Quantitative approaches to linguistic similarity

Michael Cysouw

Work-in-Progress, 9 November 2004

Part 1

Distribution of rare characteristics

- Using the WALS-data to approach some perennial questions:
- Are there languages that have many typologically rare characteristics?
- Are there regions that show a relatively high density of rare features?
- Does rarity cluster?

Rarity Index R_i

n = number of feature values

f_i = frequency of feature value i

f_{tot} = total number of languages included

For $f_i /f_{tot} < 1/n$:

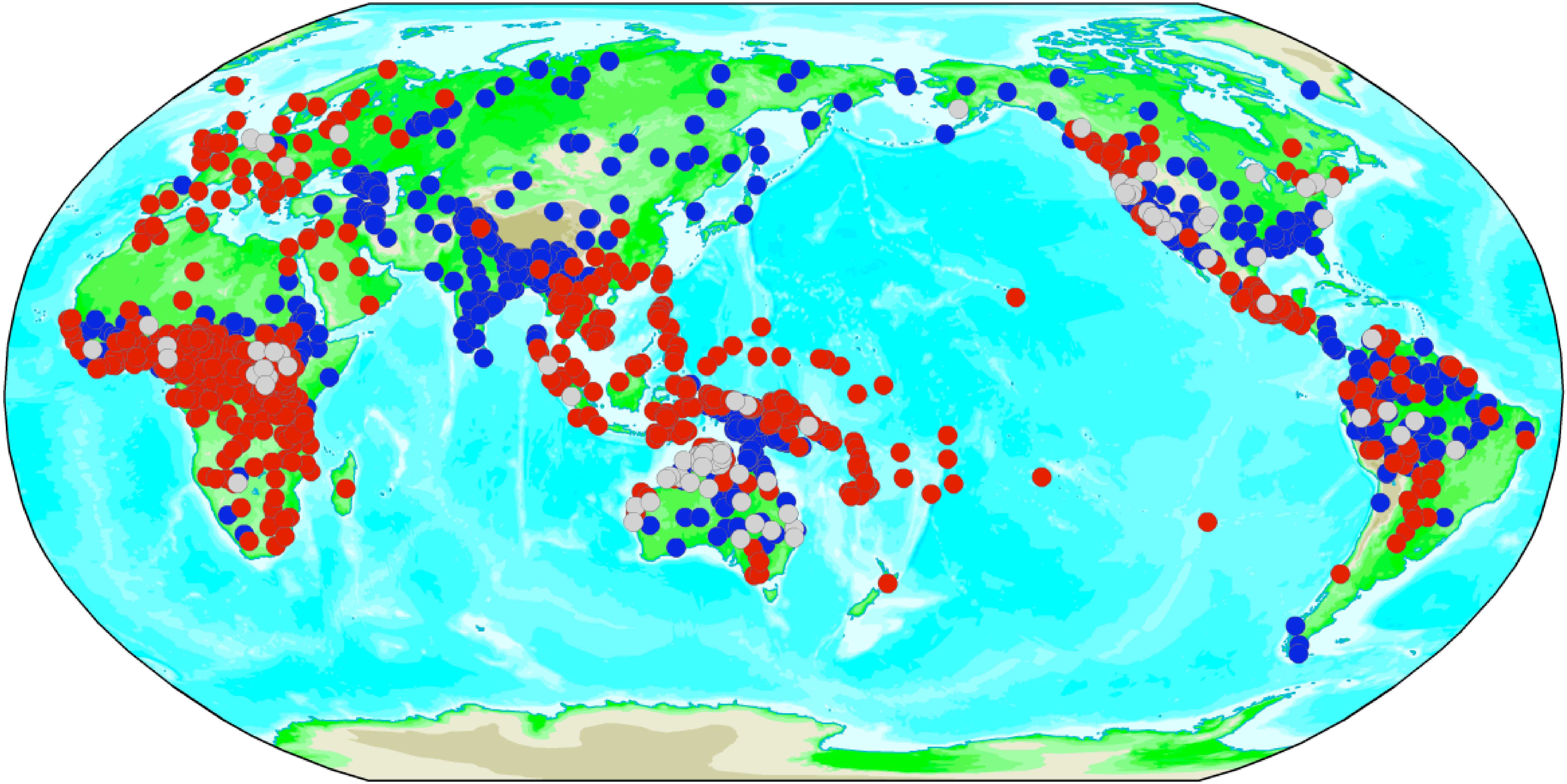
$$R_{f_i} = n \cdot \frac{f_i}{f_{tot}}$$

For $f_i /f_{tot} > 1/n$:

$$R_{f_i} = \frac{1}{n-1}\left(n \cdot \frac{f_i}{f_{tot}} - 1\right) + 1$$

Order of Object and Verb

(by Matthew Dryer)



Computing Rarity Index

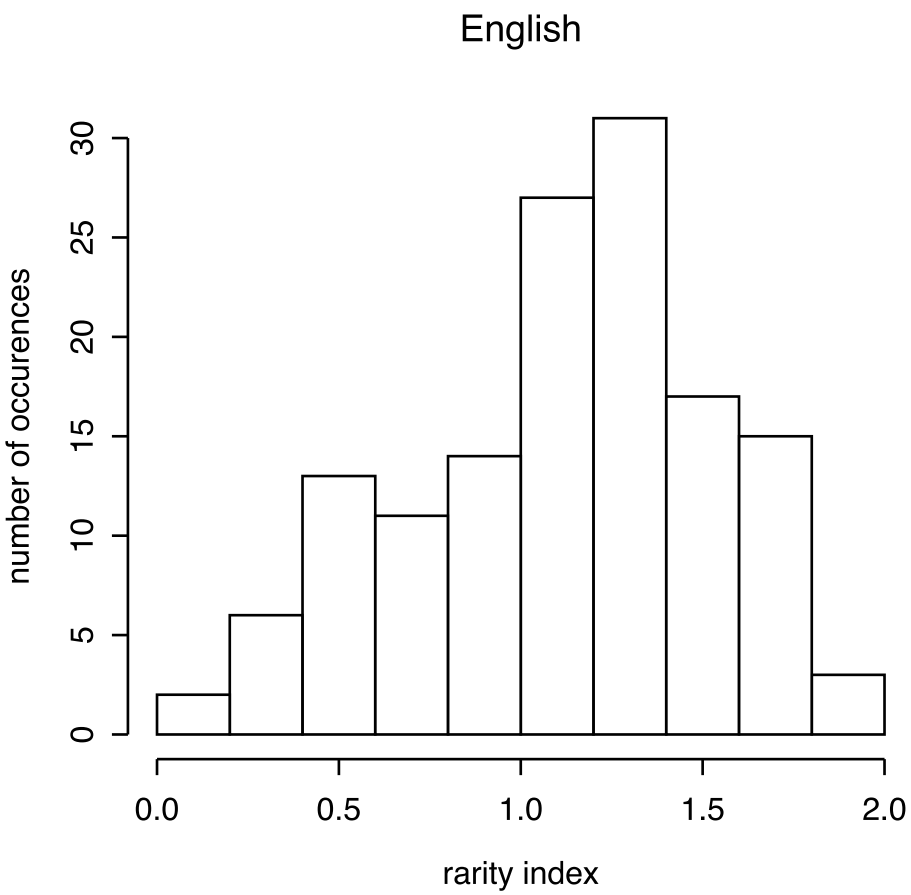
- Three feature values (n = 3)
- Frequencies f_i:

