_3	0.72	0.55	1	0.29	0.53	0.51	0.48	0.60	
L_4	0.31	0.31	0.29	1	0.38	0.36	0.26	0.27	
L_5	0.70	0.40	0.53	0.38	1	0.64	0.51	0.46	
L_6	0.61	0.44	0.51	0.36	0.64	1	0.57	0.43	
L_7	0.50	0.31	0.48	0.26	0.51	0.57	1	0.47	
L_8	0.58	0.48	0.60	0.27	0.46	0.43	0.47	1	
…									

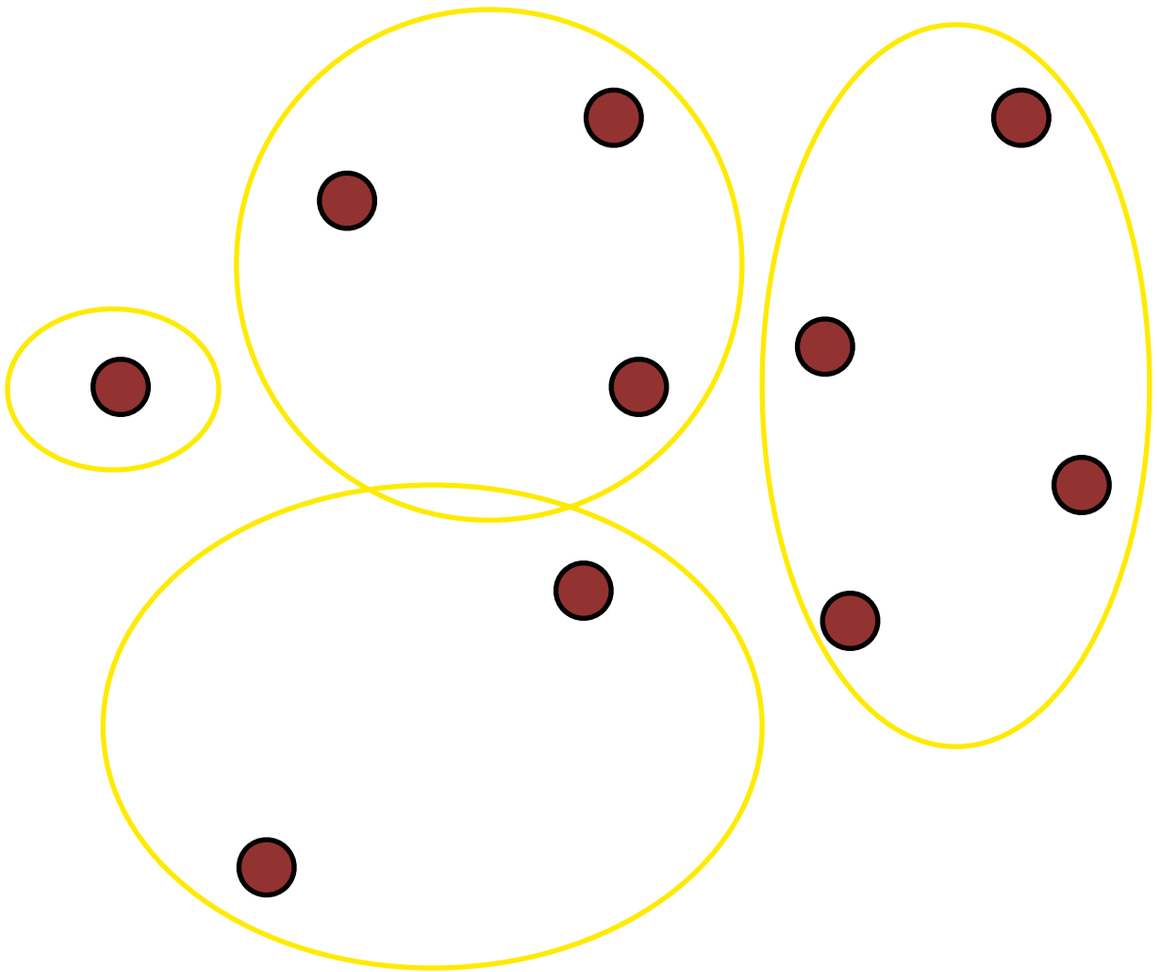
‘Deconstructed’ Typology

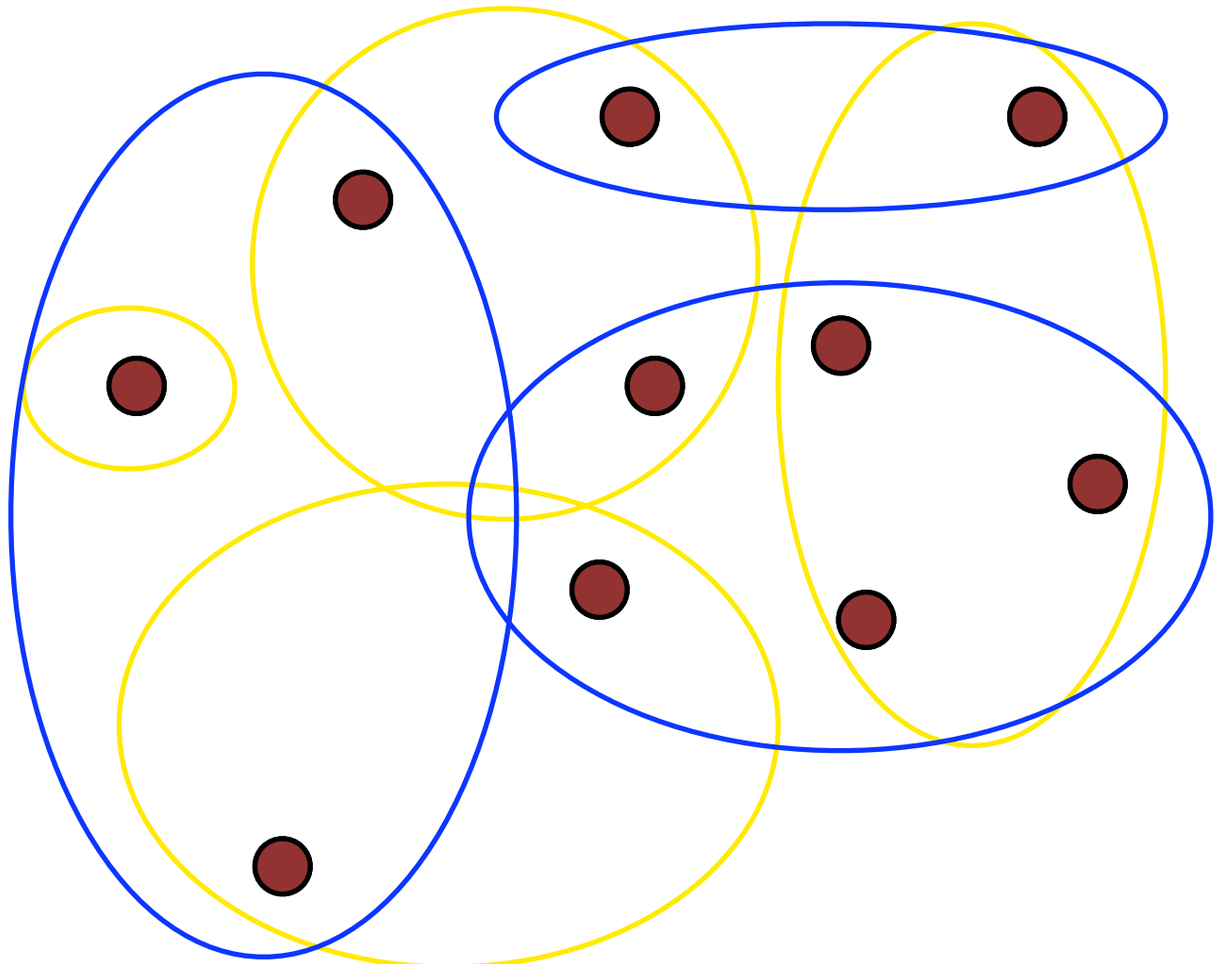
B. Wälchli's data on motion events

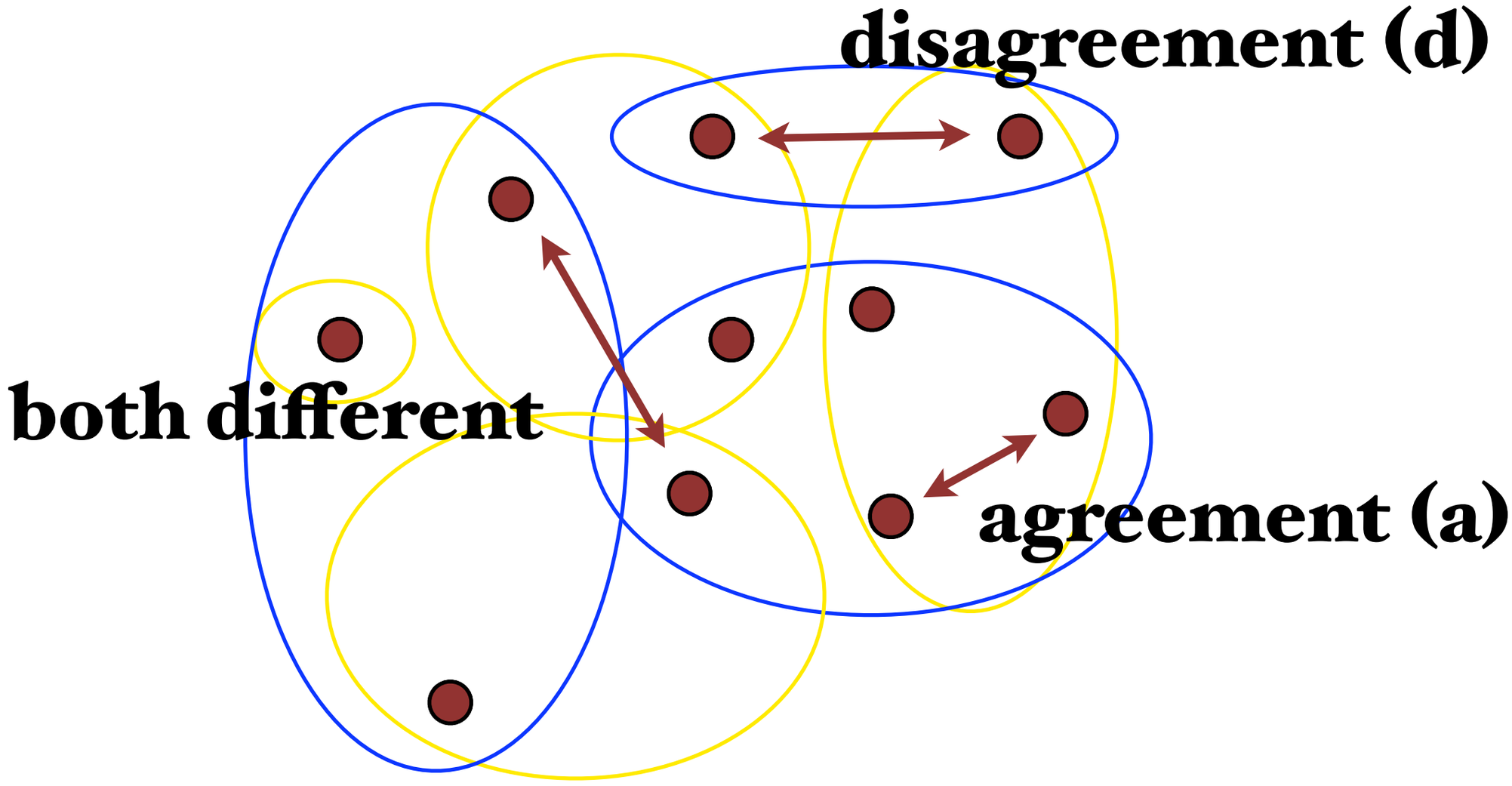
- 72 languages
- 335 clauses for each language from Bible
- clauses describing motion events
- here, only the lexical verb used is included
- ***contextually situated exemplars***

	MRD	LIT	ENG	FRE
1050	sams	eiti	go	aller
1070	sams	eiti	come	venir
1090	sams	eiti	come	venir
1104	lisems	kopti	come	sortir
1105	valgoms	zengti	descend	descendre
1114	–	–	come	se faire entendre
1120	vetjams	varyti	drive	pousser
1140	sams	eiti	come	se rendre
1160	jutams	eiti	walk	marcher



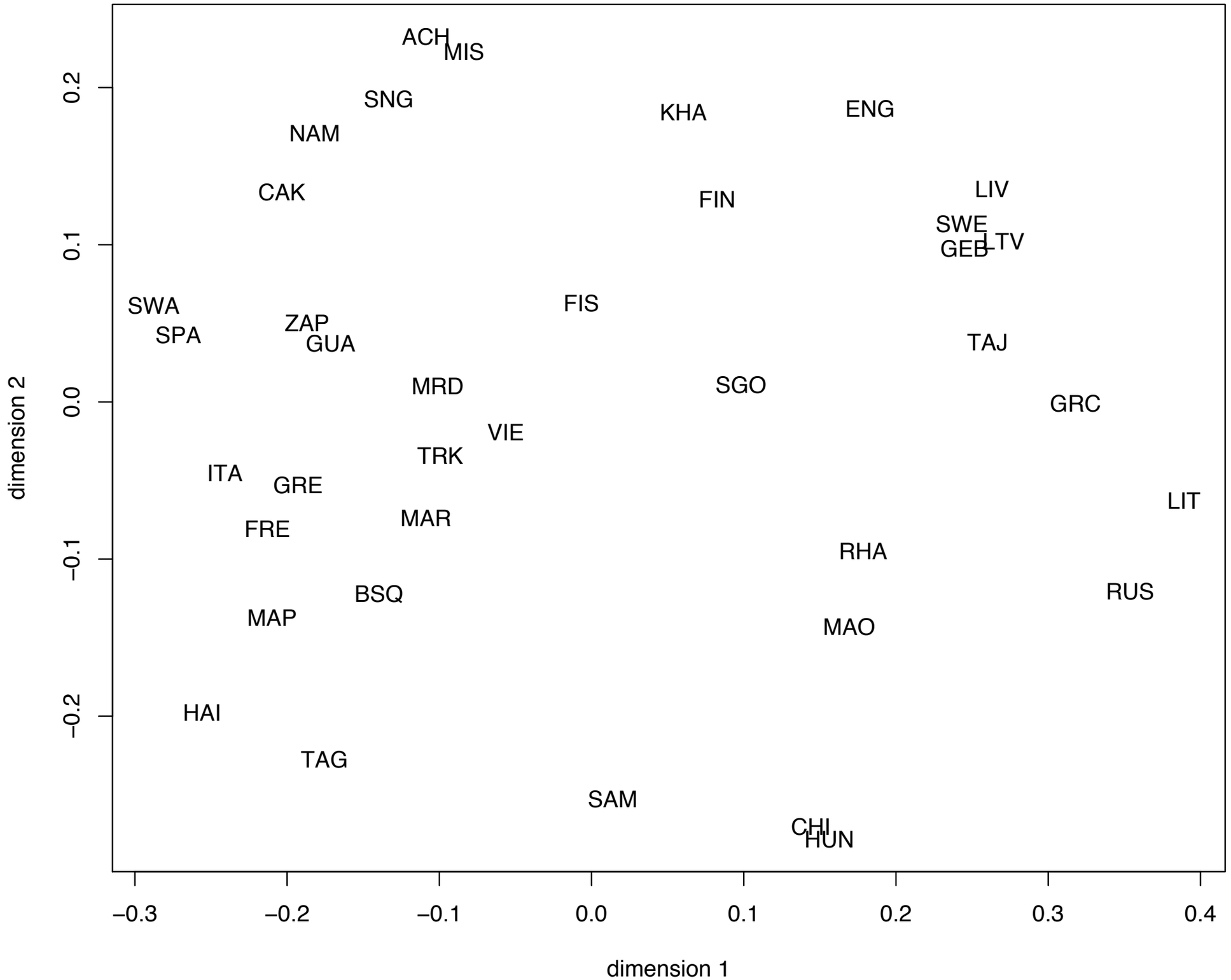


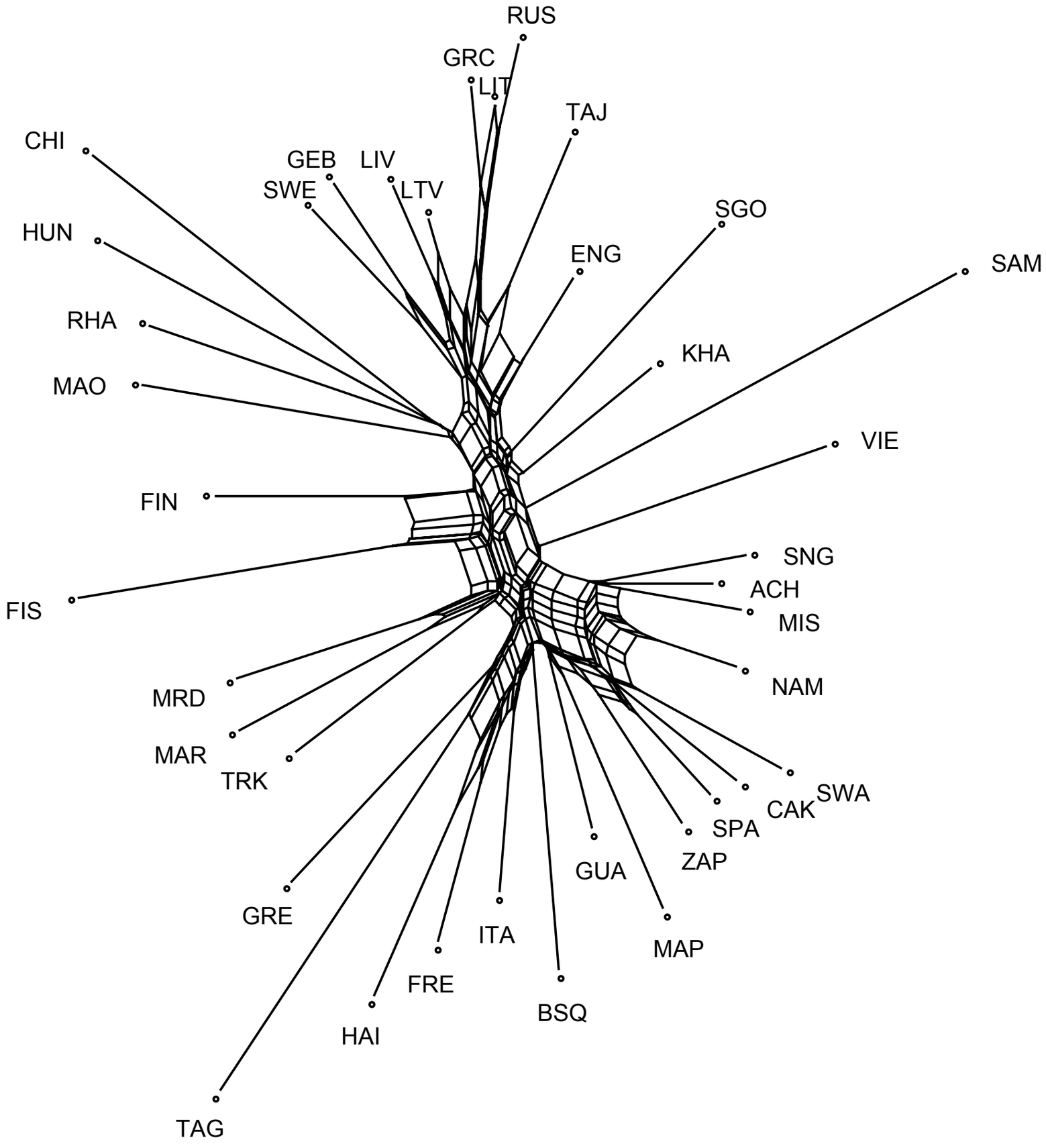




Jaccard distance:

$d/_{a+d}$


ACH
MIS
SNG
KHA
ENG
NAM
CAK
FIN
LIV
SWE
LTV
GEB
SWA
FIS
ZAP
SPA
GUA
TAJ
MRD
SGO
GRC
VIE
TRK
ITA
GRE
LIT
MAR
FRE
RHA
BSQ
RUS
MAP
MAO
HAI
TAG
SAM
CHI
HUN
0.2
0.1
0.0
−0.1
−0.2
dimension 2
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
dimension 1


RUS
GRC
LIT
TAJ
CHI
GEB
LIV
SWE
LTV
SGO
HUN
ENG
SAM
RHA
KHA
MAO
VIE
FIN
SNG
ACH
FIS
MIS
NAM
MRD
MAR
TRK
SWA
CAK
SPA
ZAP
GUA
GRE
ITA
MAP
FRE
BSQ
HAI
TAG

2. Implicational Universals (and the like)

- The typological tradition
- Statistical view of things
- Dryer's test (with variations and critique)
- Summing up parameters (don't !)

The typological tradition

- Implicational Universal
- Bidirectional Universal (Equivalence)
- Implicational Hierarchy
- Nested Implicational Universal

Greenberg (1963)

- *Universal 3*: Languages with dominant VSO order are always prepositional
- *Universal 2:* In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes

	A	B	C	D	
1	+	+	+	+	**26**
2	–	+	+	+	**78**
3	–	–	+	+	**99**
4	–	–	–	+	**20**
5	–	–	–	–	**21**
6	+	–	+	+	**3**
7	–	+	–	+	**12**
8	–	–	+	–	**4**
9	+	–	–	+	1
10	–	+	+	–	0
11	+	+	–	+	0
12	+	–	+	–	0
13	–	+	–	–	0
14	+	+	+	–	1
15	+	+	–	–	0
16	+	–	–	–	0
Total +	31	117	211	239	

Statistical view of things

What do typologists say?