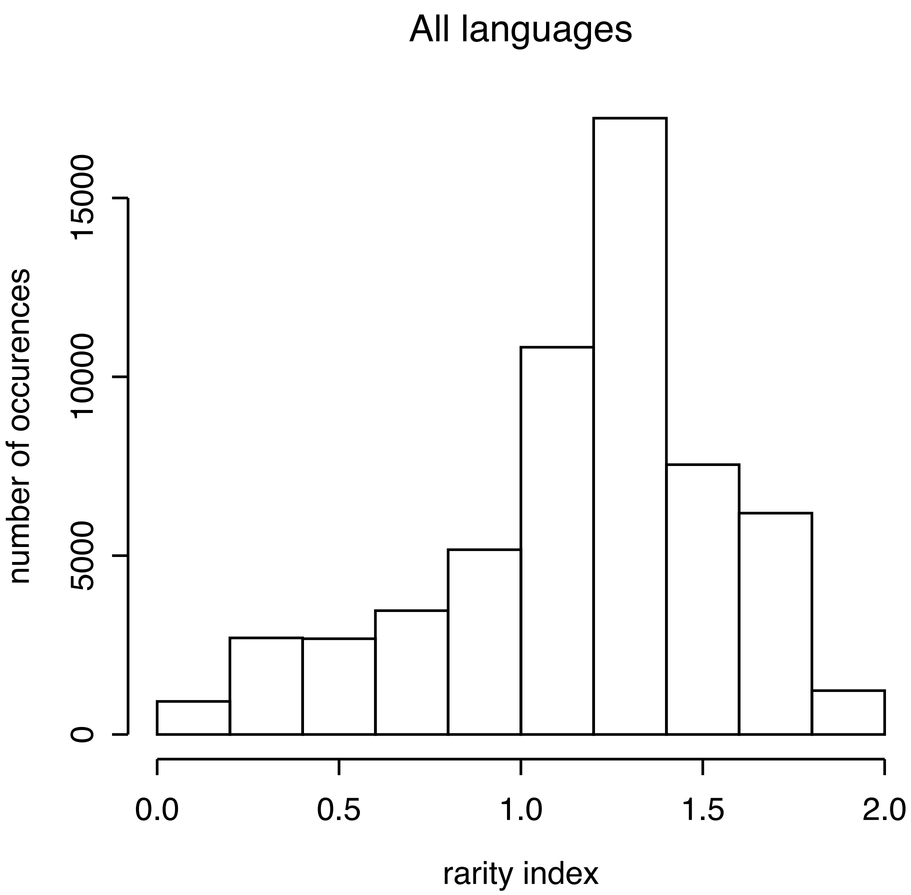
  640 (OV), 639 (VO), 91 (no preference)
- Total f_{tot} = 640 + 639 + 91 = 1370
- R_{ov} = 1.20

  R_{vo} = 1.20

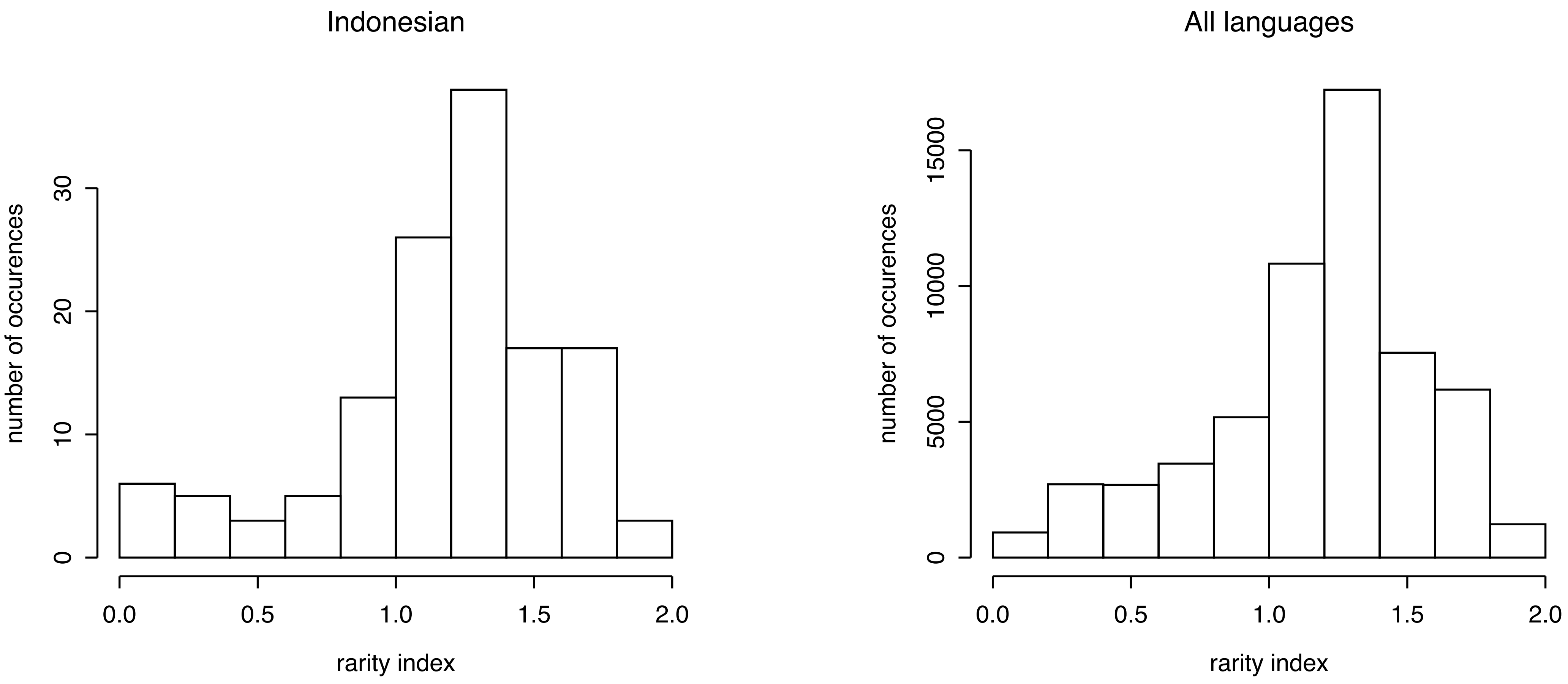
  R_{nopref} = 0.20

Indices of English

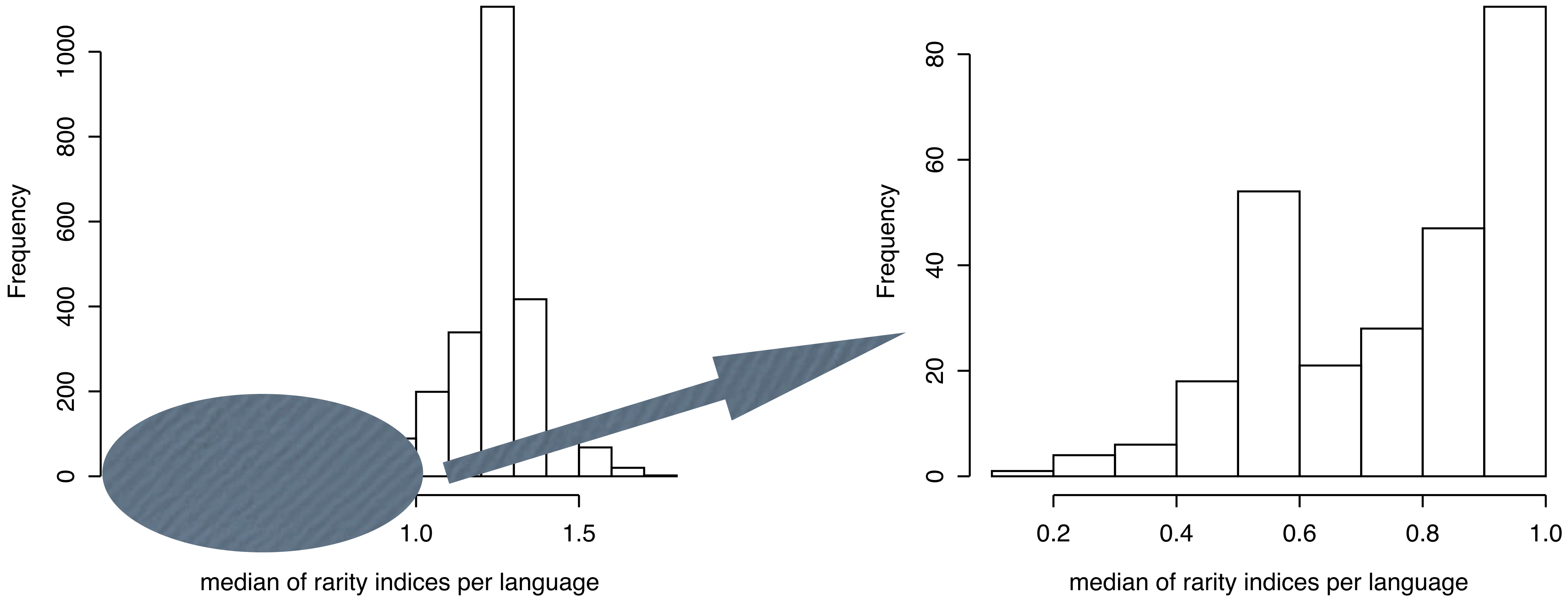




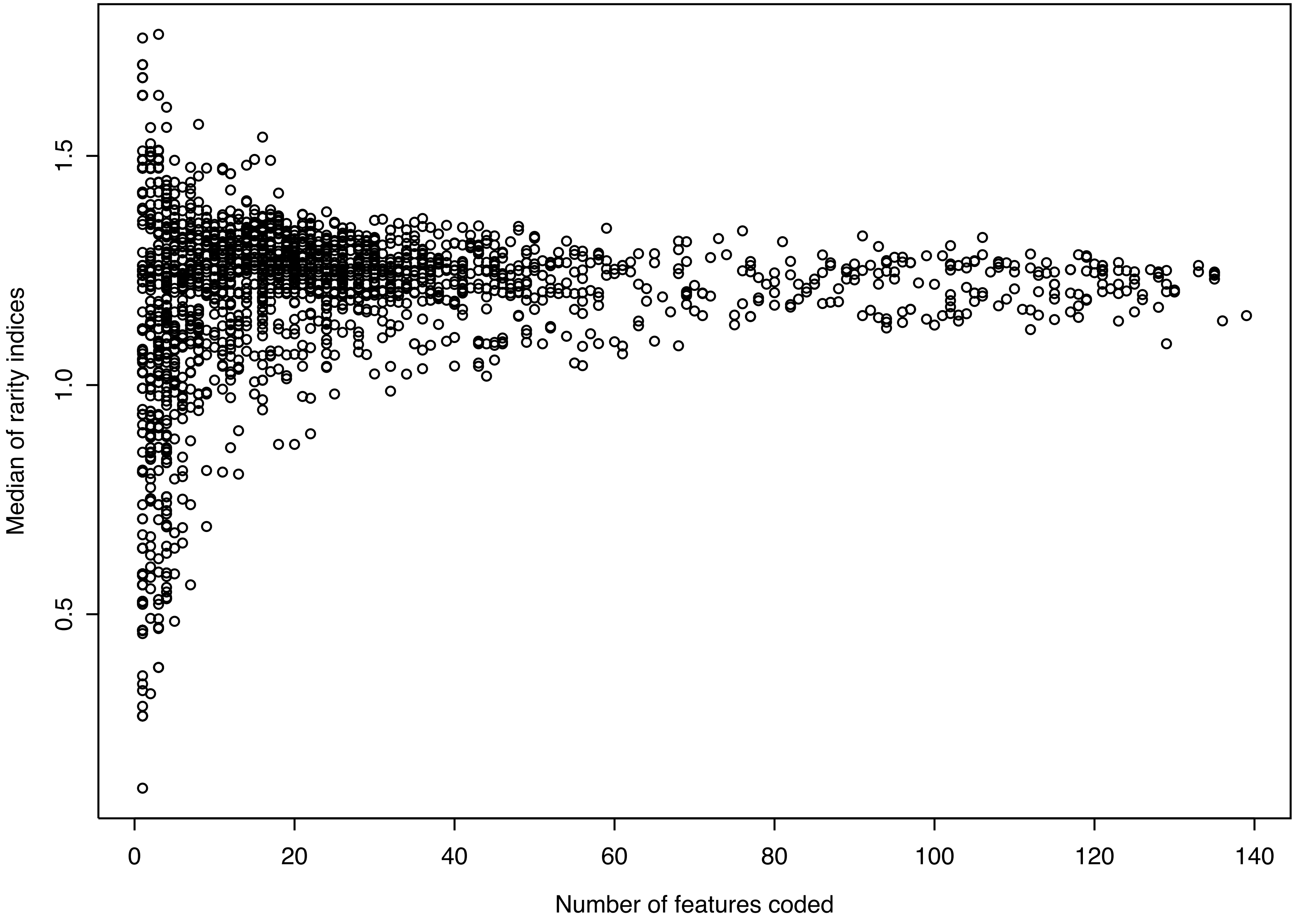