Smallest number	Kind of universal	Hypothetical distributions of a 100-language sample							
Zero	Exceptionless universal	33	34	26	48	20	60	14	72
		0	33	**0**	26	**0**	20	**0**	14
Five	Strong tendency	36	23	31	33	27	41	22	51
		5	36	**5**	31	**5**	27	**5**	22
Ten	Statistical tendency	38	14	33	24	30	30	25	40
		10	38	**10**	33	**10**	30	**10**	25
Fifteen	Maybe something			35	15	31	23	28	29
				15	35	**15**	31	**15**	28
Nineteen	Nothing					31	19	27	27
						19	31	**19**	27

What do statisticians say?

Hypothetical distributions of a 100-language sample

33	34	26	48	20	60	14	72
0	33	**0**	26	**0**	20	**0**	14
36	23	31	33	27	41	22	51
5	36	**5**	31	**5**	27	**5**	22
38	14	33	24	30	30	25	40
10	38	**10**	33	**10**	30	**10**	25
		35	15	31	23	28	29
		15	35	**15**	31	**15**	28
				31	19	27	27
				19	31	**19**	27

Kind of interaction	Very strongly significant	Strongly significant	Significant	No interaction
Fisher's Exact two-tailed	$p < 0.000001$	$p < 0.001$	$p < 0.05$	$p > 0.2$

Fisher Exact

Table:

10	23
43	21

Calculate

Help Fisher Exact

This procedure by SISA, 1989,1997,2000.

```
The total number of cases= 97
The smallest value= 10
The smallest marginal= 33

Use of * (starred) statistics is advised

The p for exactly this table= 0.000449

The p-value for the same or a stronger
 association= 0.000555*
The p-value for a stronger association= 0.000105
The mid p-value= 0.000330

The p-value for the same or the reverse
 association=  0.999894

Two sided p-values for p(O>=E|O<=E)
  p-value= 0.0010737593* (the sum of small p's)
  p-value= 0.001110 (double the single sided p)
Two sided p-value for p(O>E|O<E)
  p-value= 0.0006242531 (the sum of small p's)
Two sided p-value mid-p
```

Clear

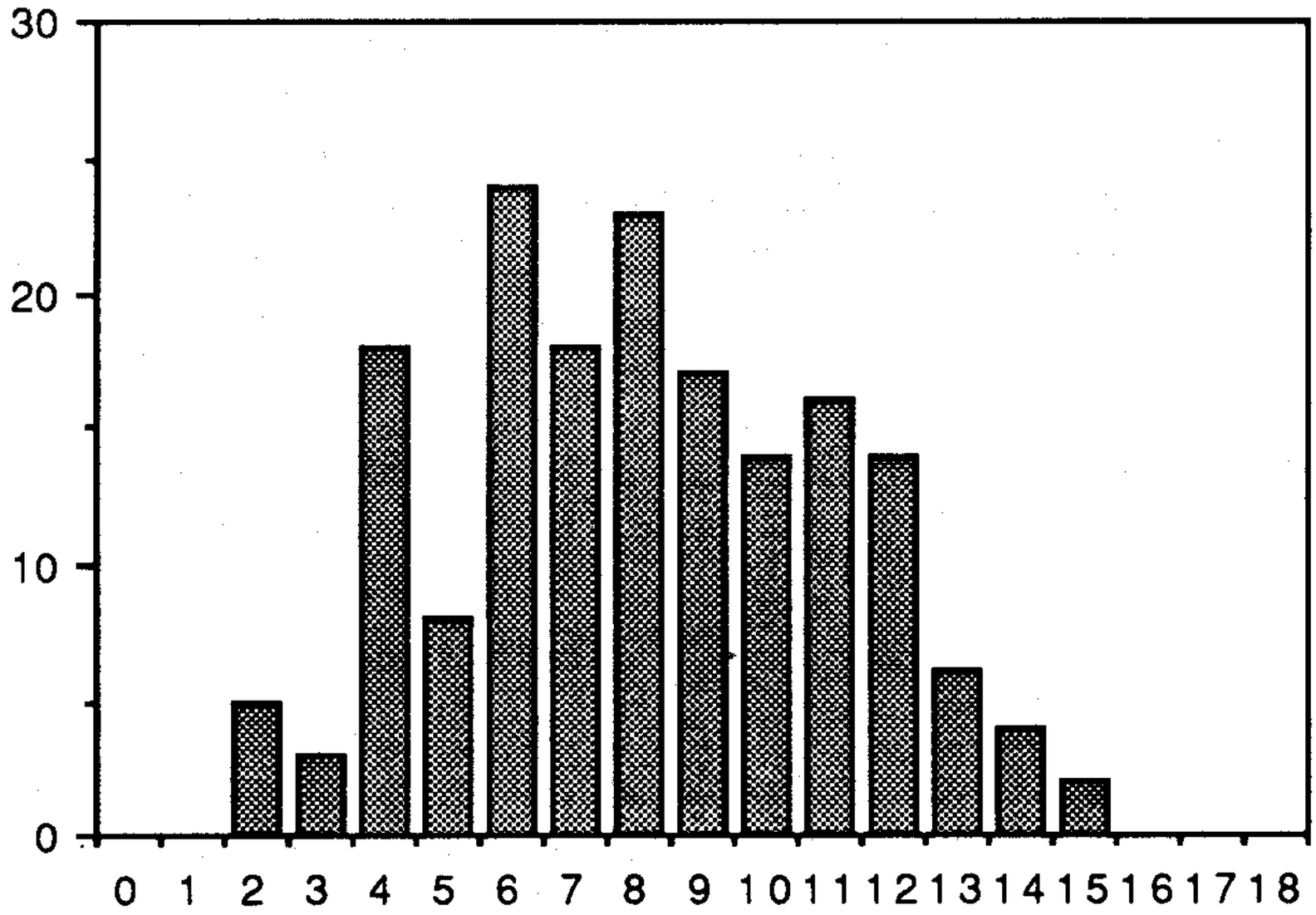
SISA home

A windows version of this procedure is available here

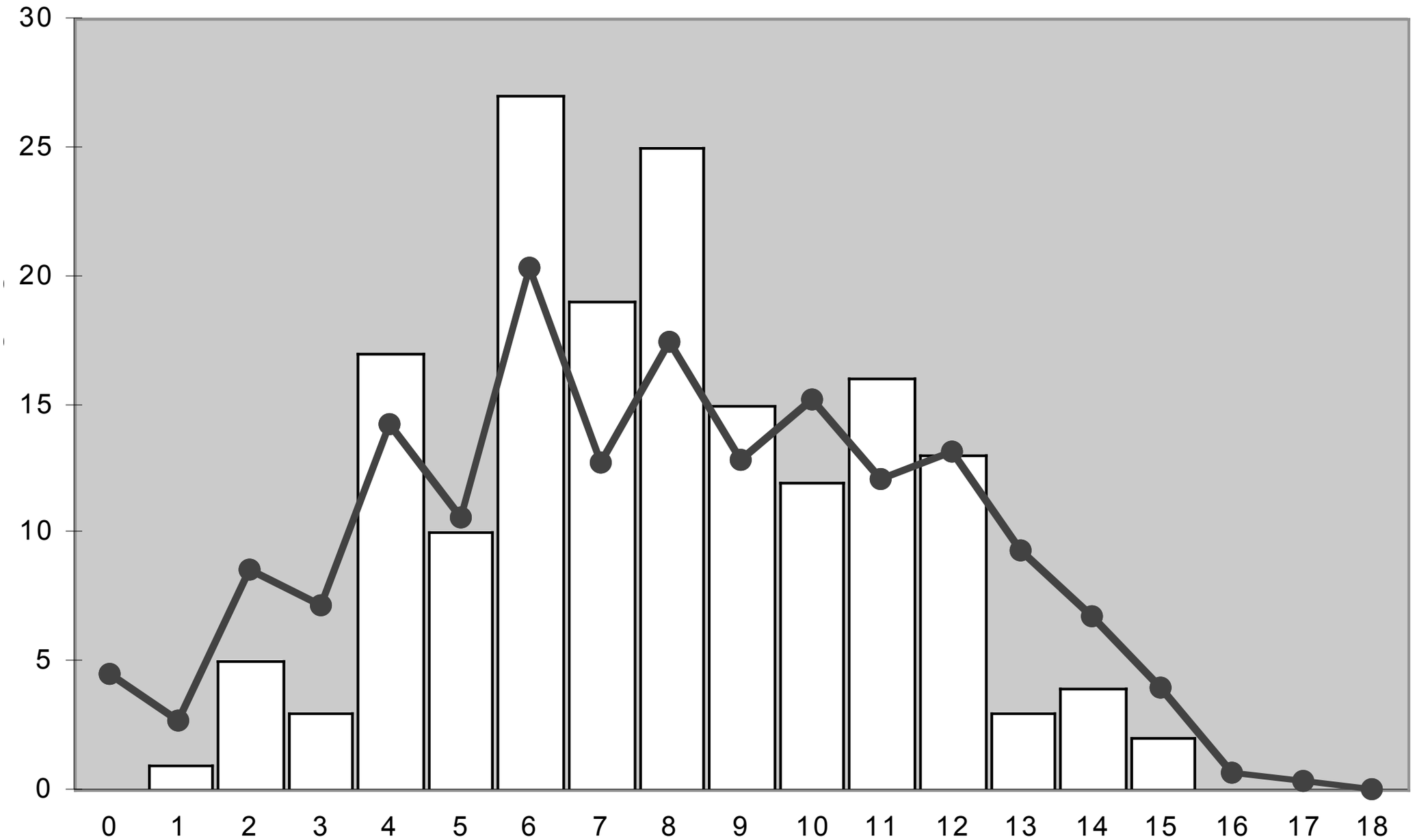
Dryer's test

Summing up parameters (**Don’t !**)

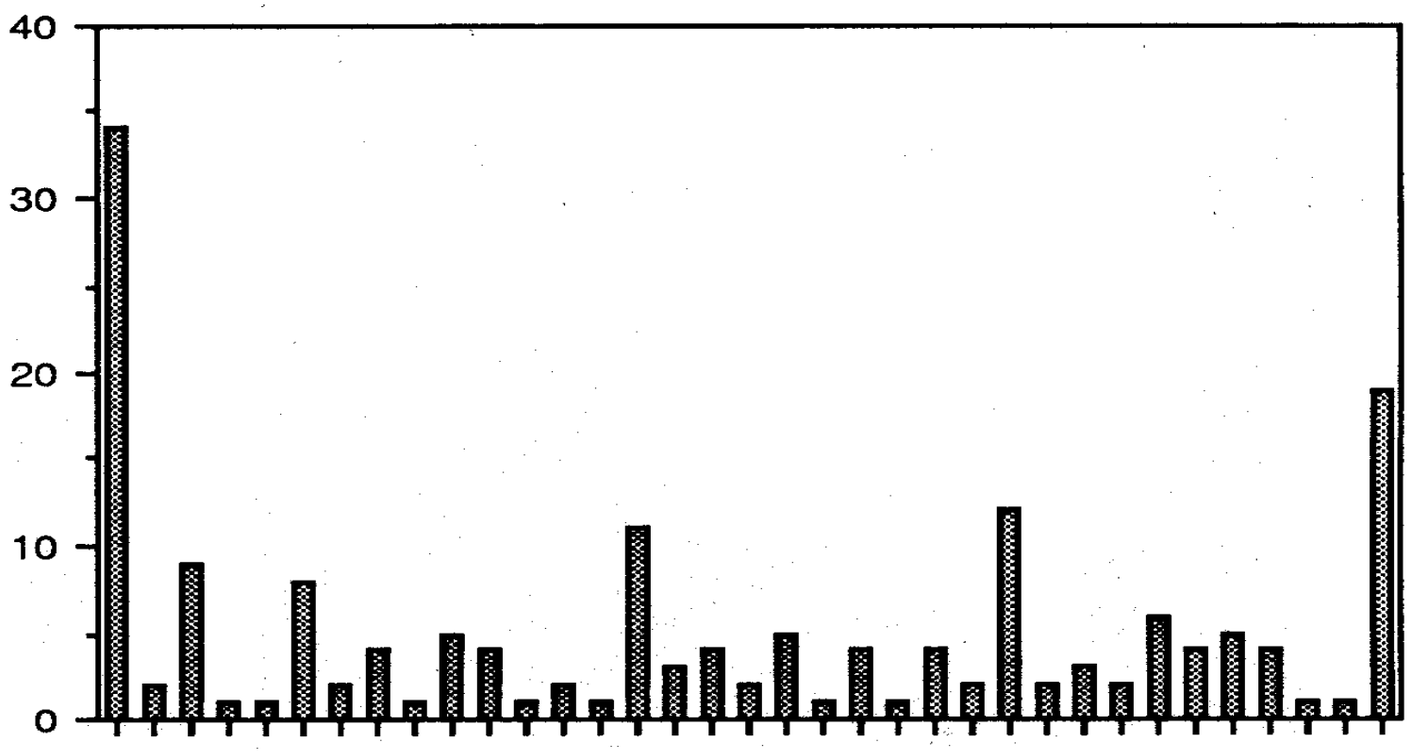
Sum of Head and Dependent marking: 'complexity':



'… the complexity (Dependent points plus Head points …) has a roughly normal distribution. Neither zero complexity nor the theoretical maximum complexity of [18] points (9 Head points plus 9 Dependent points …) occurs. the highest attested complexity is 15, found in only two languages. Figure 4 shows the complexity values attested in my sample. … The normal distribution and preference for moderate complexity shown in the overall sample are echoed in most … areas, with high complexity predominating in only two.' (Nichols 1992: 88-89)


30
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

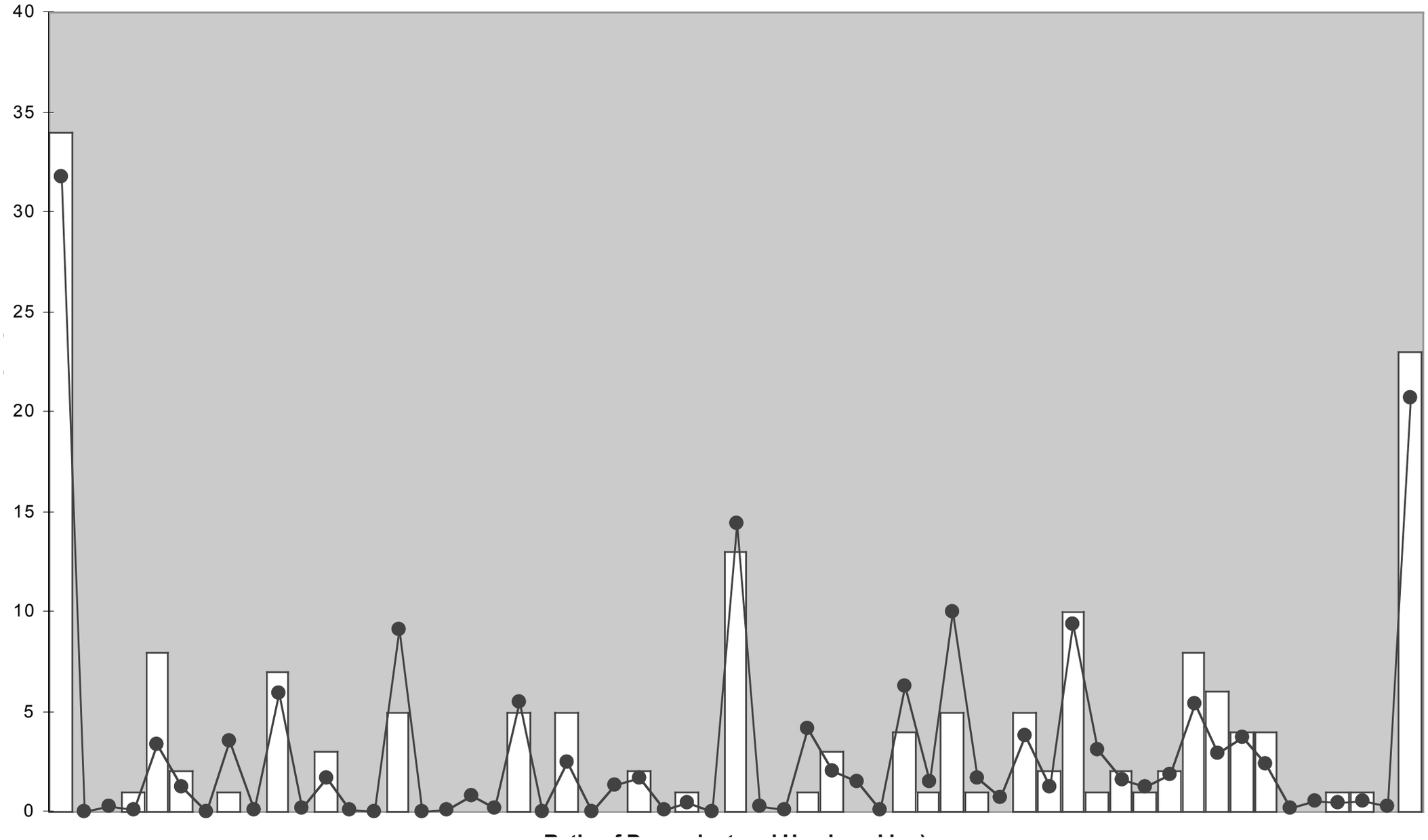
Ratio of Dependent and Head points: indicating the relative strength of head or dependent marking in a language.



'… computing the ration of dependent to head marking … gives us 35 different ratios among the 174 sample languages. Their distribution is shown in figure 1. It is bimodal, with the greatest peaks at the extremes of exclusive head marking (ration of zero since D = 0) and exclusive dependent marking (since H = 0, an actual ratio cannot be computed as it has a zero denominator). The other ratios, whose without zeroes, run from 0.14 (two languages) to 8.00 (one language). The highest frequencies are:

0.00	34 languages (radically head marking)	
0.17	9 languages	
0.50	8 languages	[should be '0.33', MC]
1.00	11 languages	
2.00	12 languages	
H = 0	19 languages (radically dependent marking)	

… The other three frequency peaks suggest that preferred patterns cluster at perceptually simple ratios: two to one, one to one, and one to two. Overall, then , we have a preferecne for neatness of some sort: polar types, two-to-one ratios and even splits.' (Nichols 1992: 72-73)


40
35
30
25
20
15
10
5
0

3. Semantic Maps

- Traditional view
  - Extension of Implicational Hierarchy
- Multidimensional Scaling

Extension of Hierarchy

Guarani: [first/second prn] – third prn – human N – animate N – inanimate N

Usan: [first/second prn – third prn] – human N – animate N – inanimate N

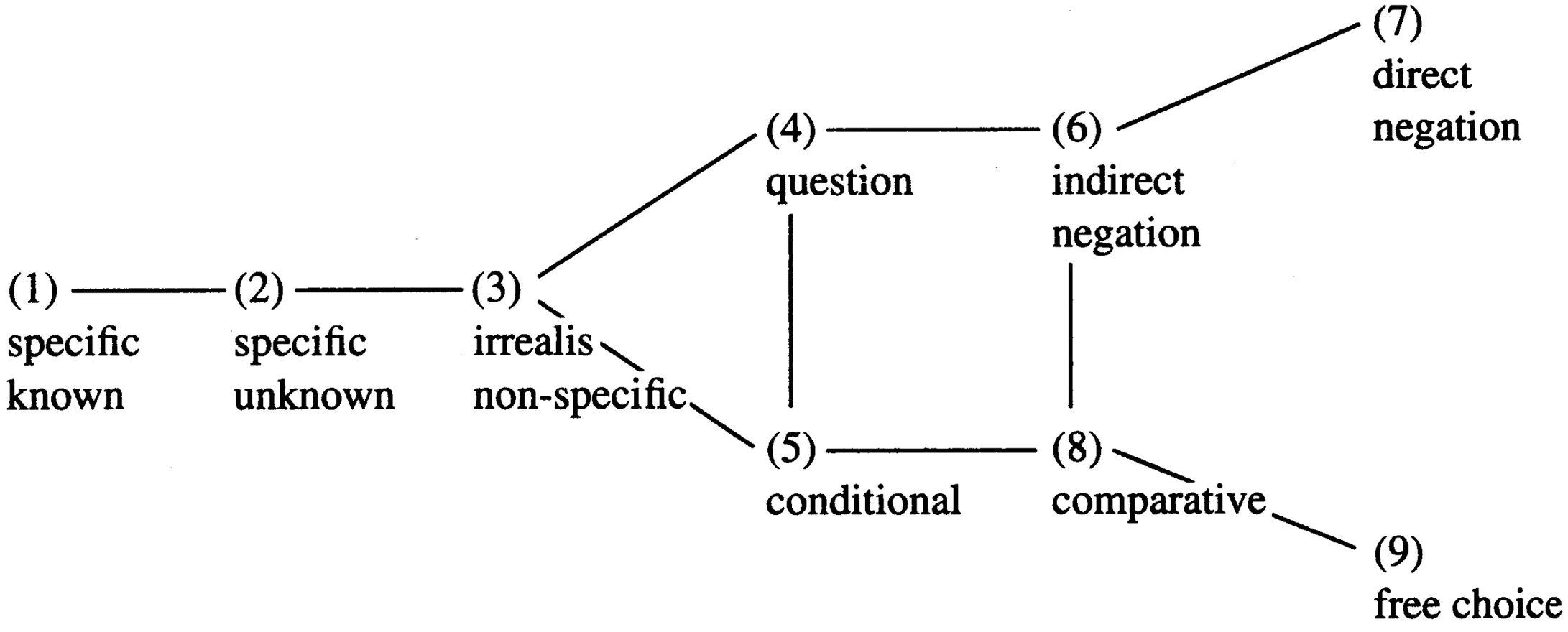
Tiwi: [first/second prn – third prn – human N] – animate N – inanimate N