Indices of Indonesian



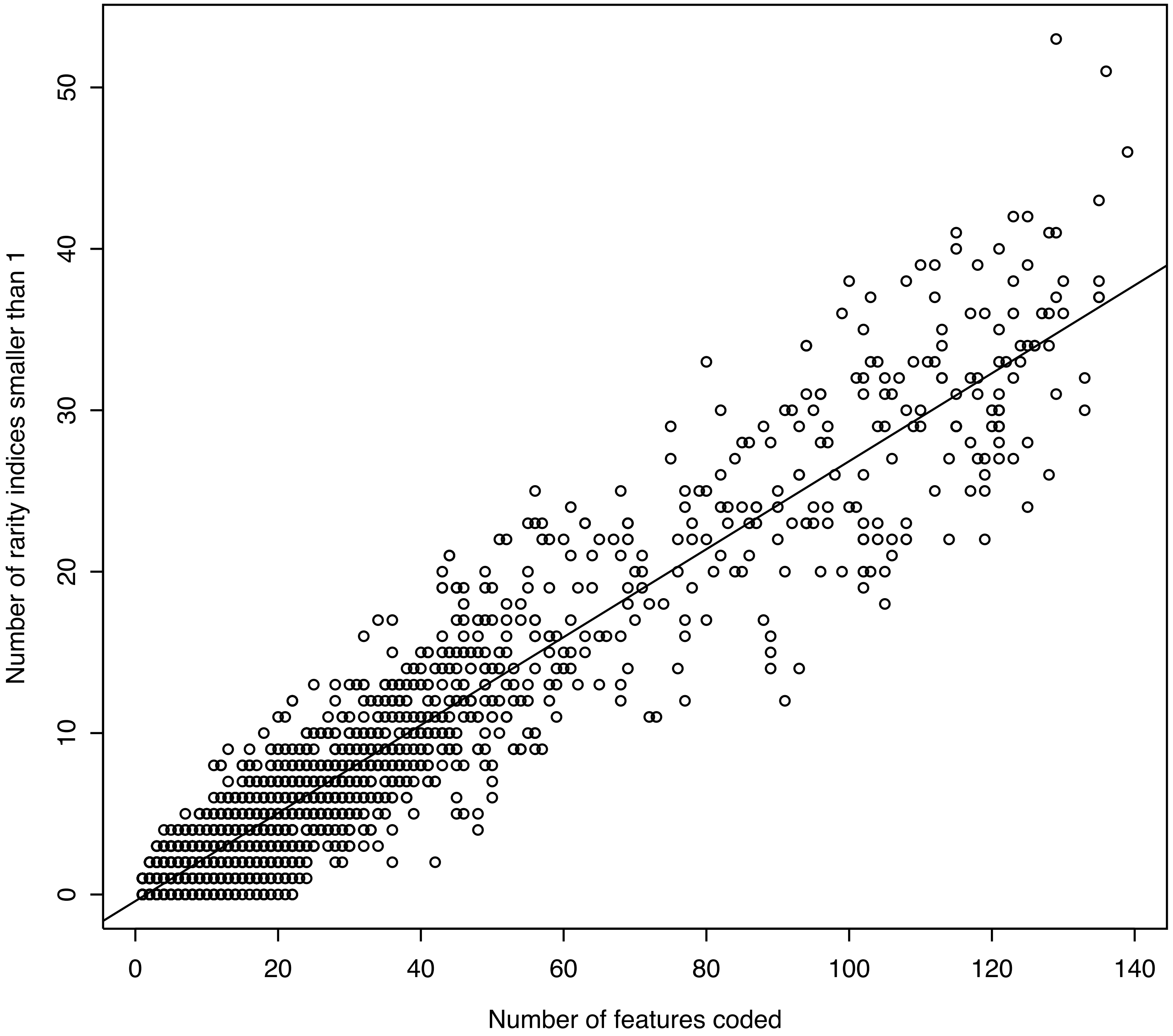
Median of Rarity Indices



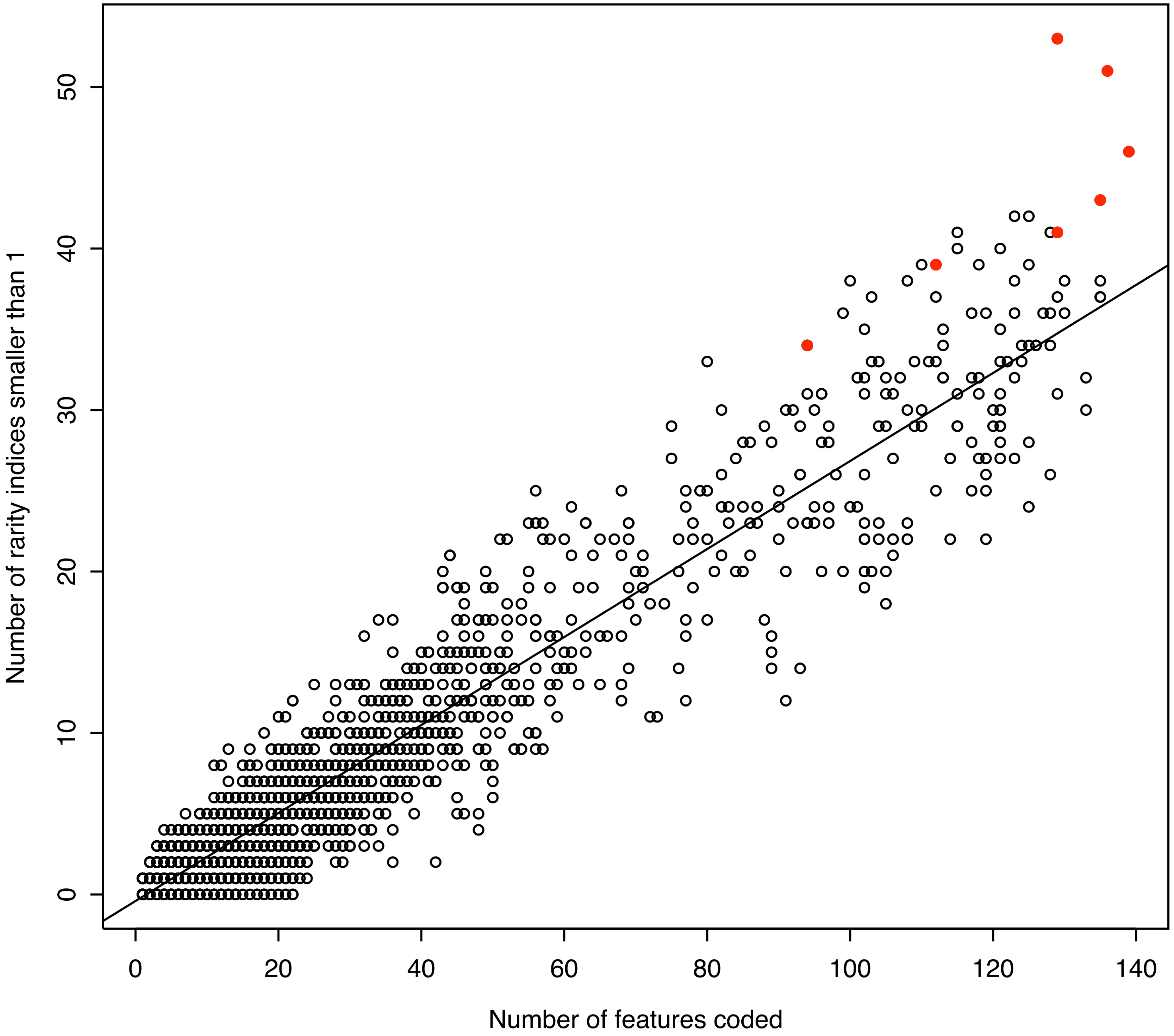
Influence of amount of data



Number of indices smaller than one

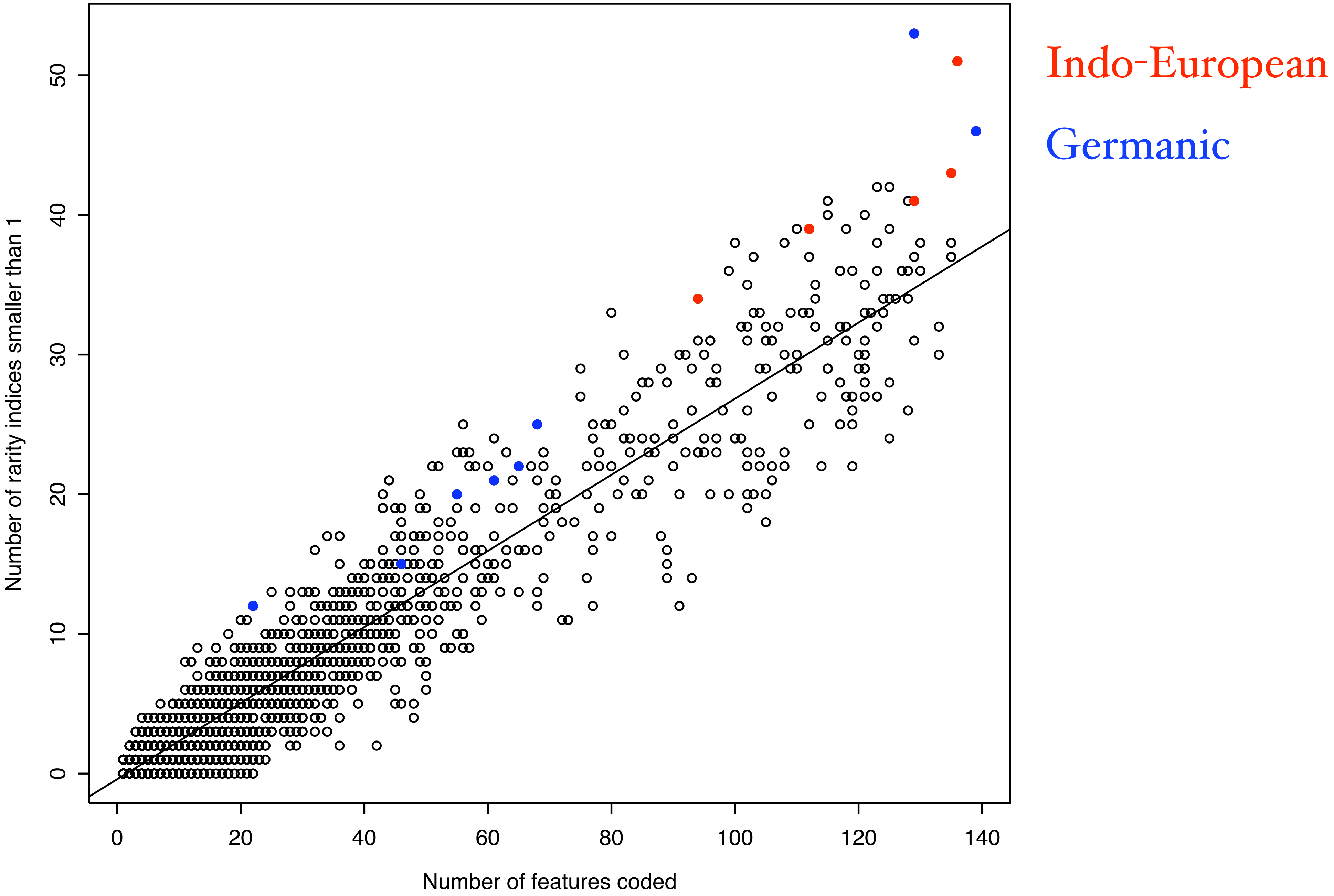


Indo-european languages



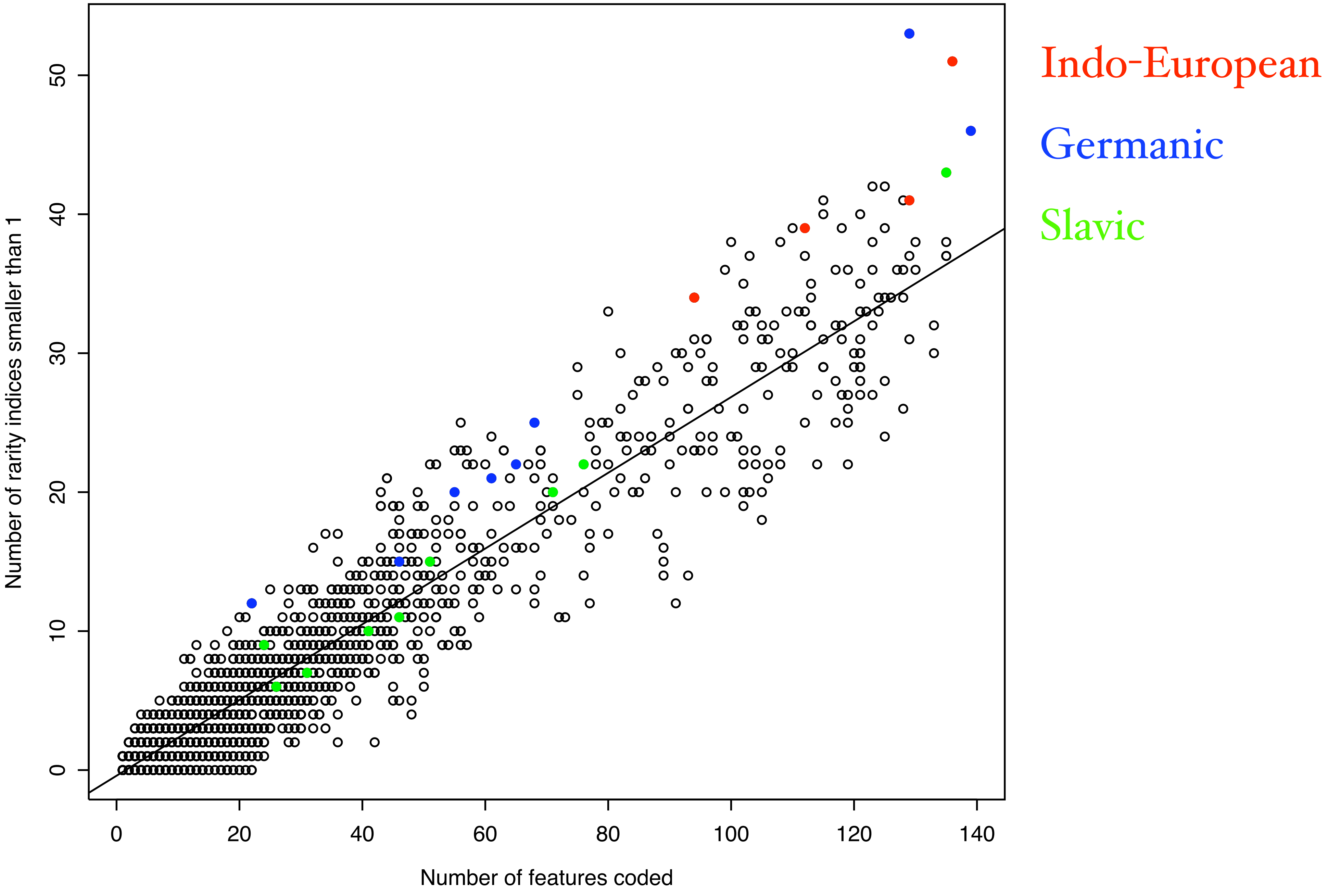
German,
French,
English,
Russian,
Greek,
Irish,
Latvian