Kharia: [first/second prn – third prn – human N – animate N] – inanimate N

English: [first/second prn – third prn – human N – animate N – inanimate N]

Figure 5.1 *Semantic maps of plural inflection in various languages*

Haspelmath (1997)



Examples of indefinite Pronouns

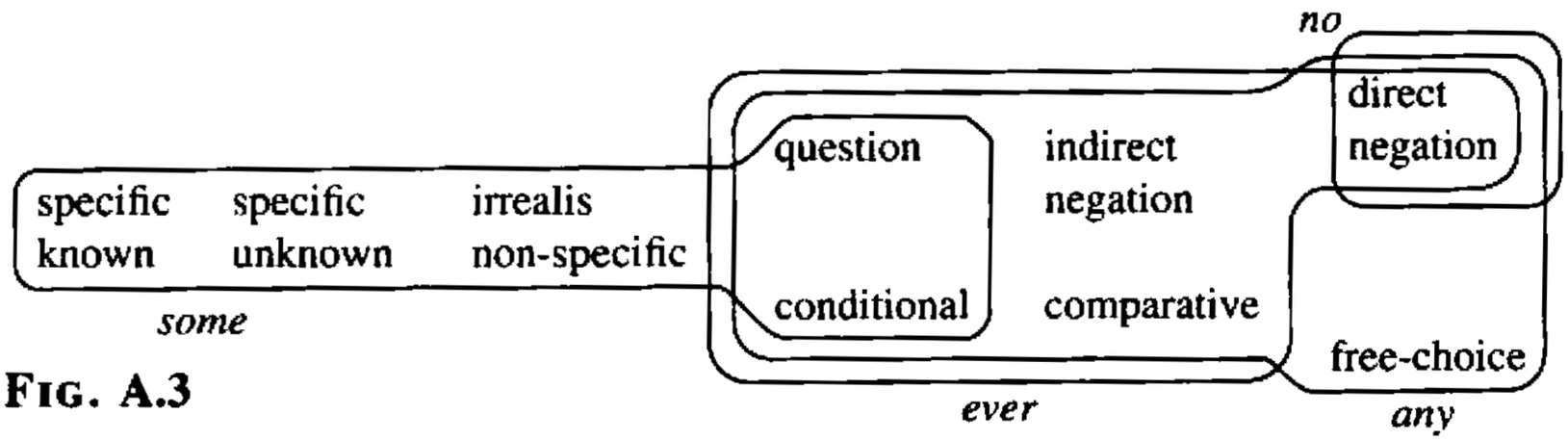


FIG. A.3

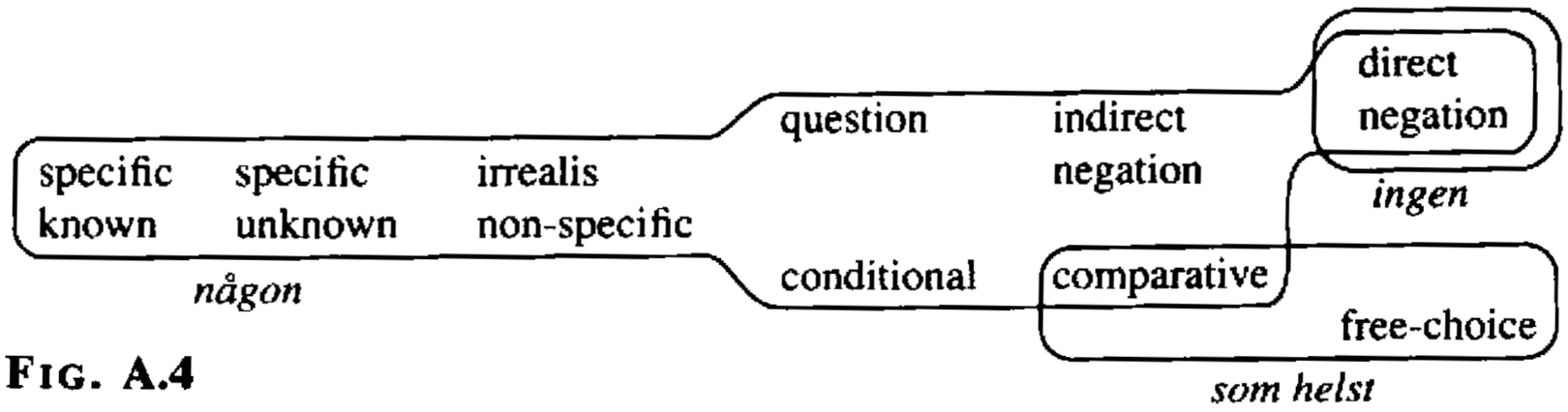
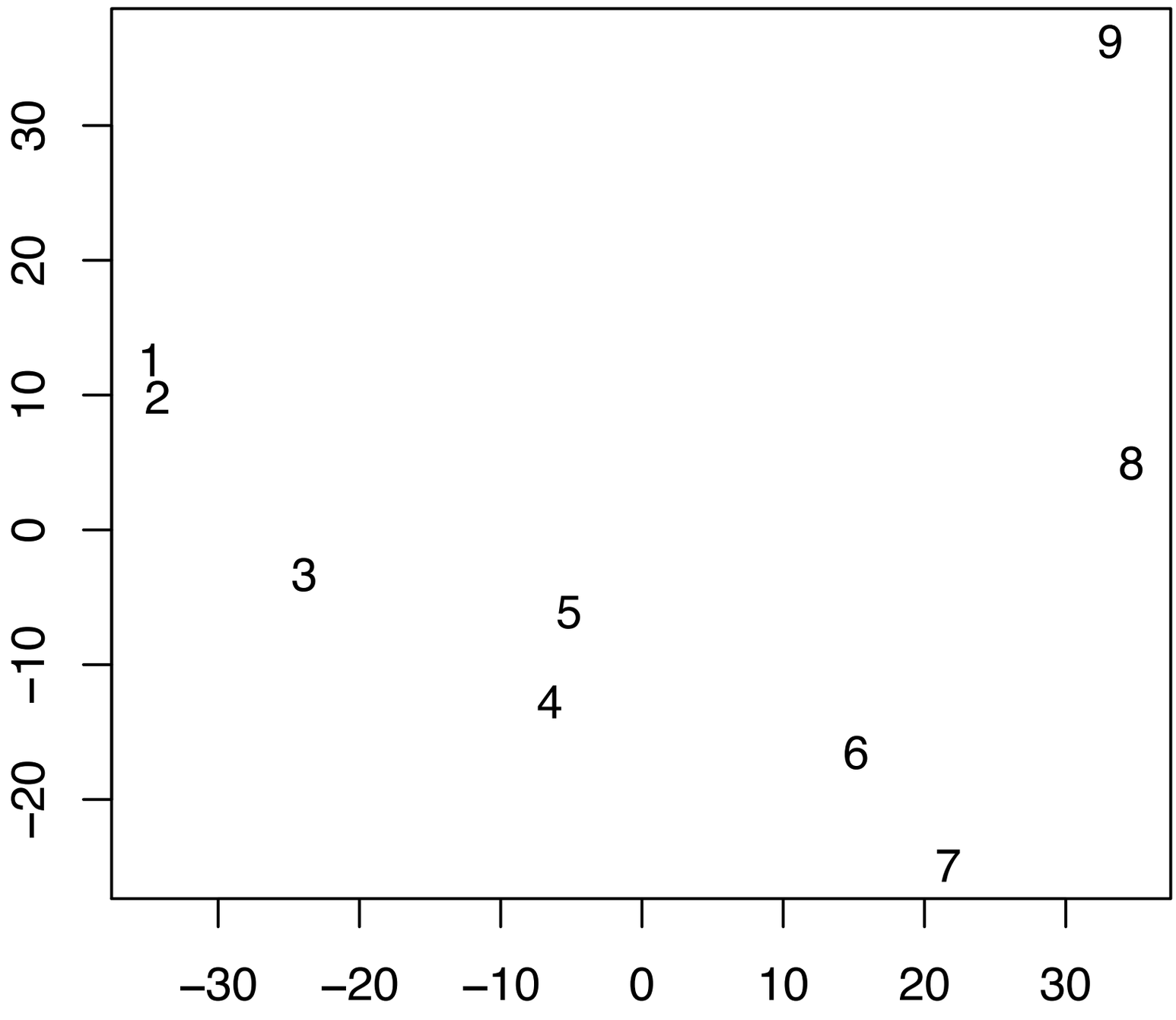


FIG. A.4

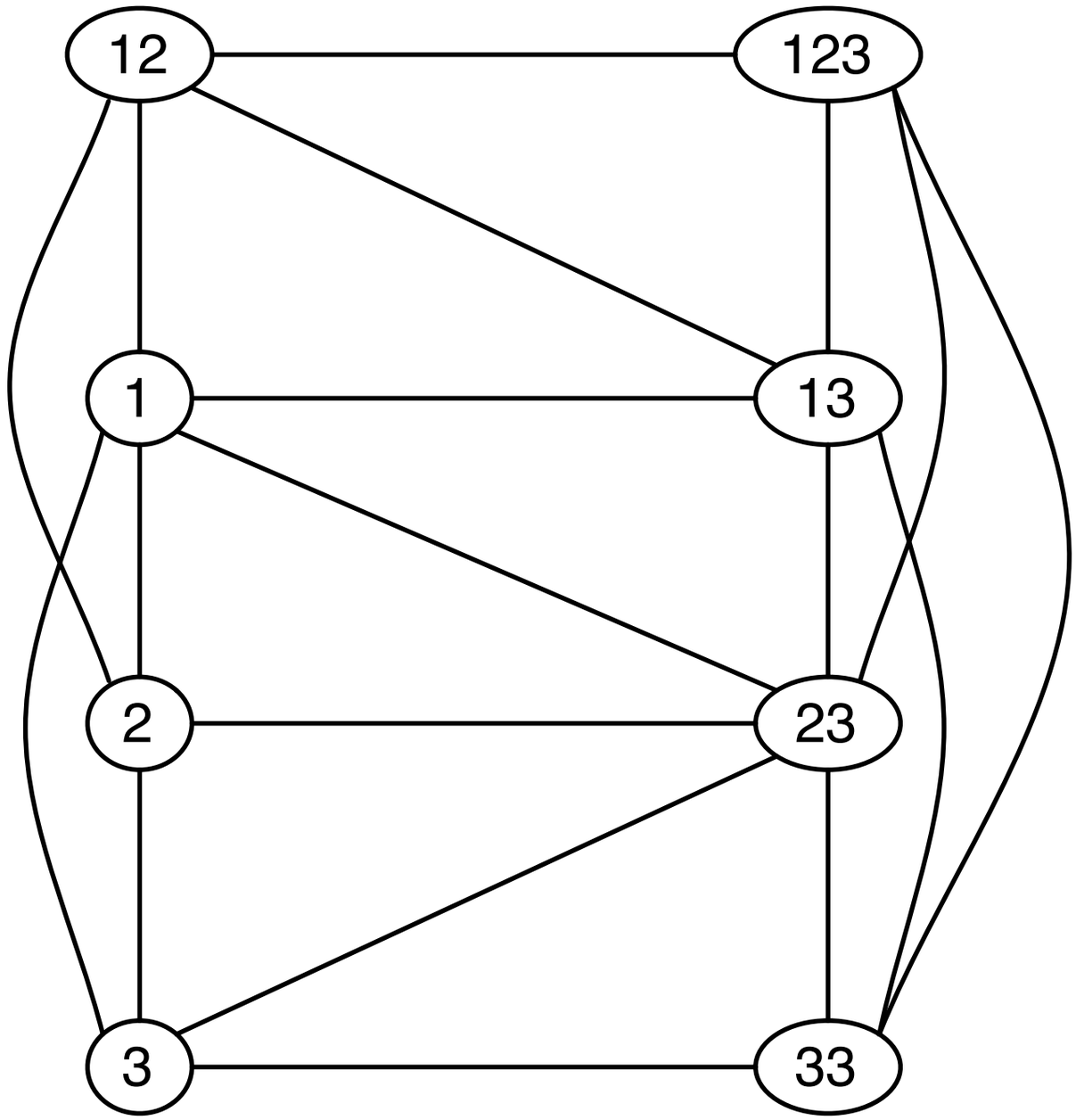
Evaluation

- Good: only 10 out of 45 (= 9x8/2) possible lines needed
- But: one line is indeterminated
- But: 105 groups predicted, though only 39 attested
- But: frequencies do not play a role