Germanic



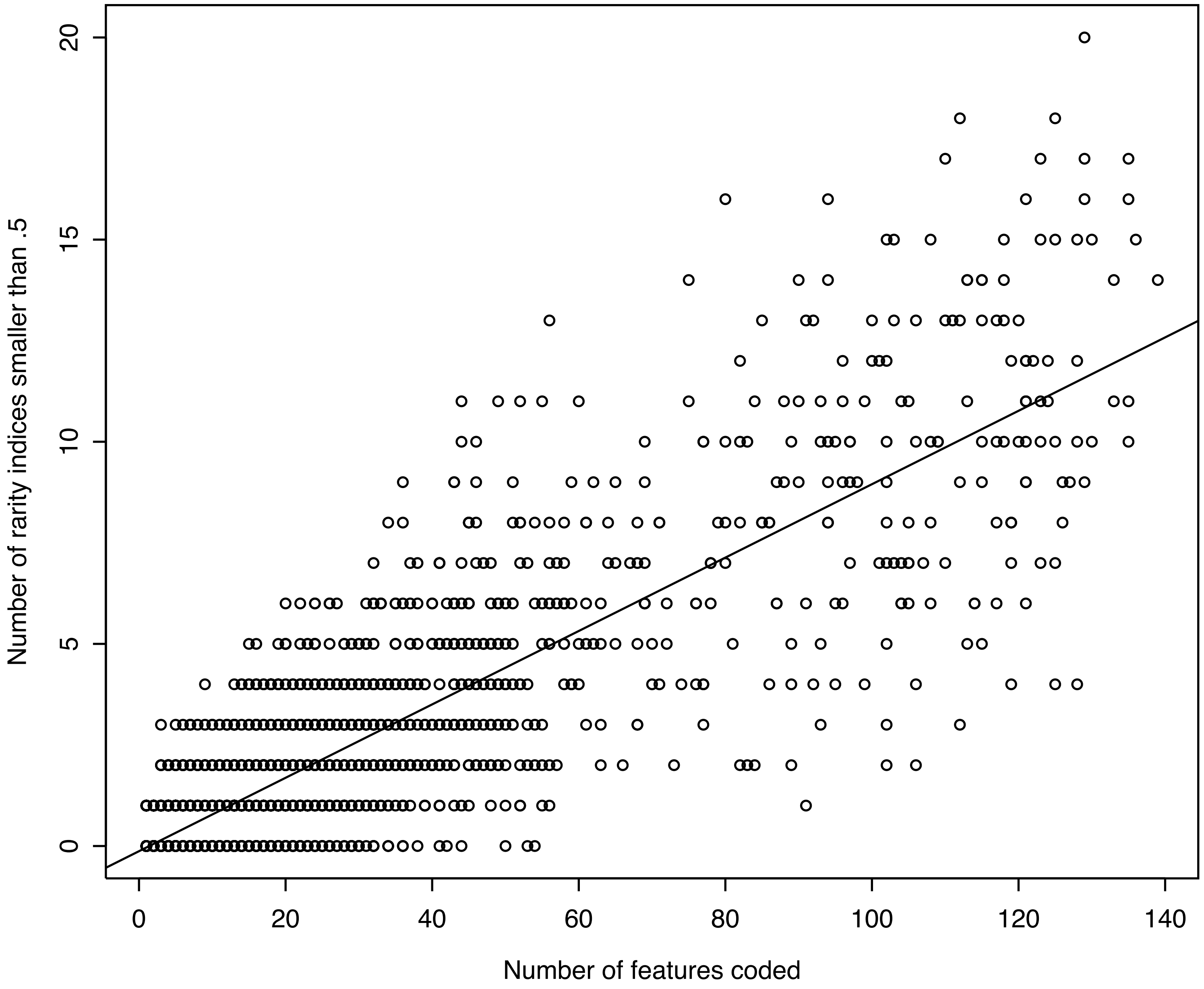
Number of rarity indices smaller than 1
0
10
20
30
40
50
0
20
40
60
80
100
120
140
Number of features coded
Indo-European
Germanic