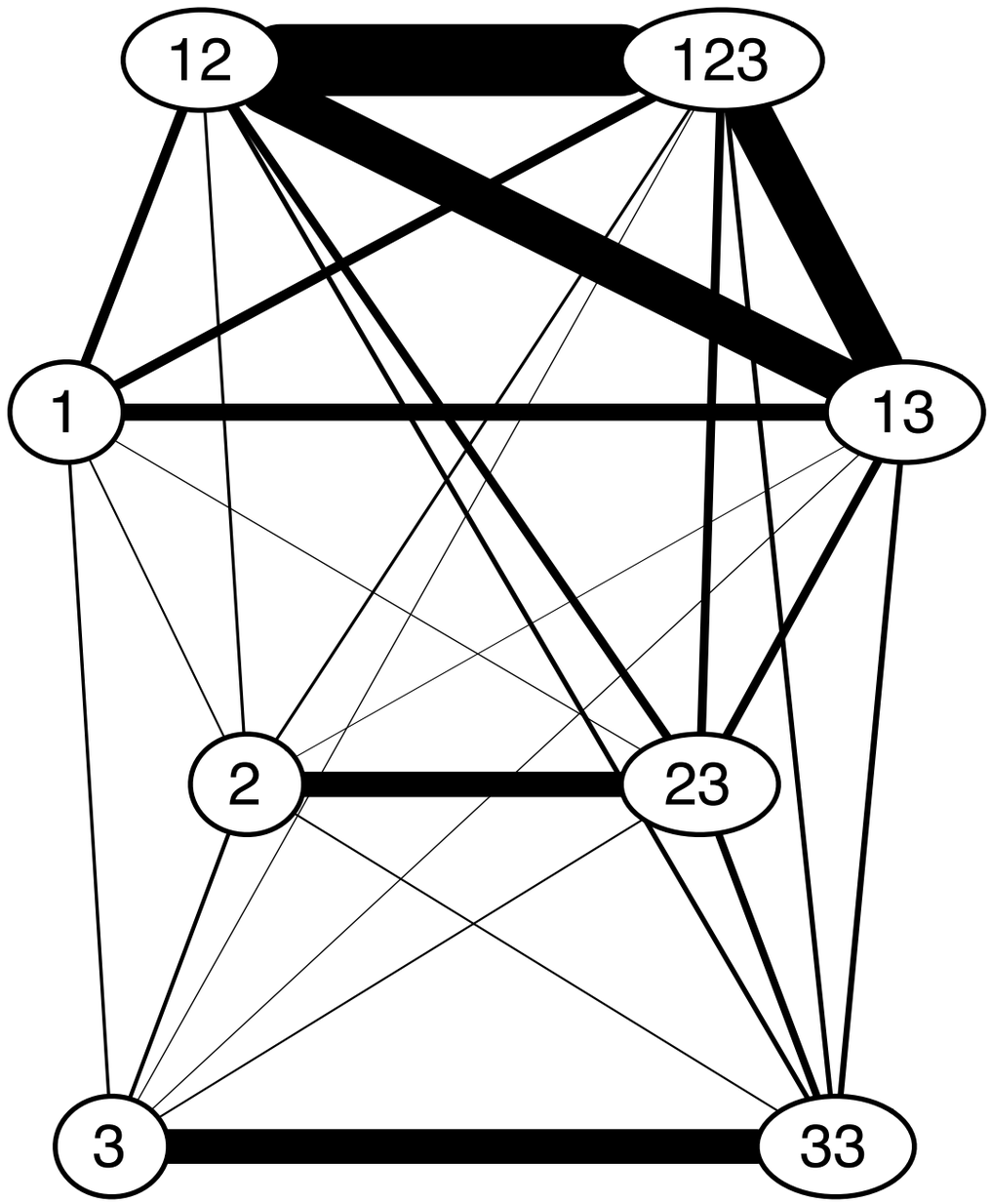
Multidimensional scaling



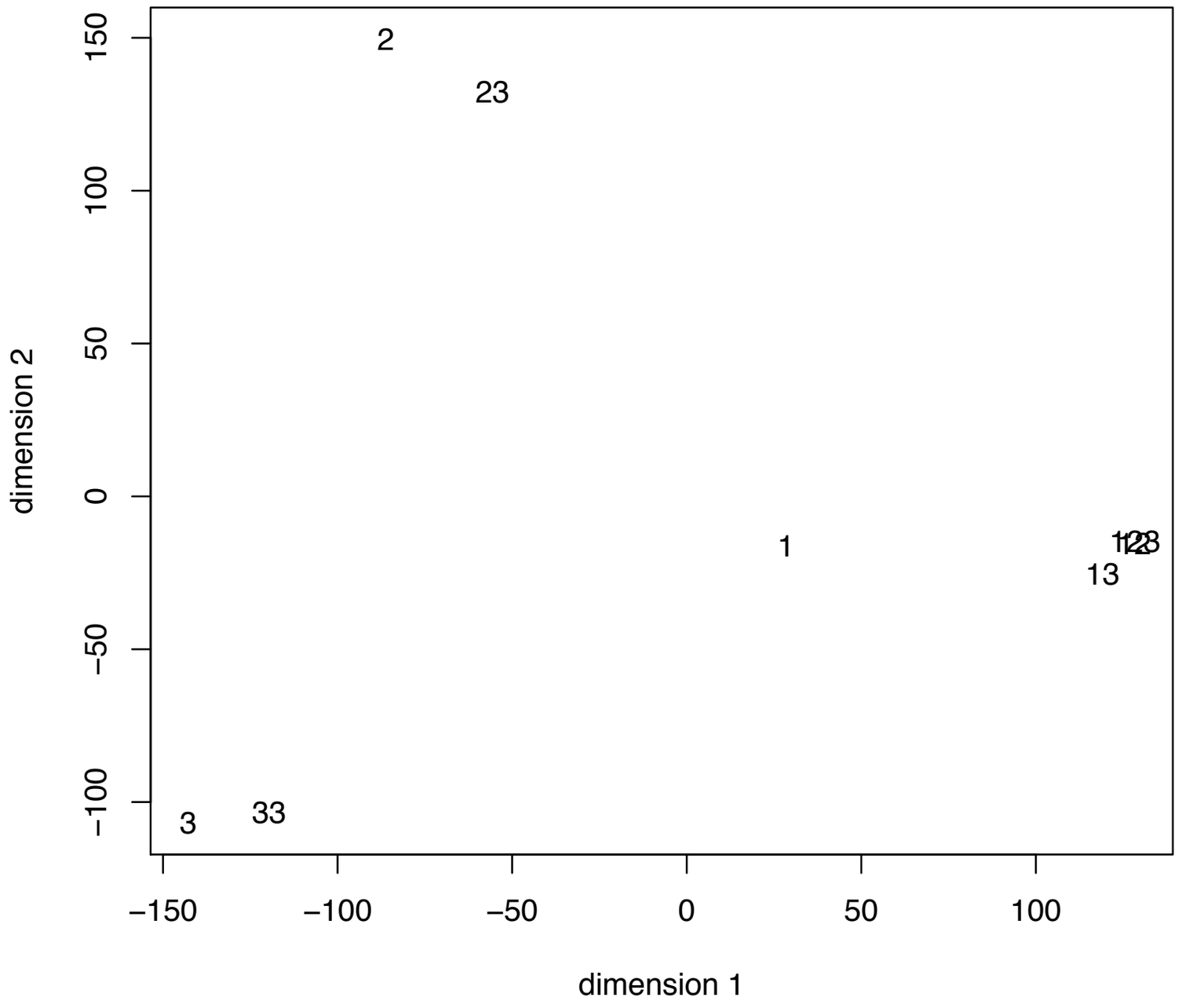
Semantic Map of Person Marking

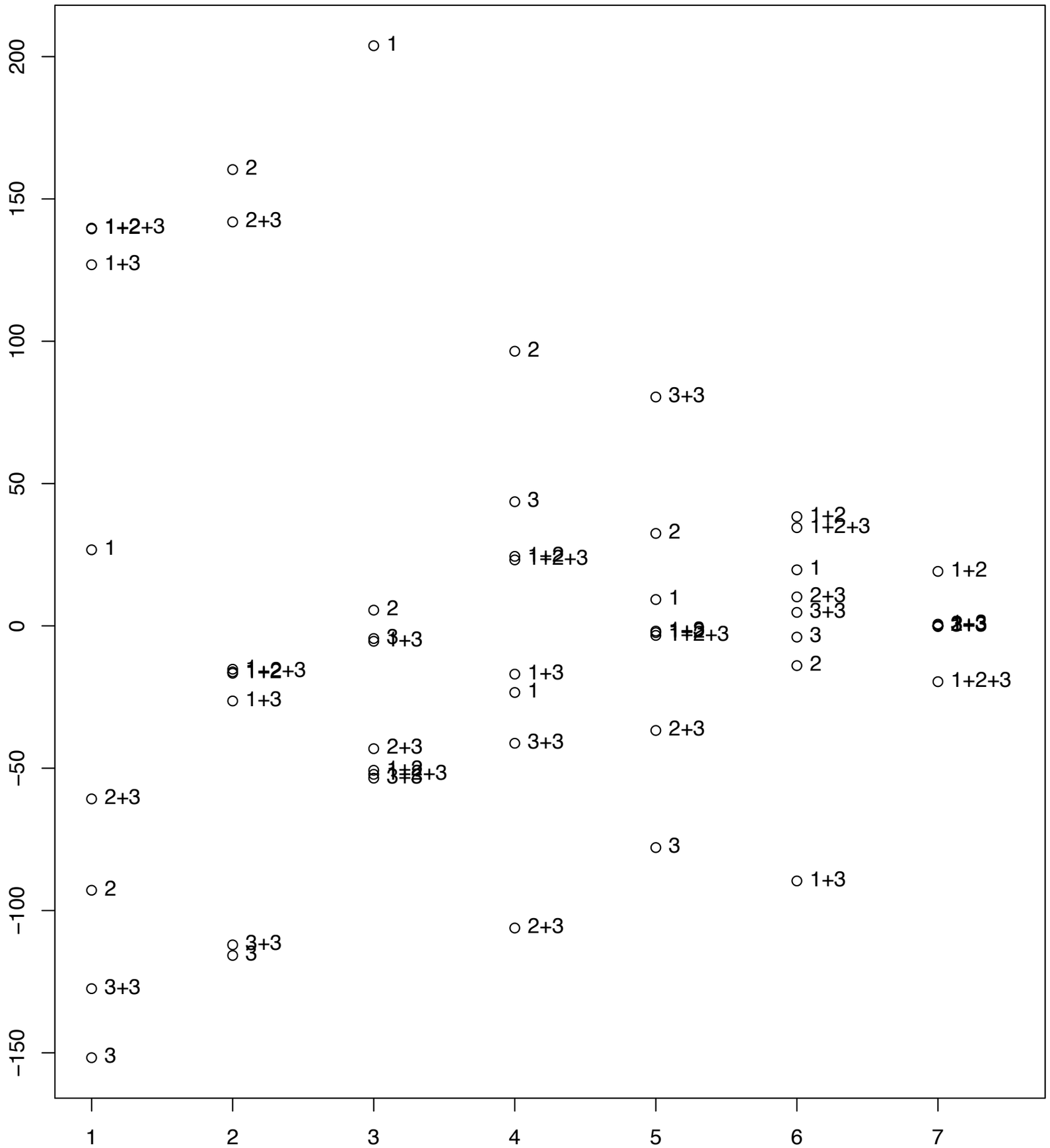


Frequencies included



Multidimensional scaling

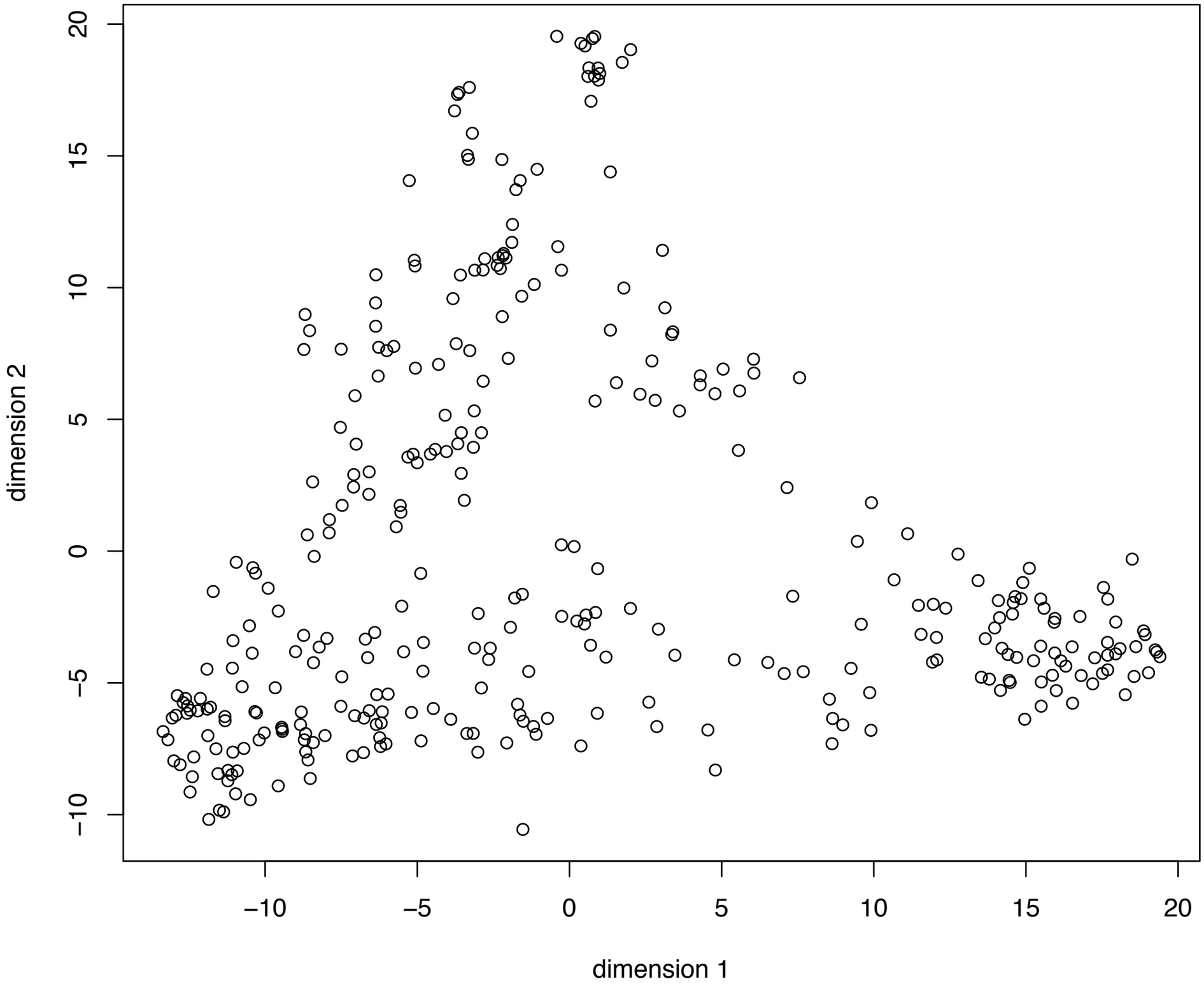


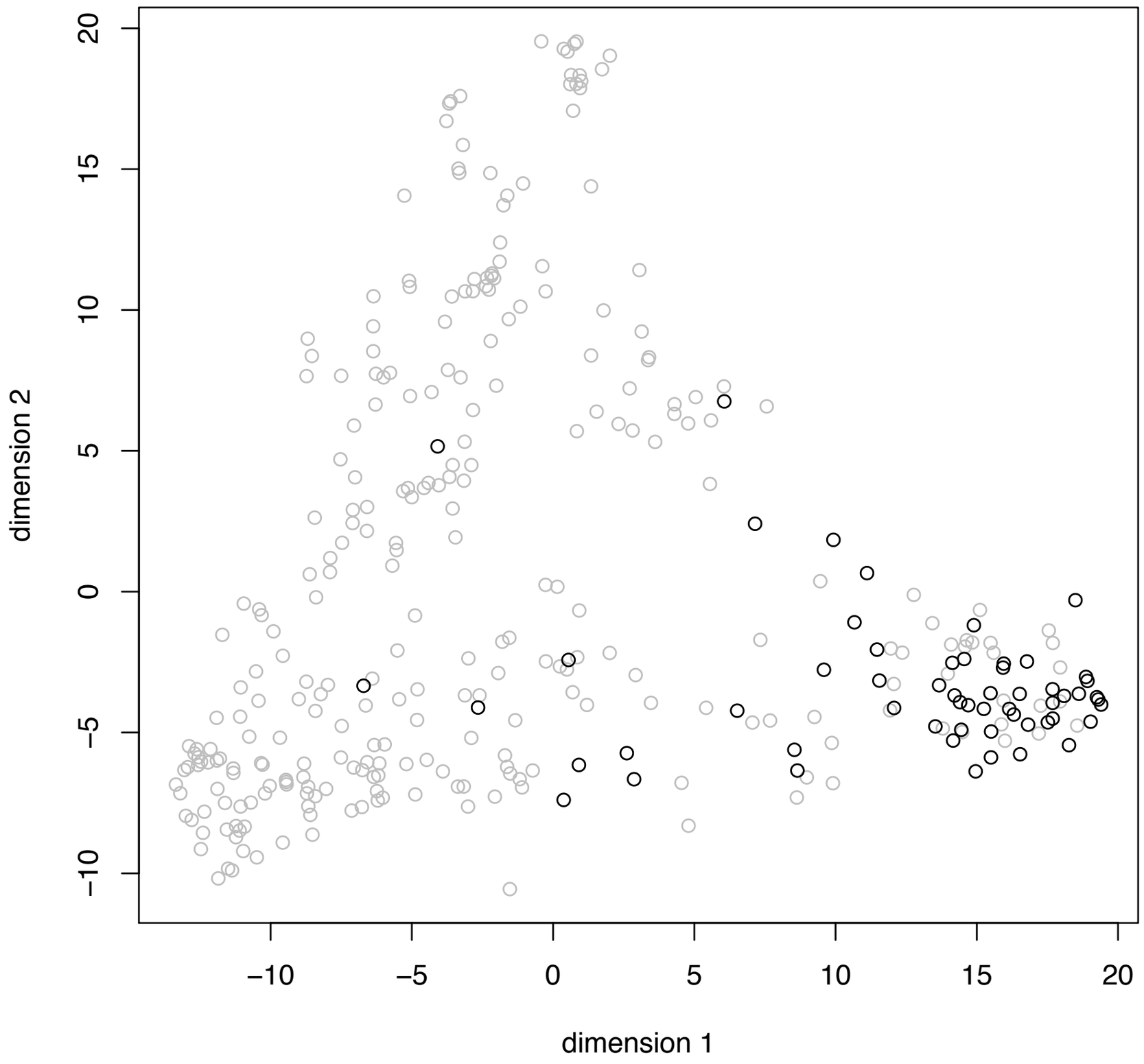

200
150
100
50
0
−50
−100
−150
1
2
3
4
5
6
7
dimensions

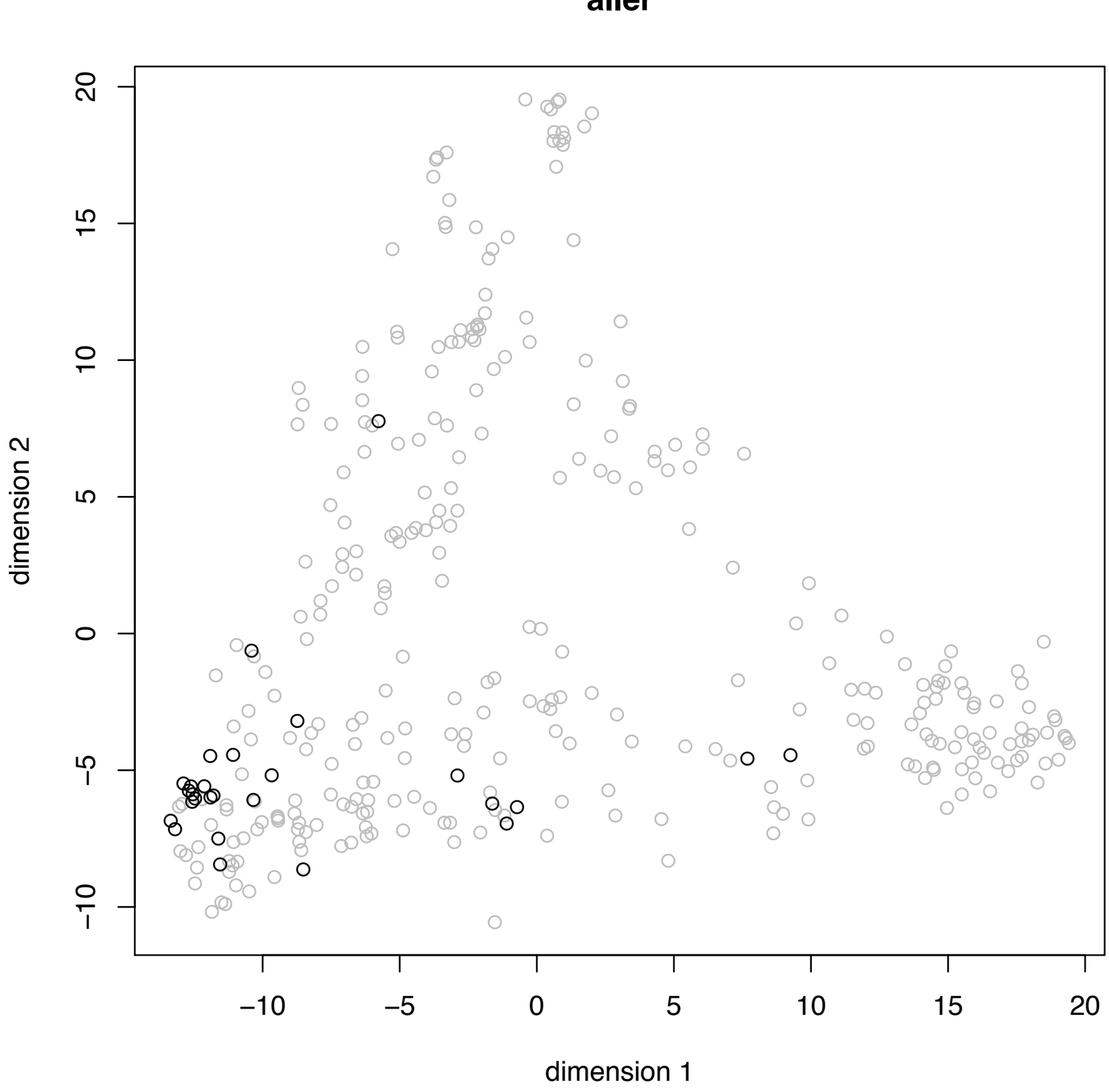
B. Wälchli's data on motion events

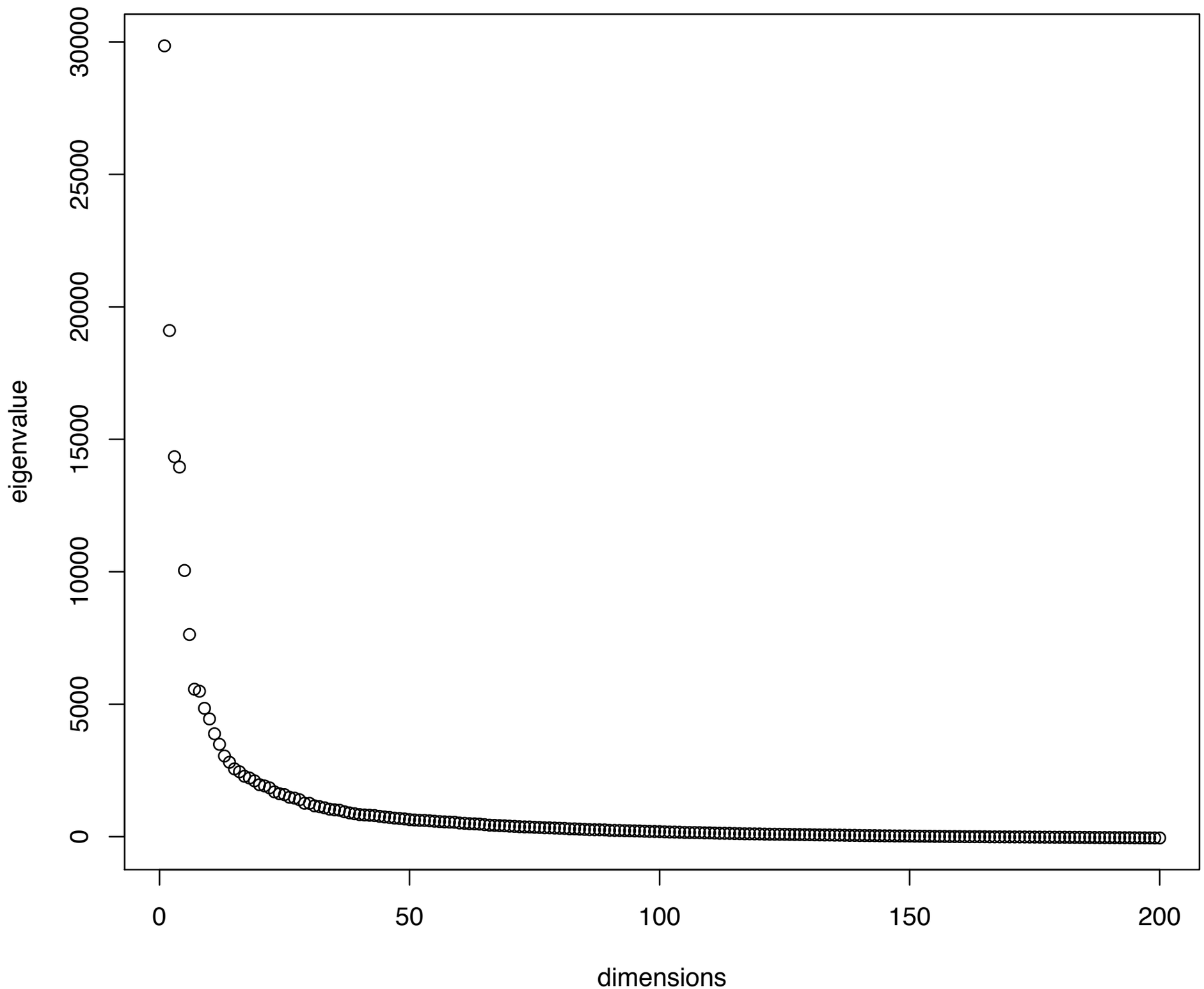
- 72 languages
- 335 clauses for each language from Bible
- clauses describing motion events
- here, only the lexical verb used is included

	MRD	LIT	ENG	FRE
1050	sams	eiti	go	aller
1070	sams	eiti	come	venir
1090	sams	eiti	come	venir
1104	lisems	kopti	come	sortir
1105	valgoms	zengti	descend	descendre
1114	–	–	come	se faire entendre
1120	vetjams	varyti	drive	pousser
1140	sams	eiti	come	se rendre
1160	jutams	eiti	walk	marcher


dimension 2
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20
dimension 1


venir
dimension 2
dimension 1
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20


aller
dimension 2
dimension 1
20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20


screeplot
eigenvalue
0
5000
10000
15000
20000
25000
30000
0
50
100
150
200
dimensions

4. Relationships between languages

- Universality: linguistic diversity interpreted as a-historical generalisations
- Contingency: linguistic diversity interpreted as result of historical processes

Using Typological Data for Genealogical Investigations

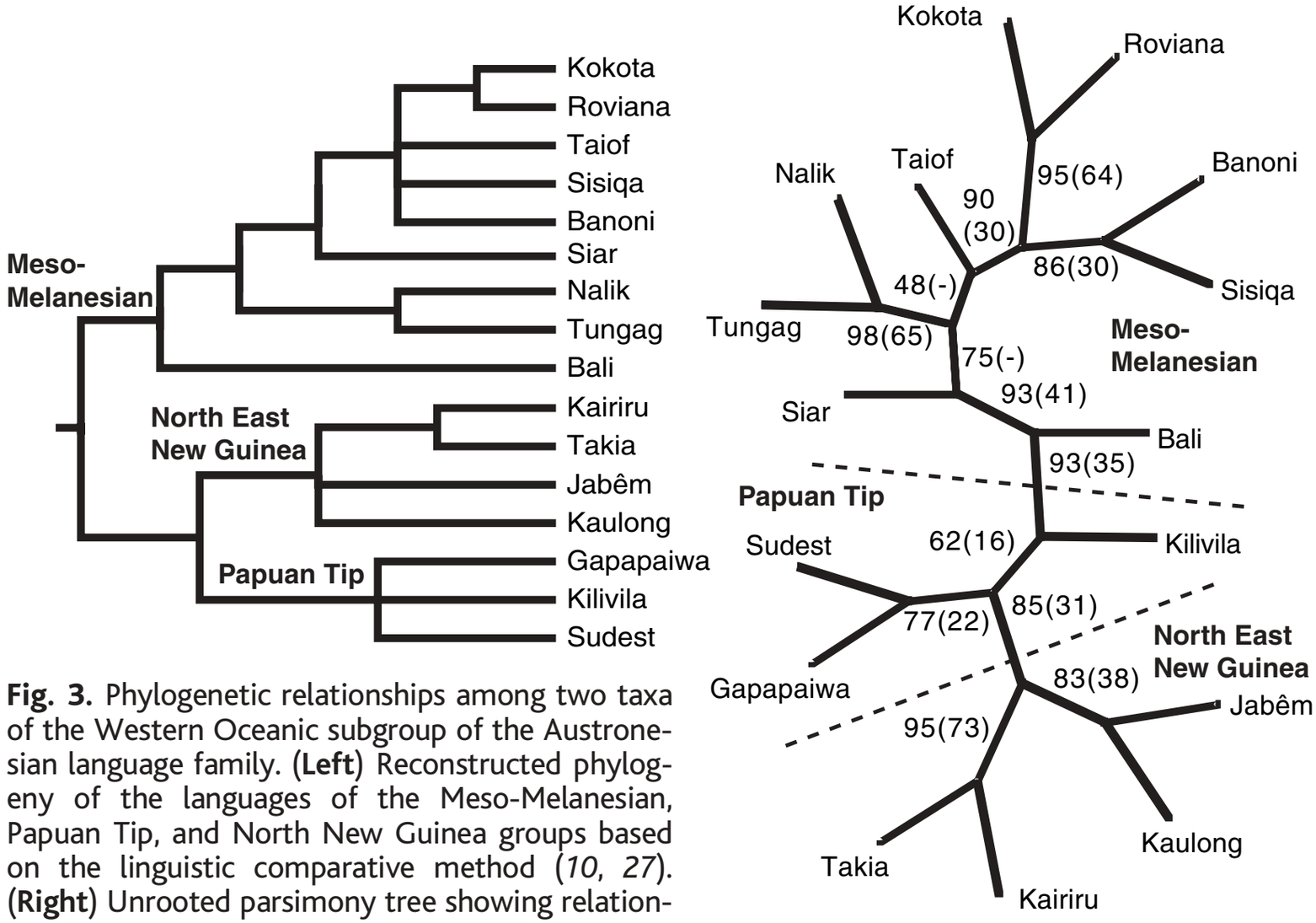
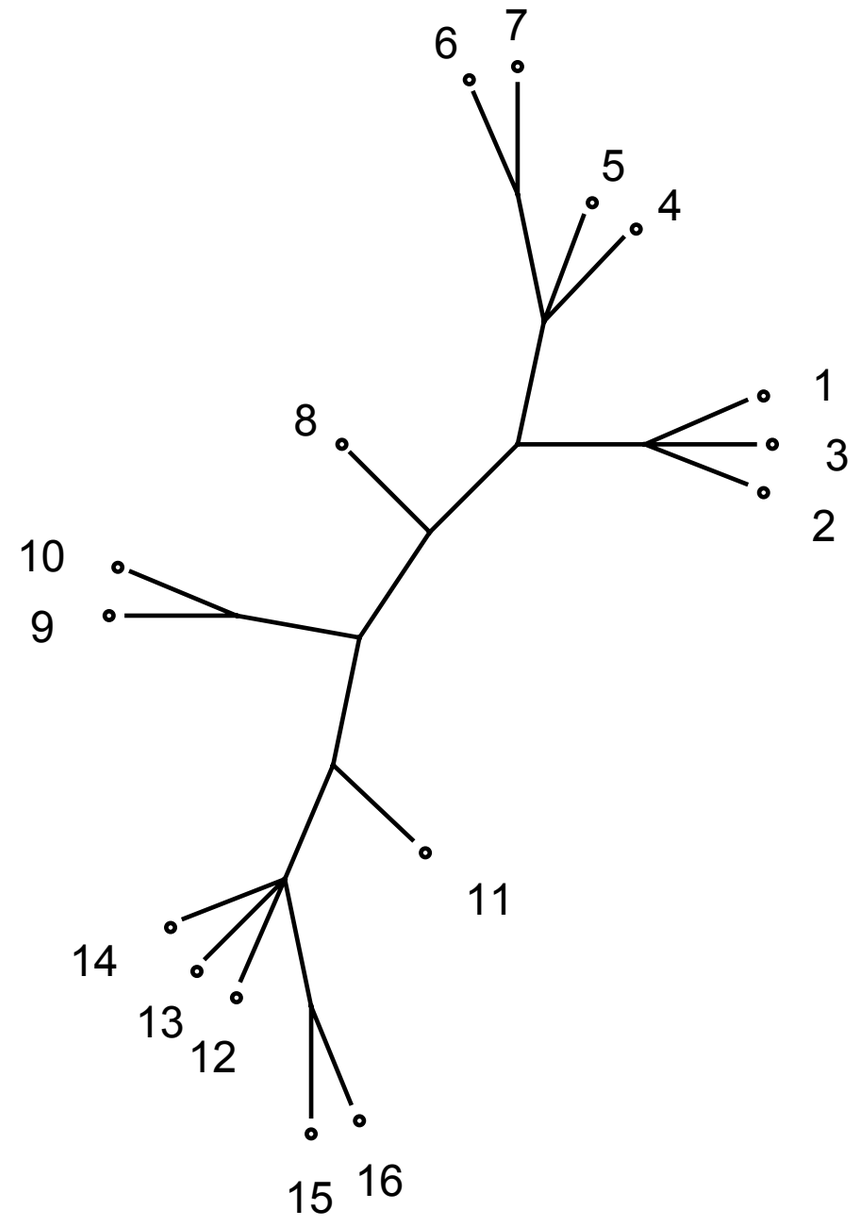


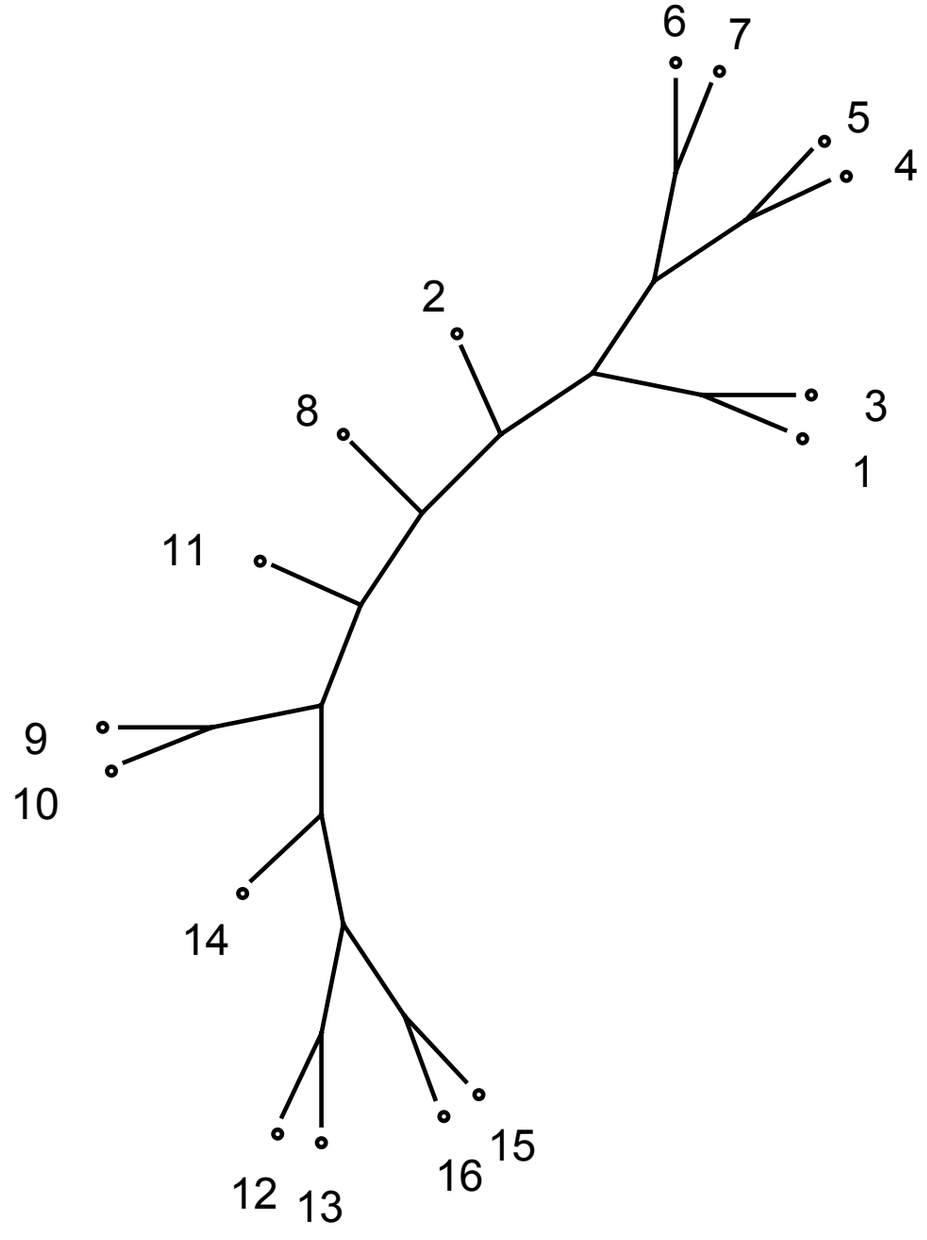
Fig. 3. Phylogenetic relationships among two taxa of the Western Oceanic subgroup of the Austronesian language family. (**Left**) Reconstructed phylogeny of the languages of the Meso-Melanesian, Papuan Tip, and North New Guinea groups based on the linguistic comparative method (*10*, *27*). (**Right**) Unrooted parsimony tree showing relationships among the Meso-Melanesian and Papuan Tip groups based on grammatical traits only (that is, discarding abundant lexical evidence) (the figure shows reweighted and raw bootstrap values). The two trees show a high degree of concordance, with monophyly in both major taxa and the similar geographical structuring of within-taxon diversity.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science* 309: 2072-2075.

Traditional Tree



Dunn *et al.* tree based on typological data

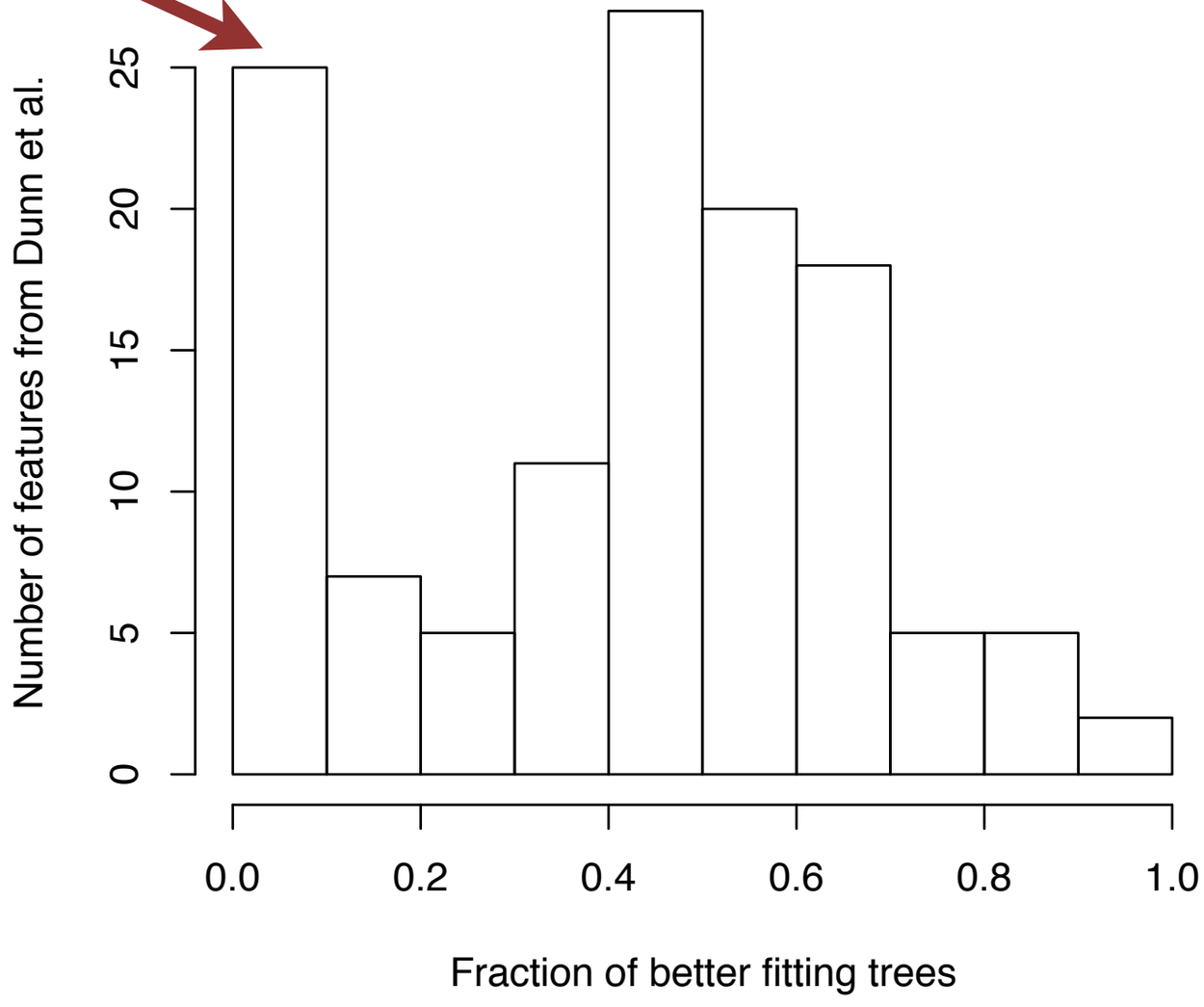


Measuring typological stability

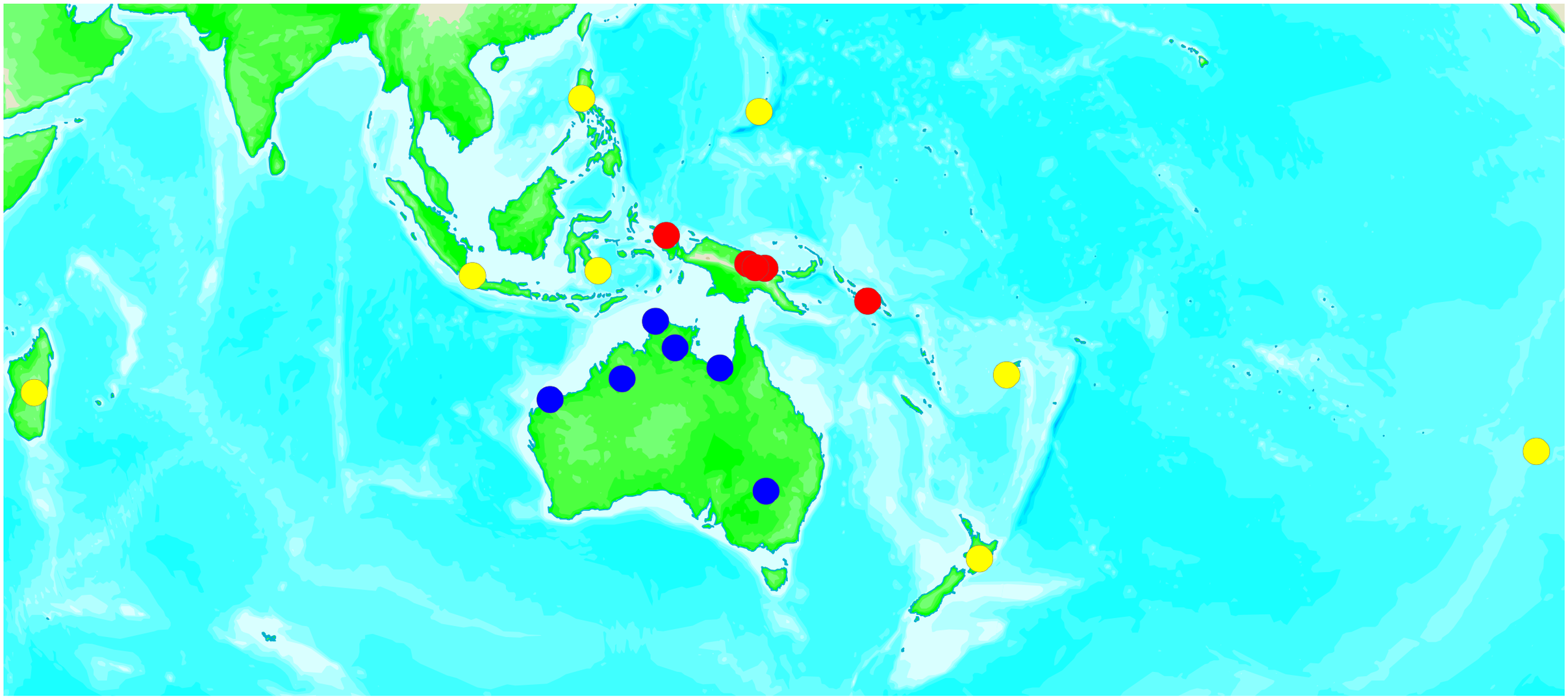
- Given an accepted tree, with many languages sampled from this tree
  - how good does a typological feature predict this tree
  - Energy-based measurement of fit between a dataset and a tree (work by Mihai Albu)
  - Take a large set of random trees, and determine how good the "real" tree fits

Distribution of fits of all 125 features

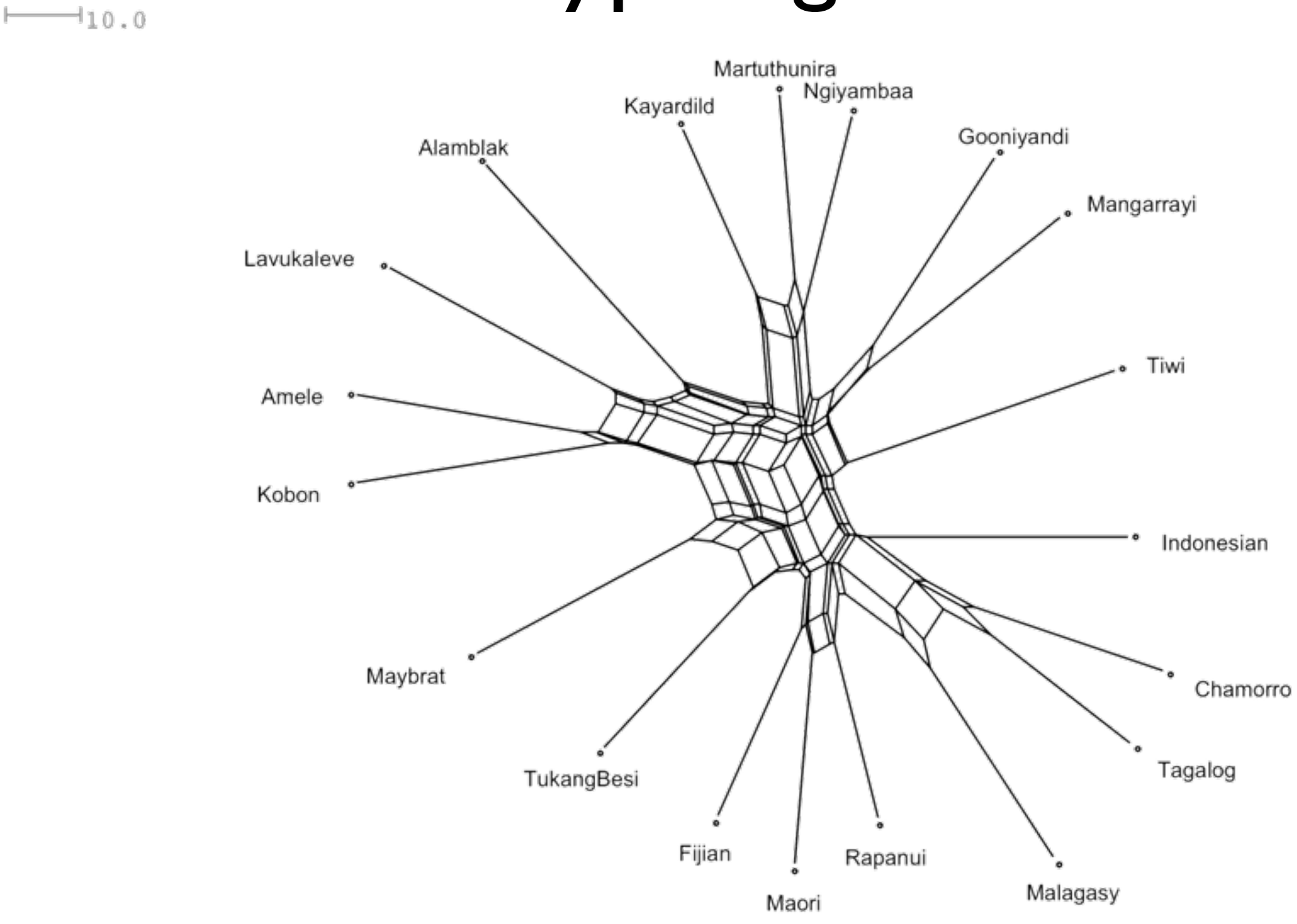
(Too) many good fits!