Slavic

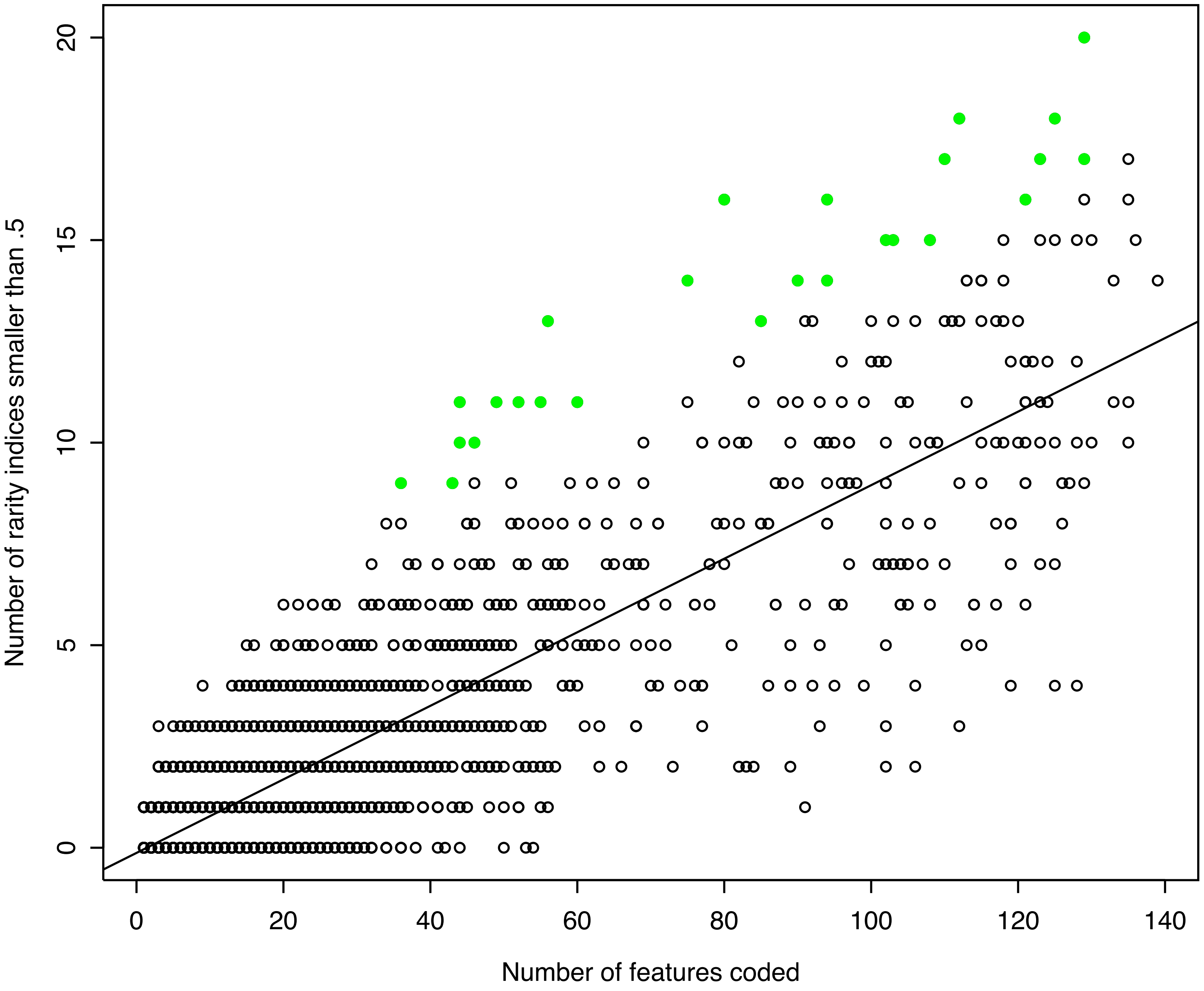

Number of rarity indices smaller than 1
0
10
20
30
40
50
0
20
40
60
80
100
120
140
Number of features coded
Indo-European
Germanic
Slavic

Going more extreme...