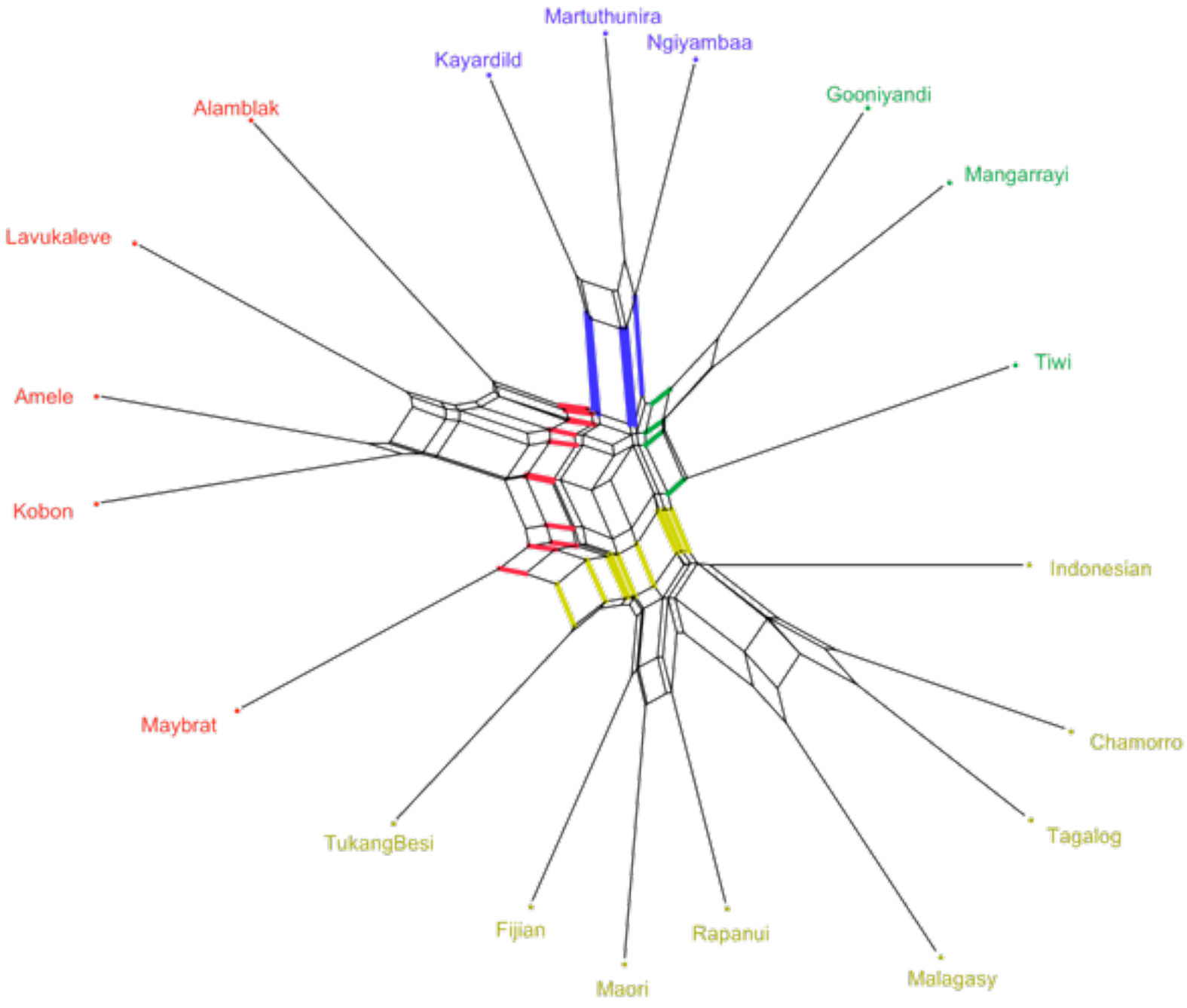
Oceania



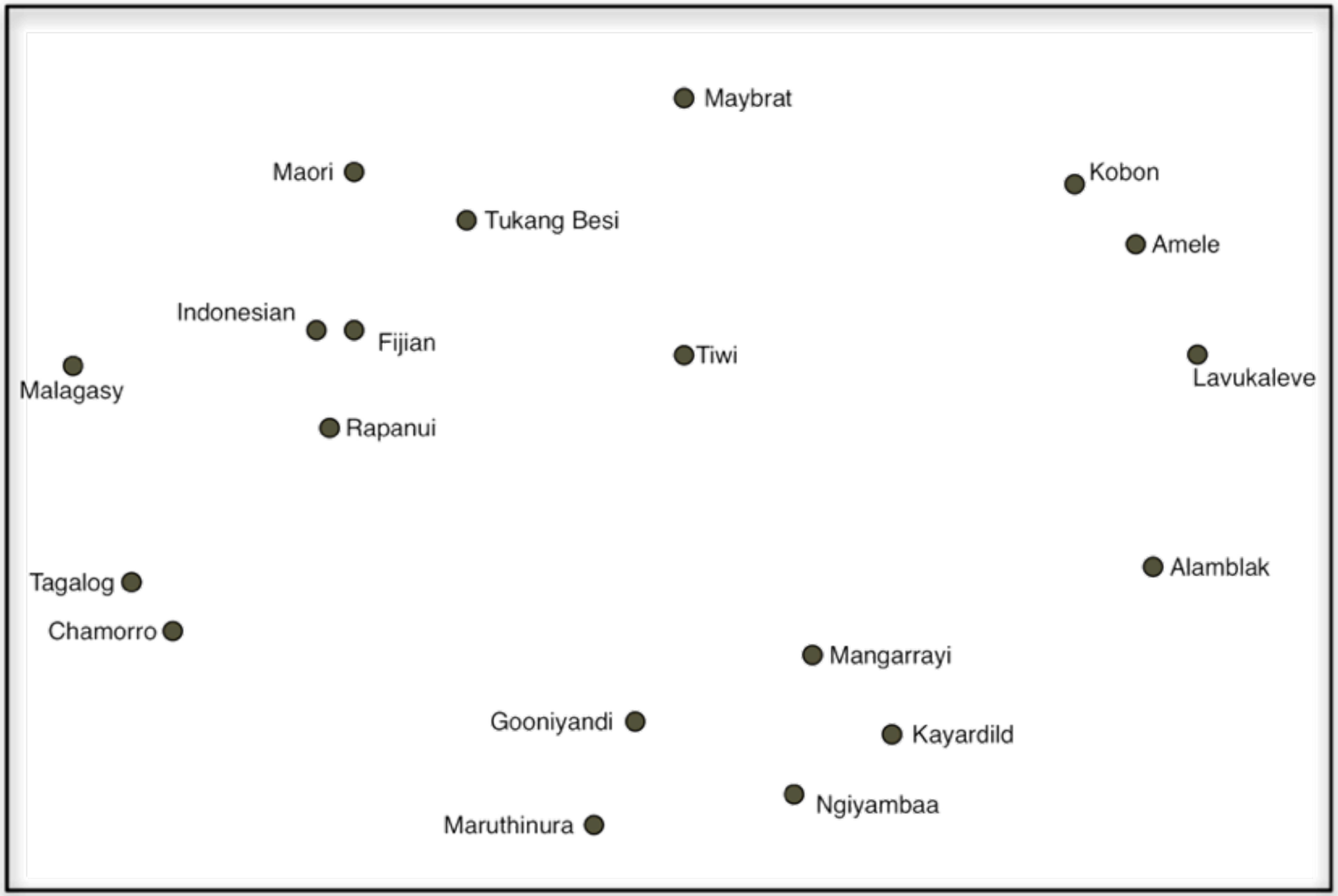
NNet of typological distances



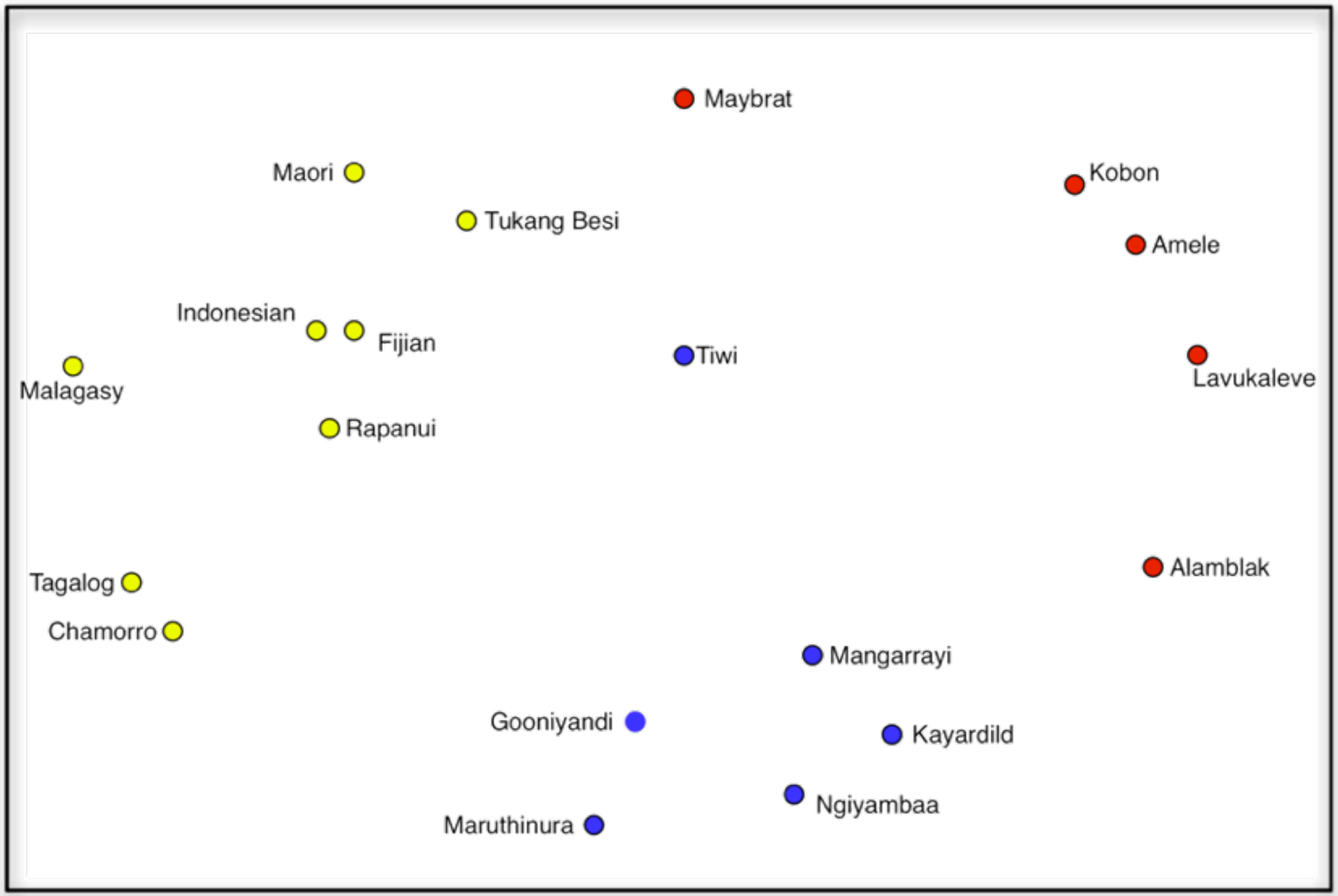
NNet of typological distances



MDS of typological distances



MDS of typological distances



Maybrat

Kobon

Tukang Besi

Amele

Tiwi

Lavukaleve

Alamblak

Mangarrayi

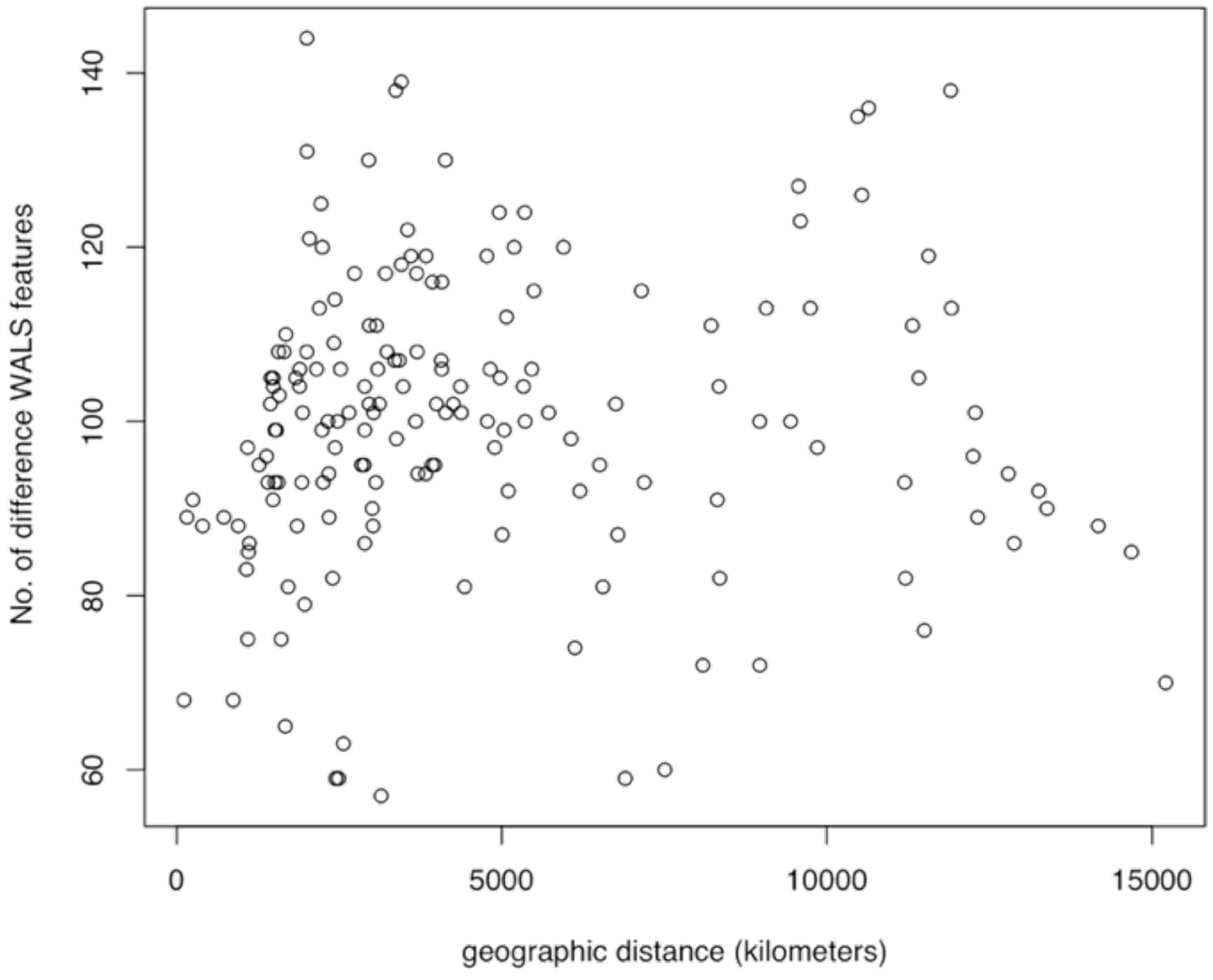
Gooniyandi

Kayardild

Ngiyambaa

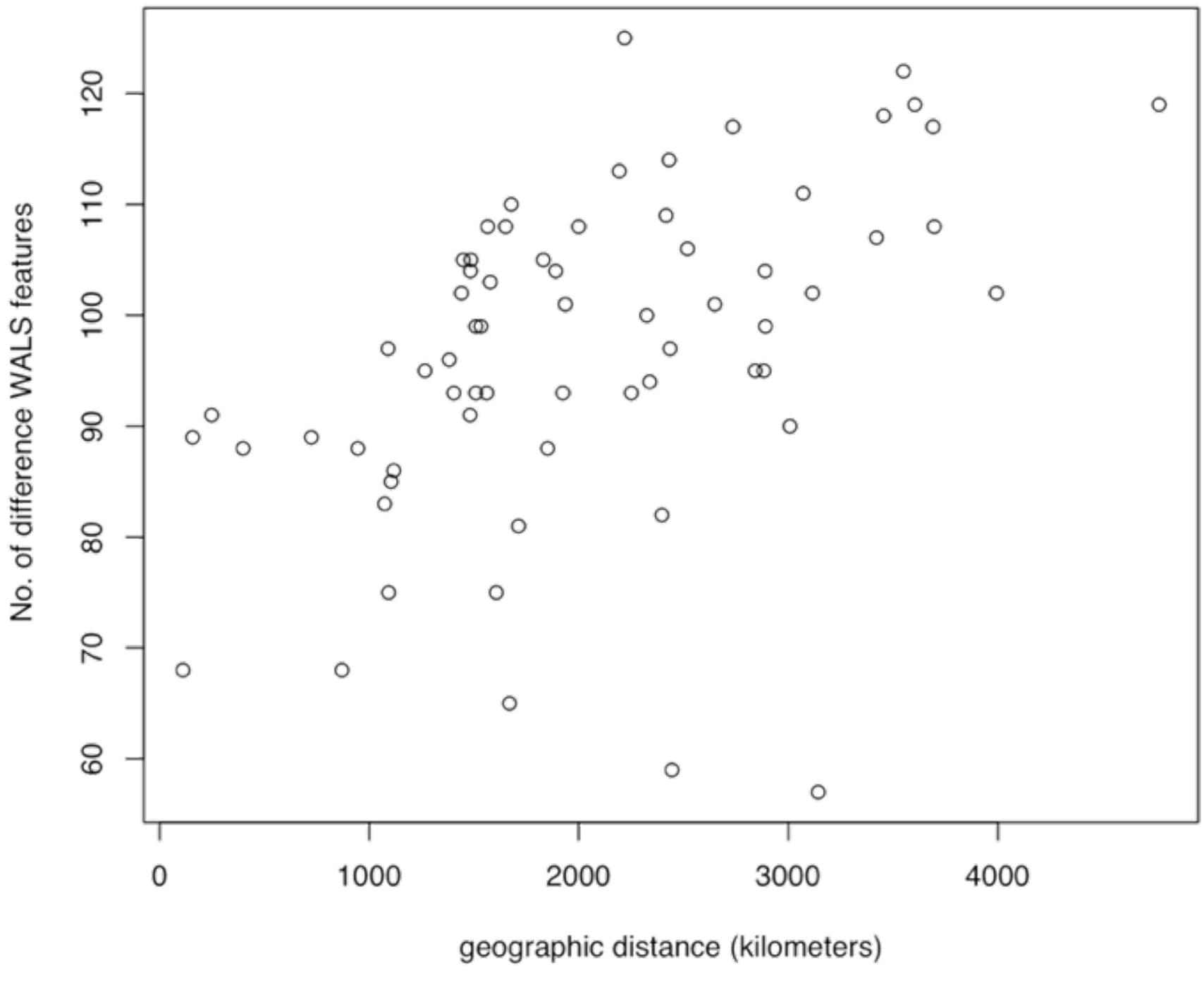
Maruthinura

Typology/geography correlation



Mantel test p = .349

Correlation for selection only



Mantel test p = .001

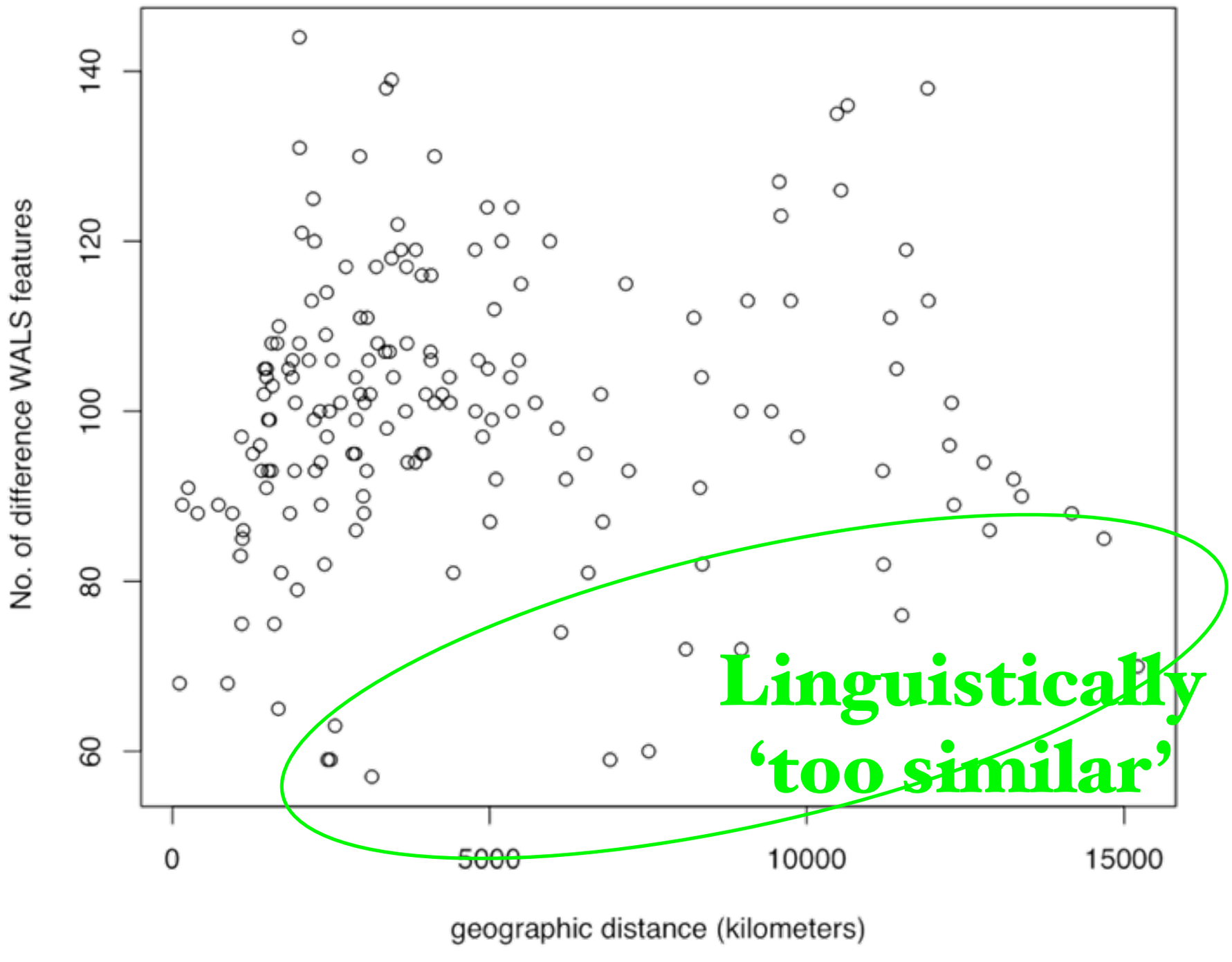
When does correlation improve?

	Pearson's *r*
Nothing removed	.035

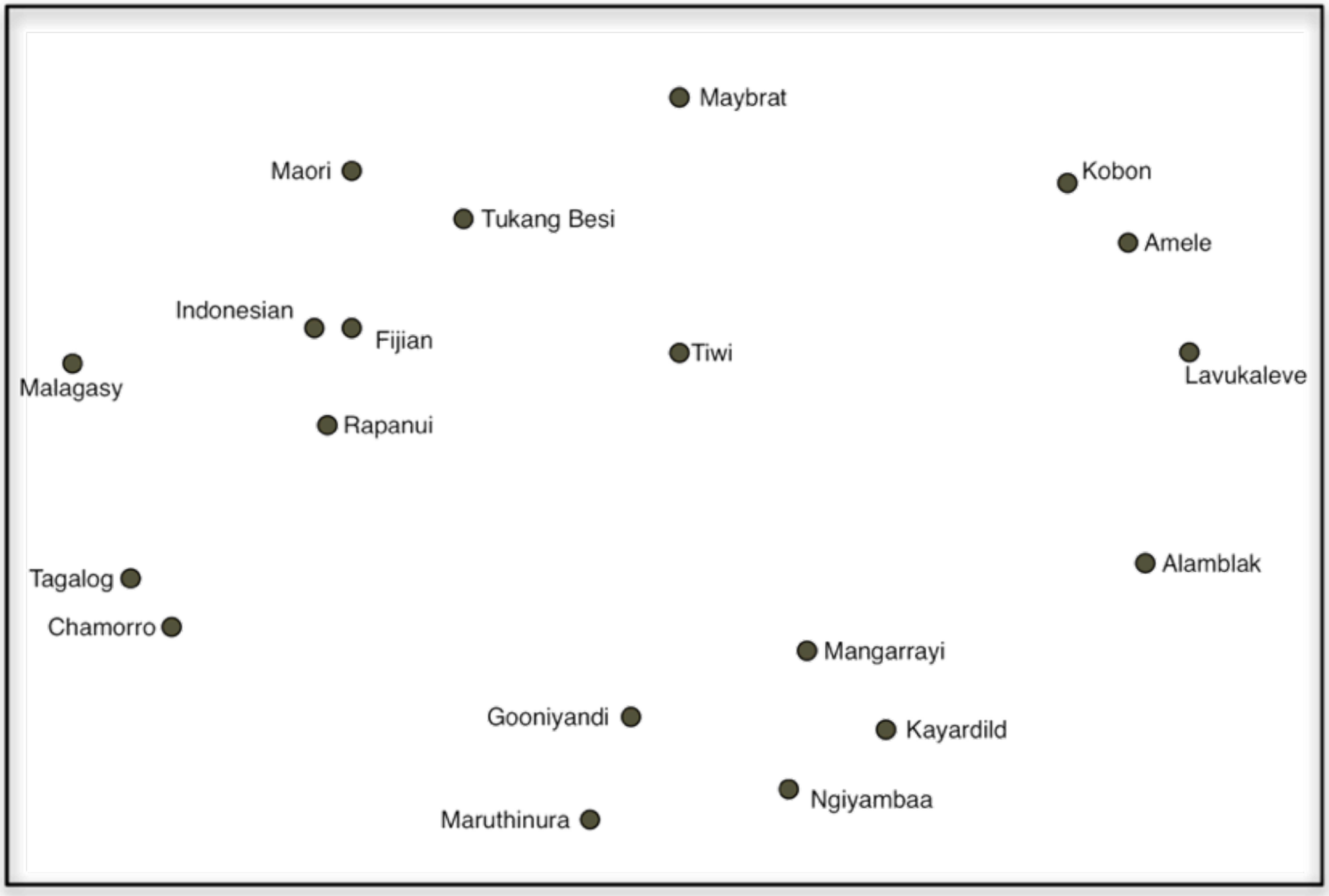
When does correlation improve?

	Pearson's *r*
Nothing removed	.035
Rapanui	.186
Chamorro	.086
Indonesian	.076
Fijian	.073
Tagalog	.071
Maori	.062
Tukang Besi	.048

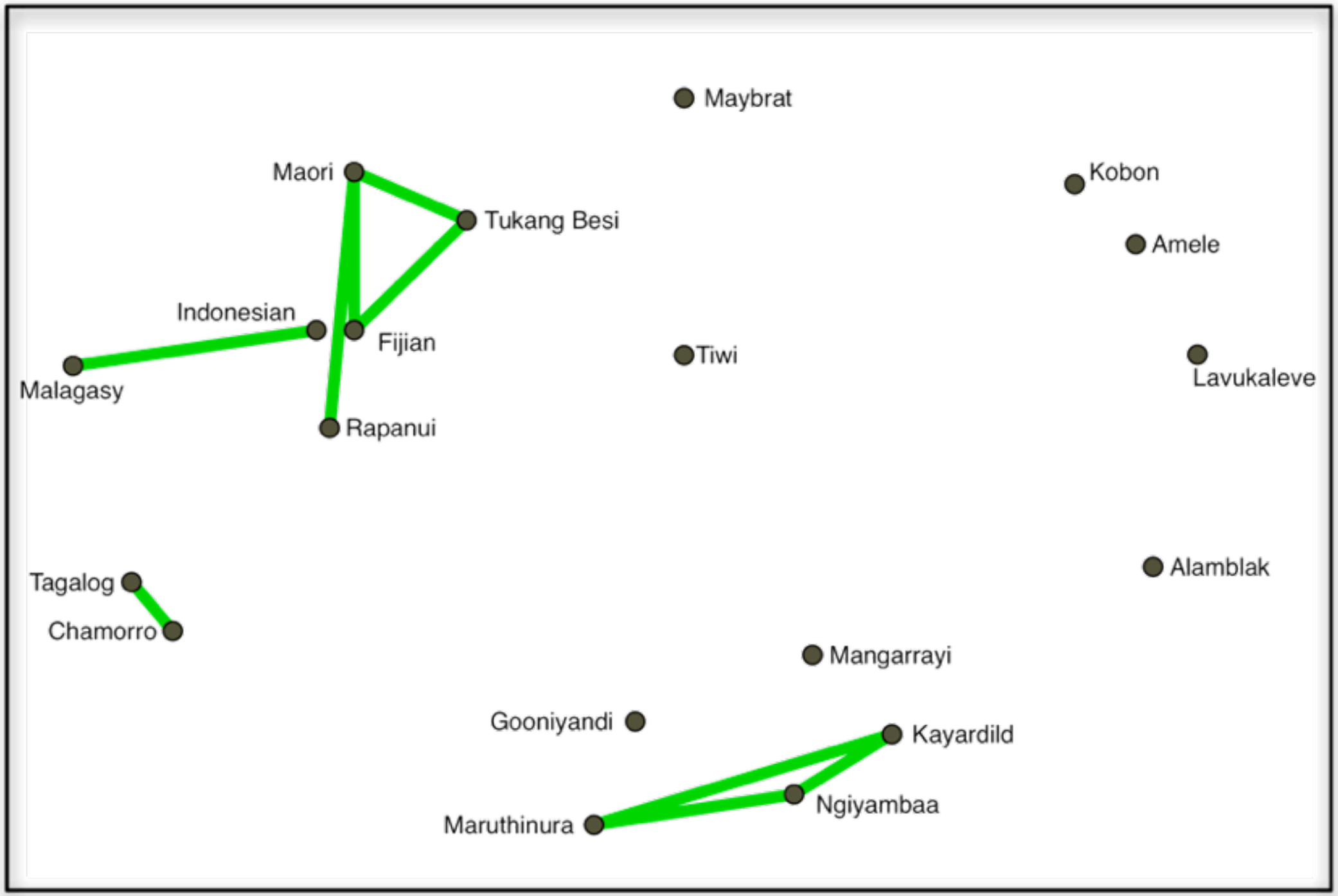
Investigation typology/geography relation



Linguistically ‘too similar’



Linguistically ‘too similar’



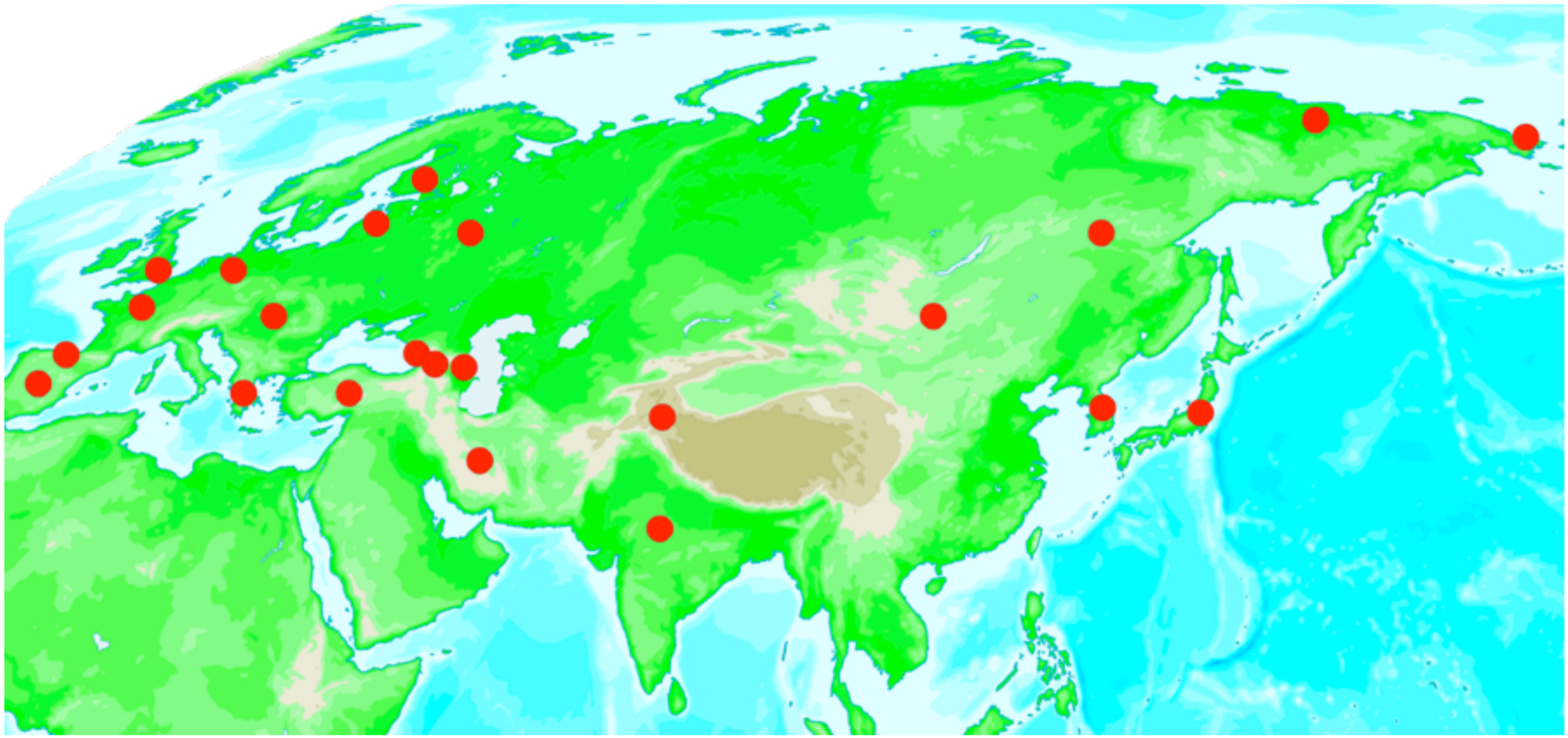
Summary

- Typology is correlated to genealogy
- but: typology is also correlated to geography
- When removing the (genealogically related) Austronesian languages, the typology/ geography correlation improves
- The language-pairs that are typologically more similar than expected from geography are genealogically related

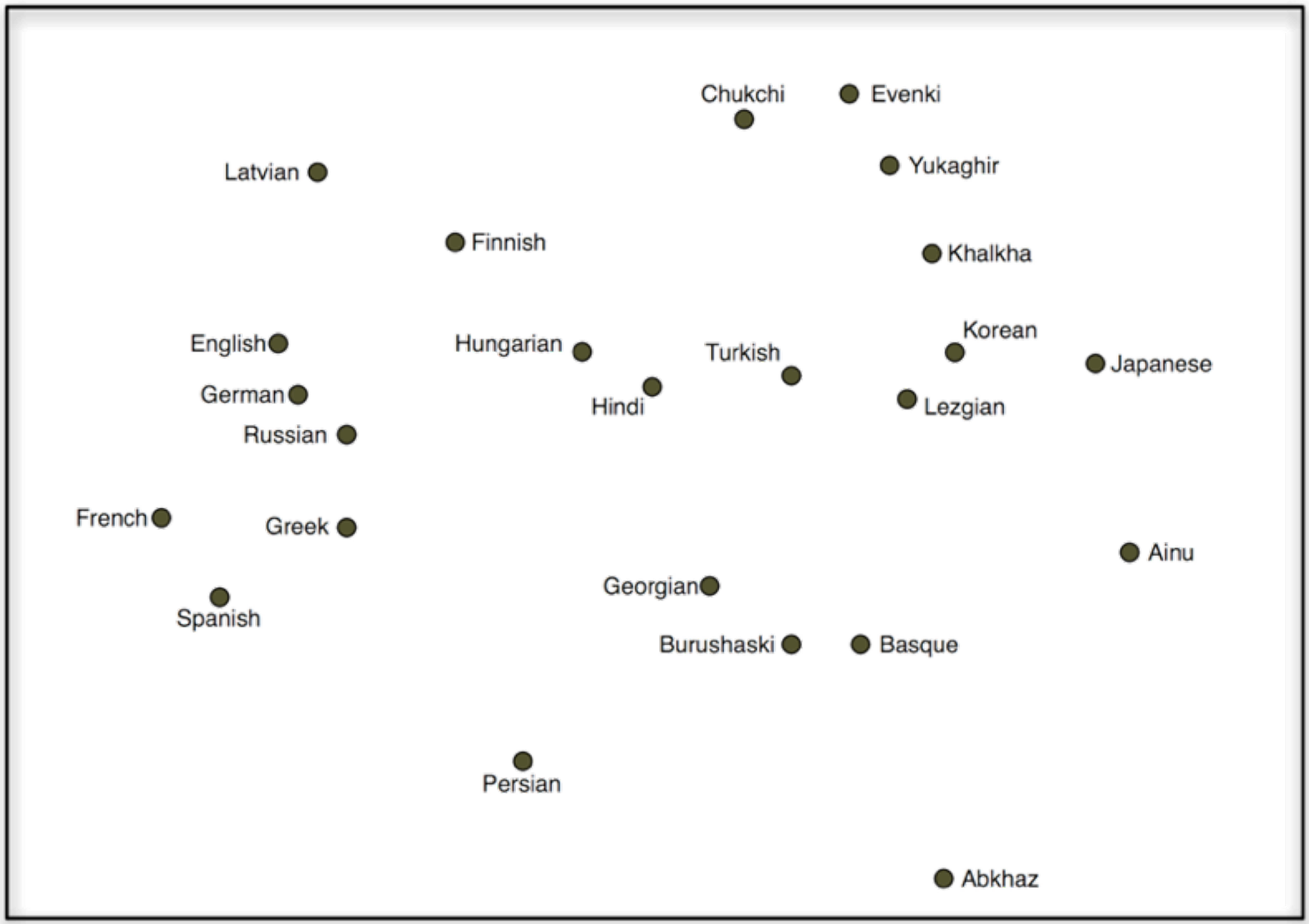
Towards an interpretation

- In longterm static (areal) interaction typological features diffuse individually, leading to regular geographical clines
- In relatively recent (genealogical) spread bundles of features ‘move’ together, leading to stronger similarities as expected from geography

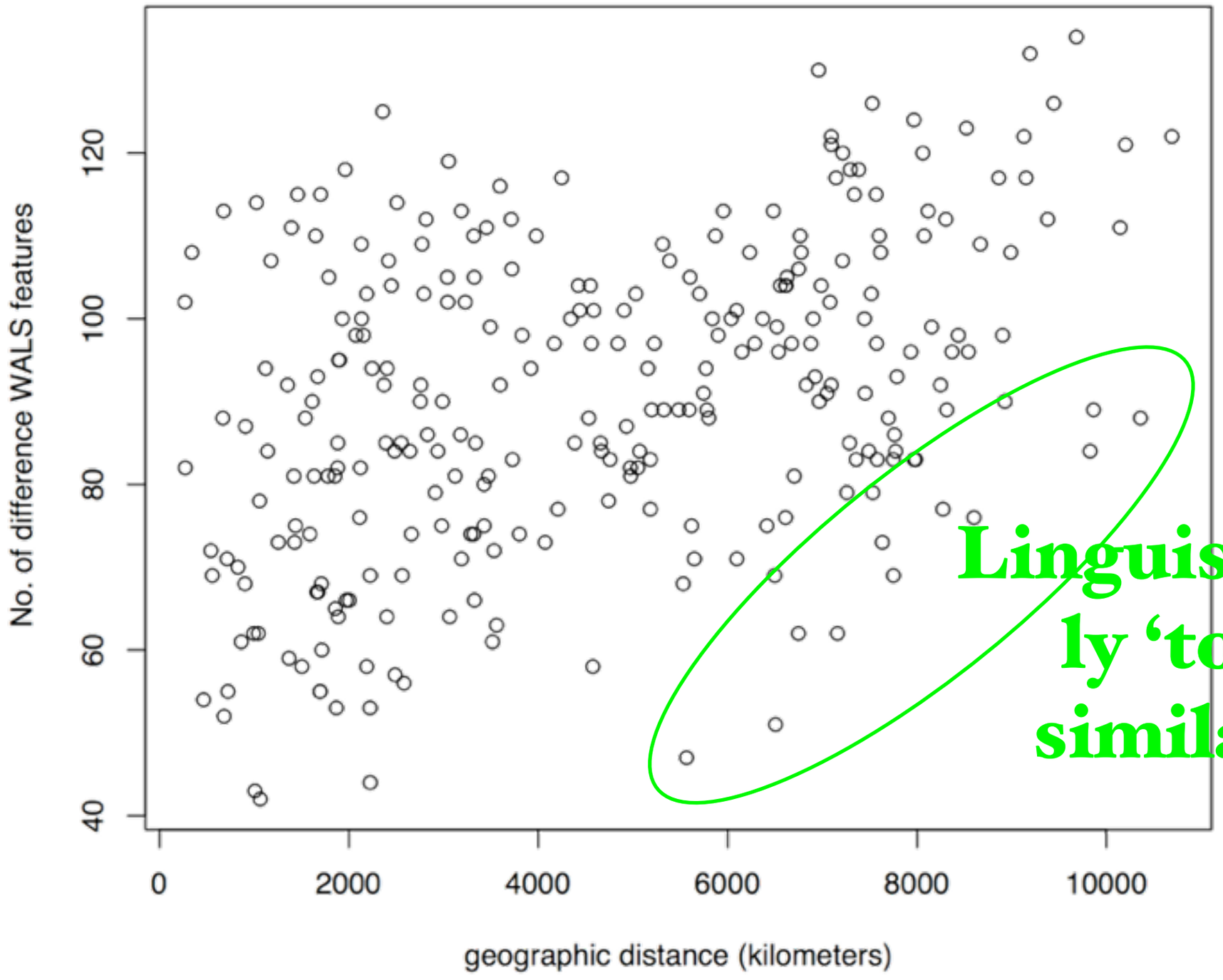
Eurasia



MDS of typological distances

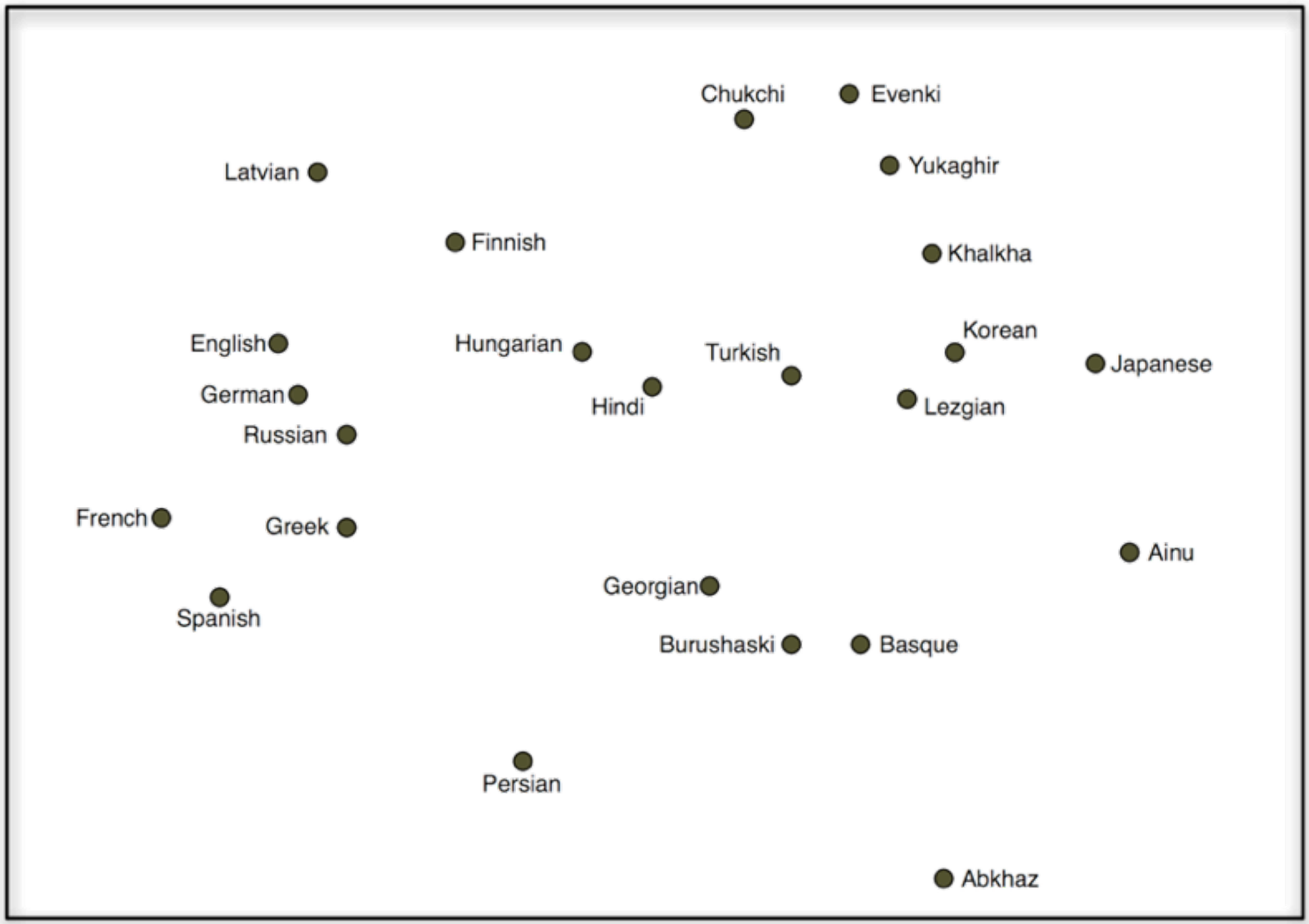


Typology/geography correlation

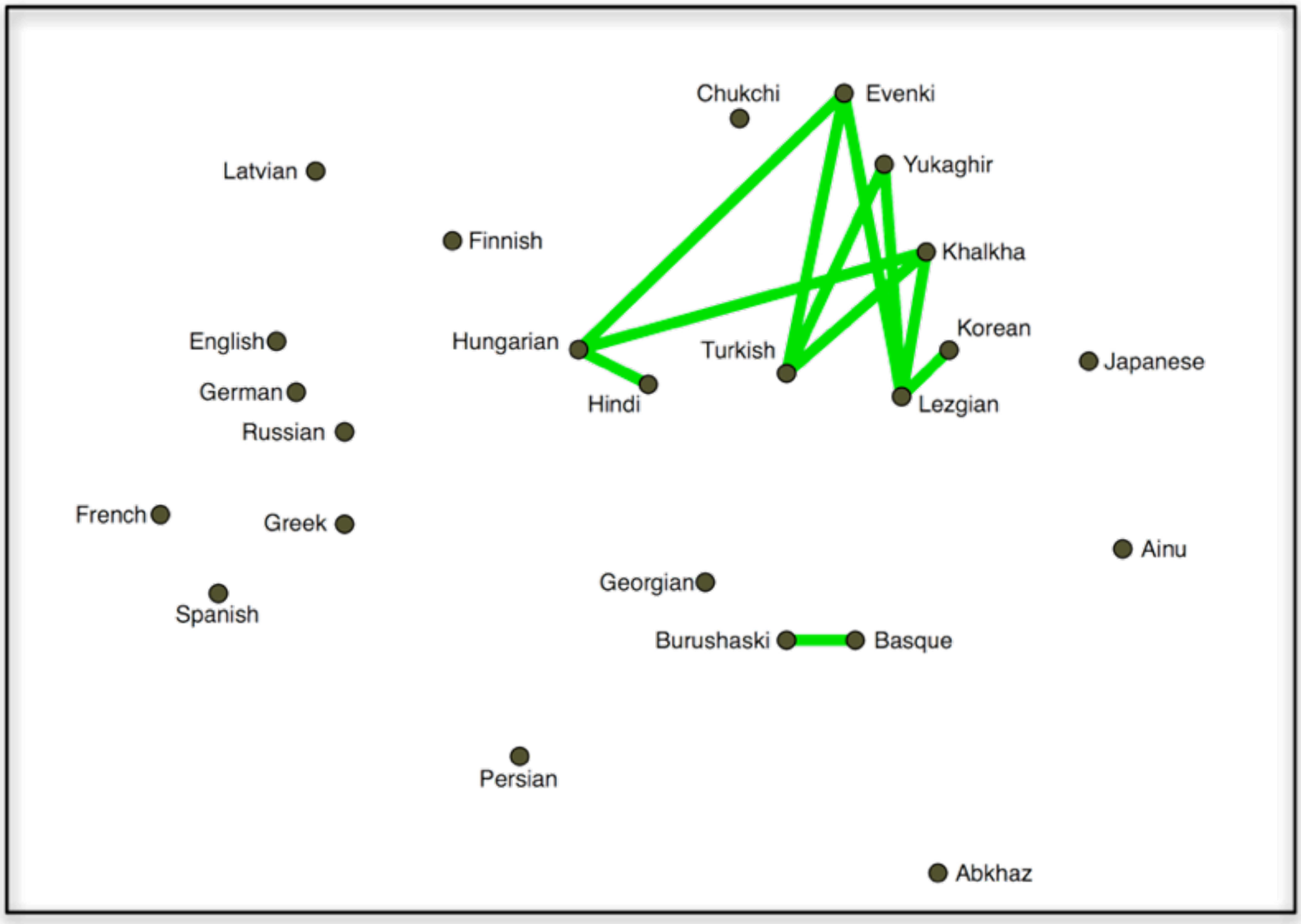


Mantel test p = .001

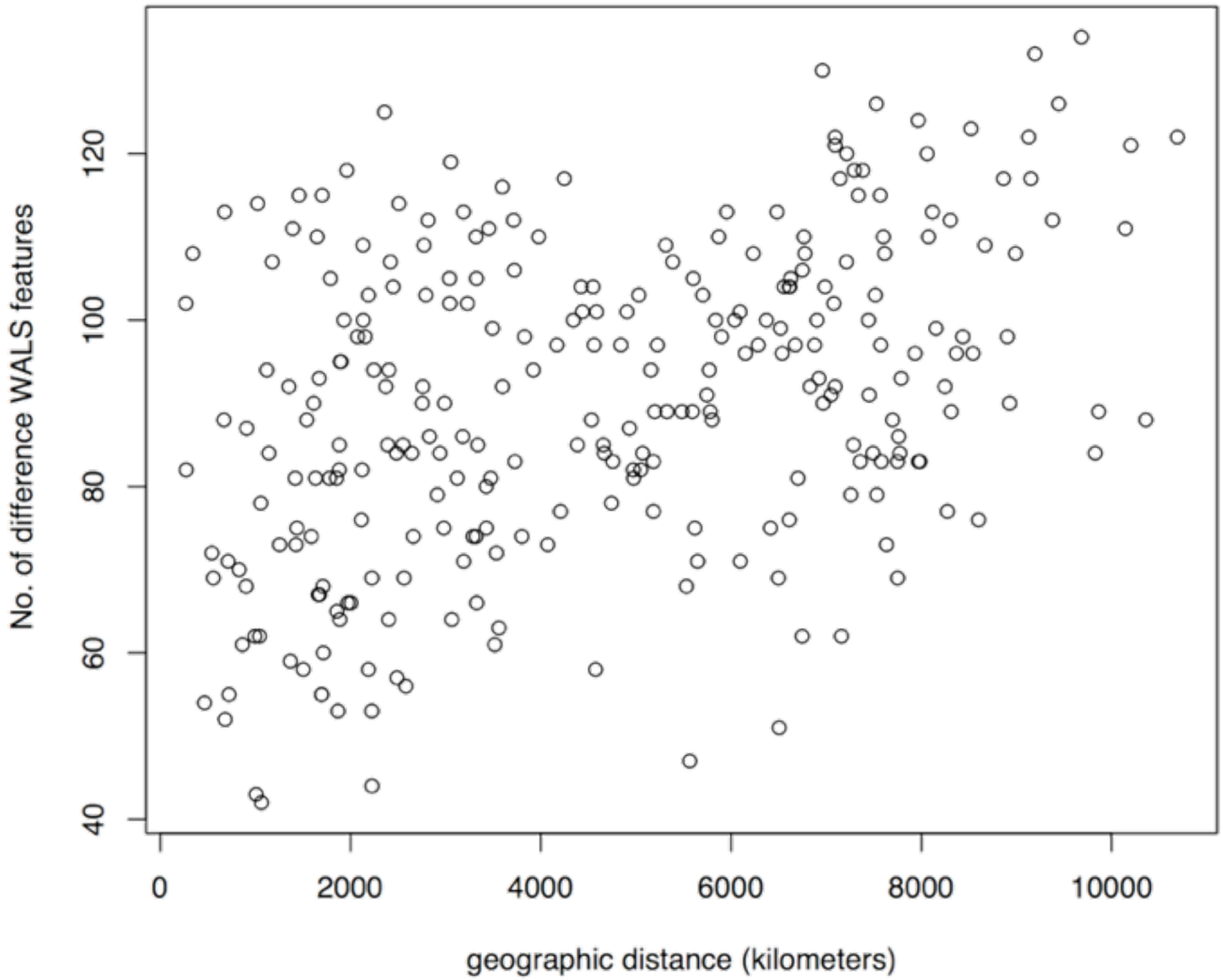
MDS of typological distances



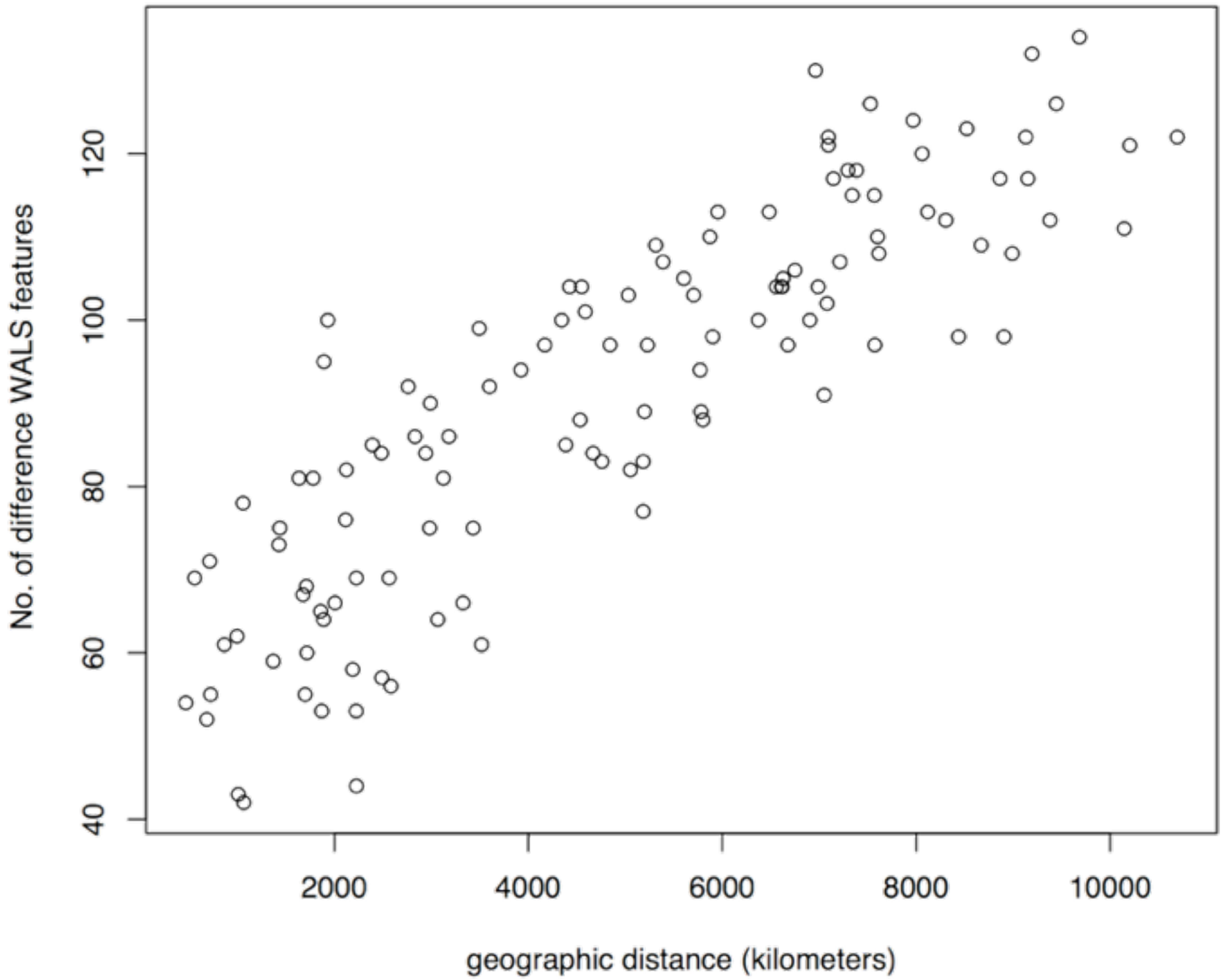
MDS of typological distances



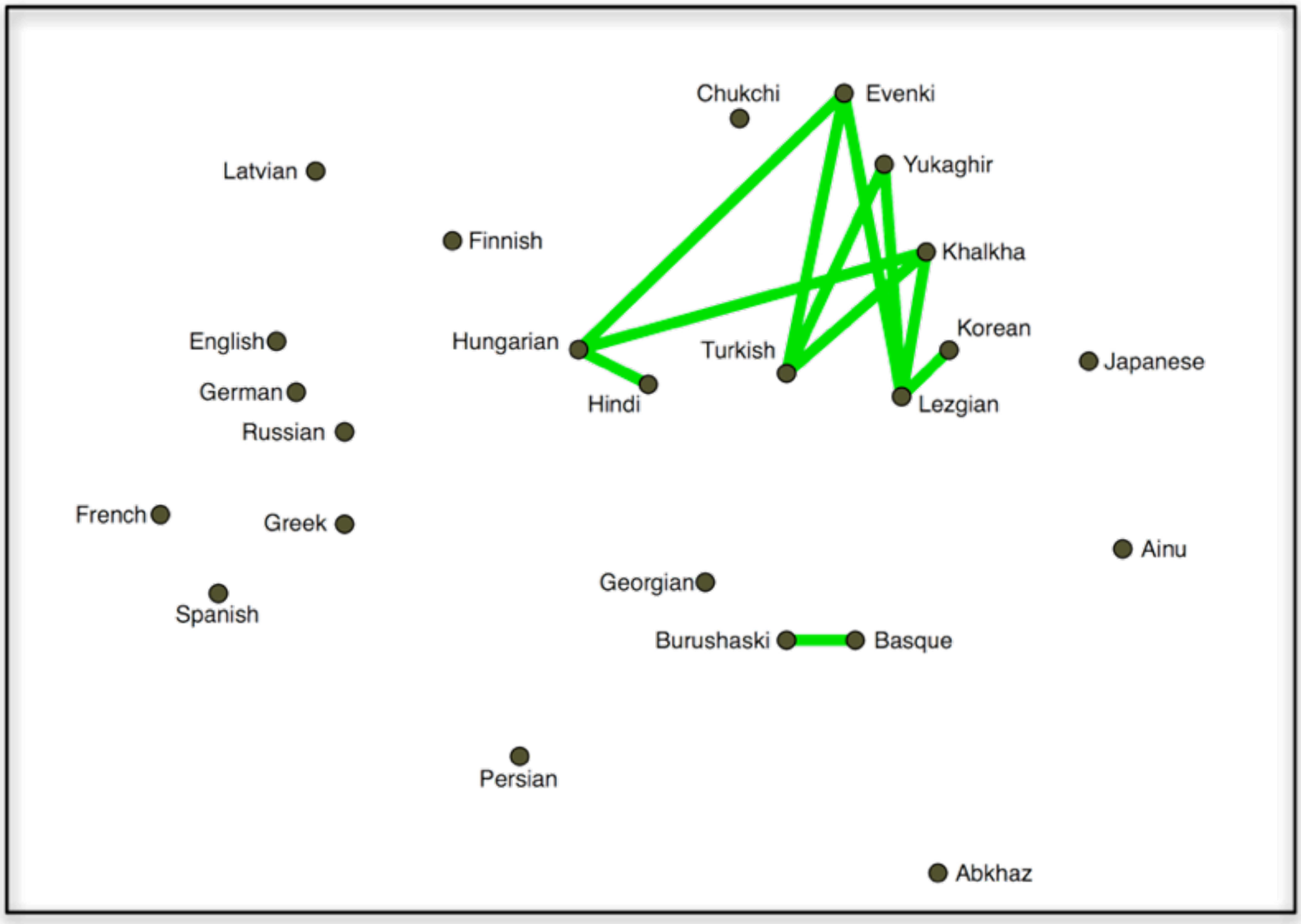
Remove ‘worst-fitting’ languages



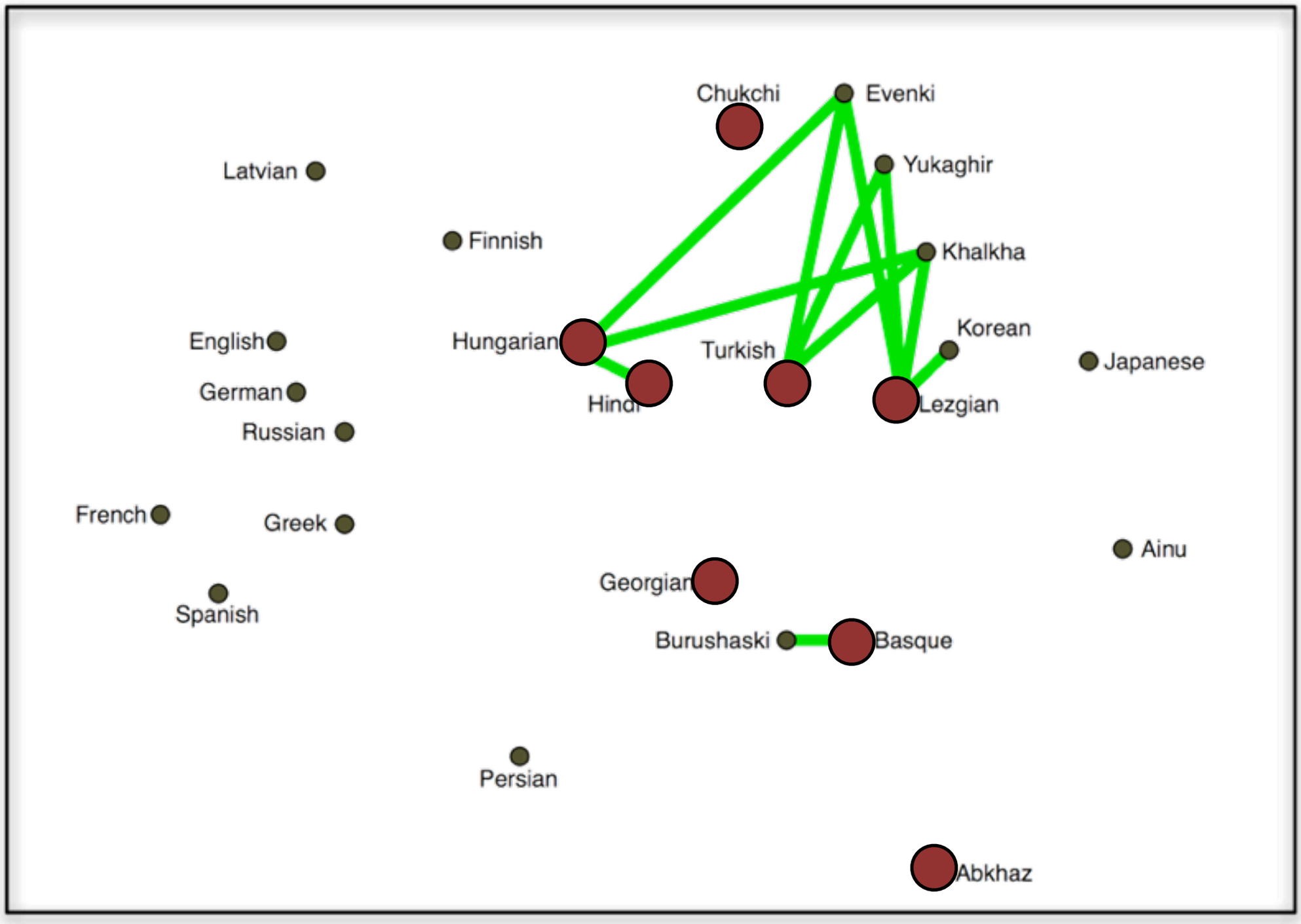
Remove ‘worst-fitting’ languages



MDS of typological distances



MDS of typological distances



Some interpretation

- Turkish and Hungarian are cases of relatively recent movement of whole languages
- But Lezgian (probably) not
- Link Hindi-Hungarian is unclear, and Burushaski-Basque is too cranky a speculation
- Chukchi, Georgian, Abkhaz simply unrelated, both genealogical and areal