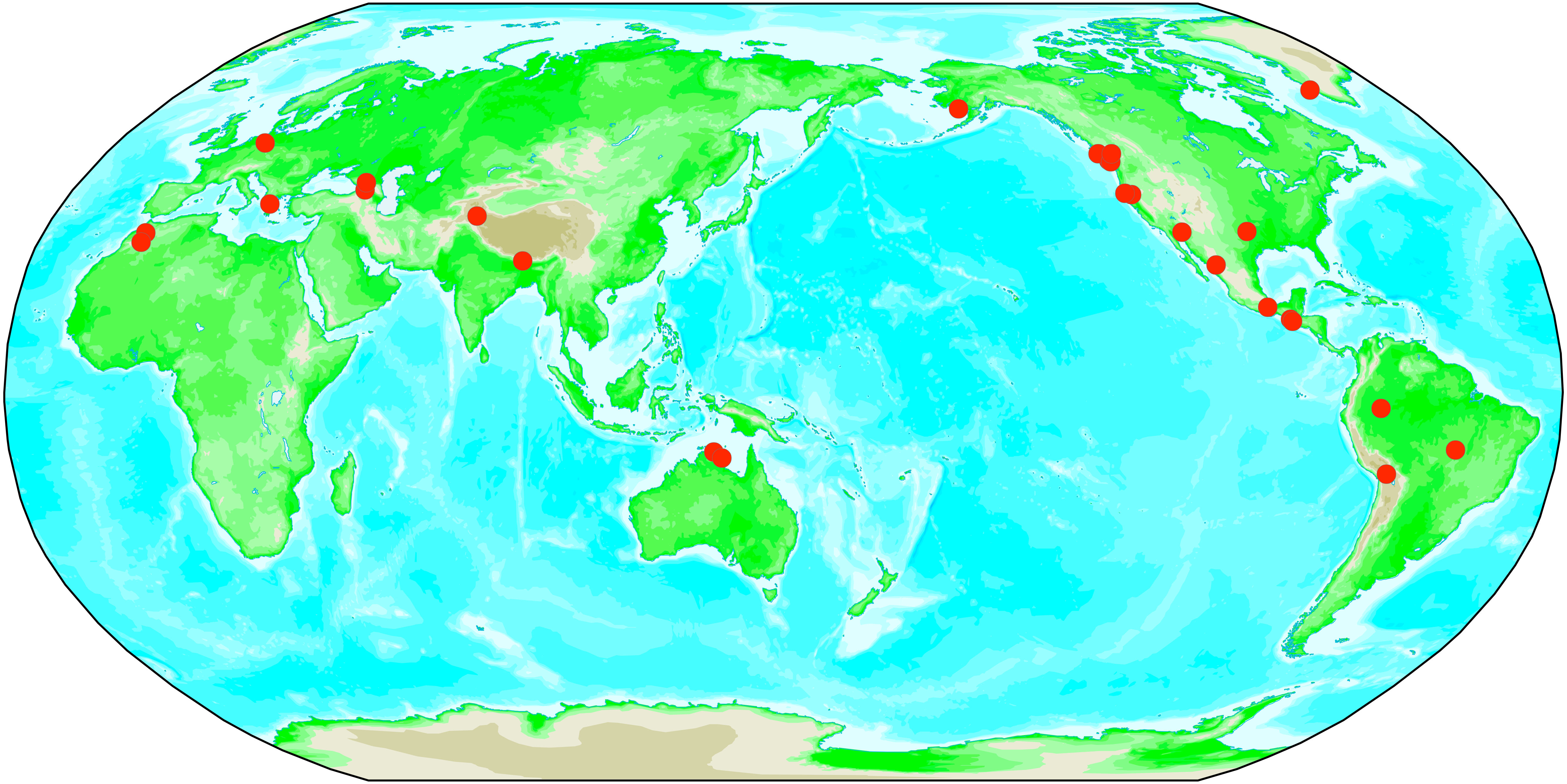
(Number of rarity indices smaller than .5)



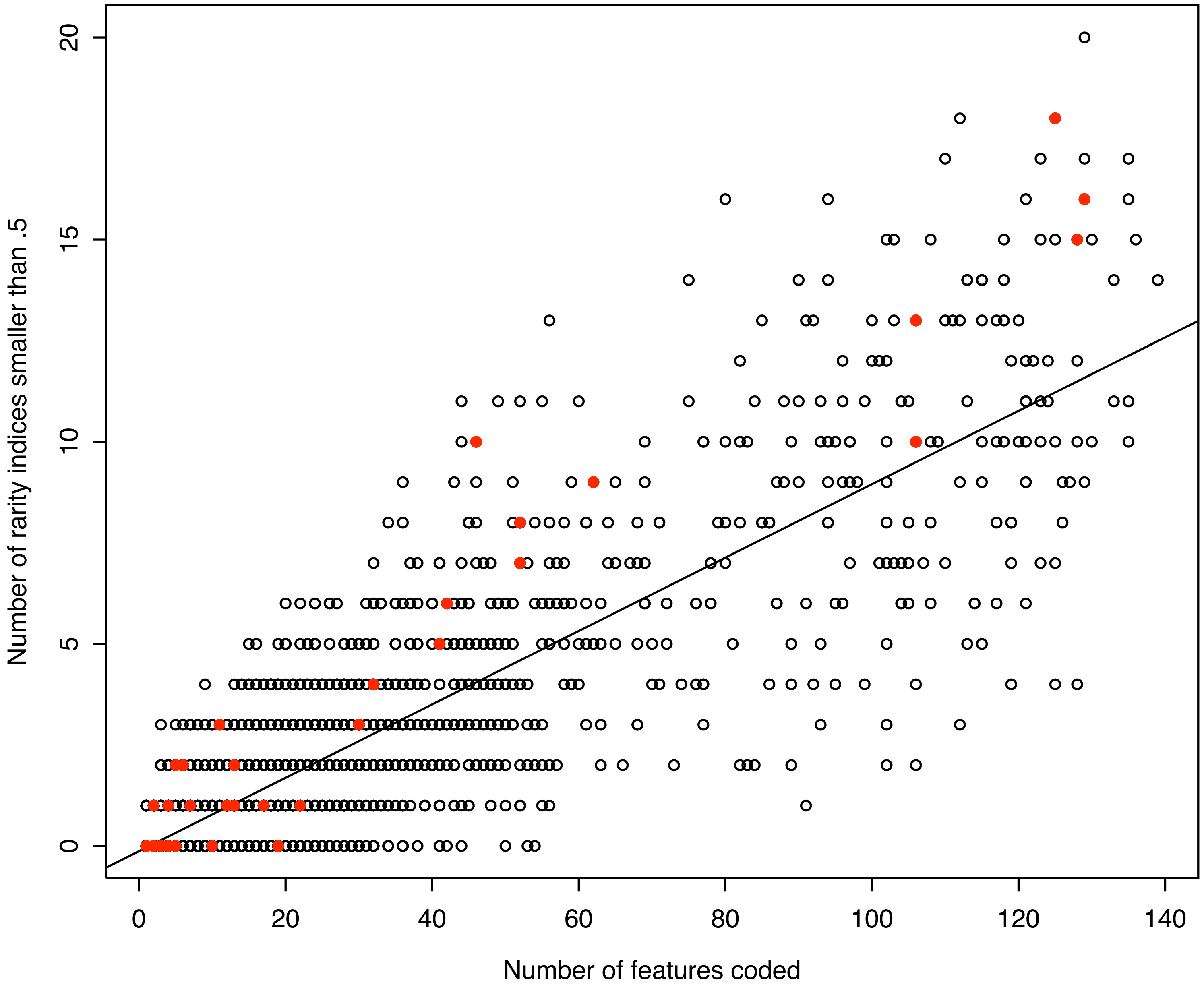
Languages with many rare features



Languages with many rare features



Caucasian languages



Next steps

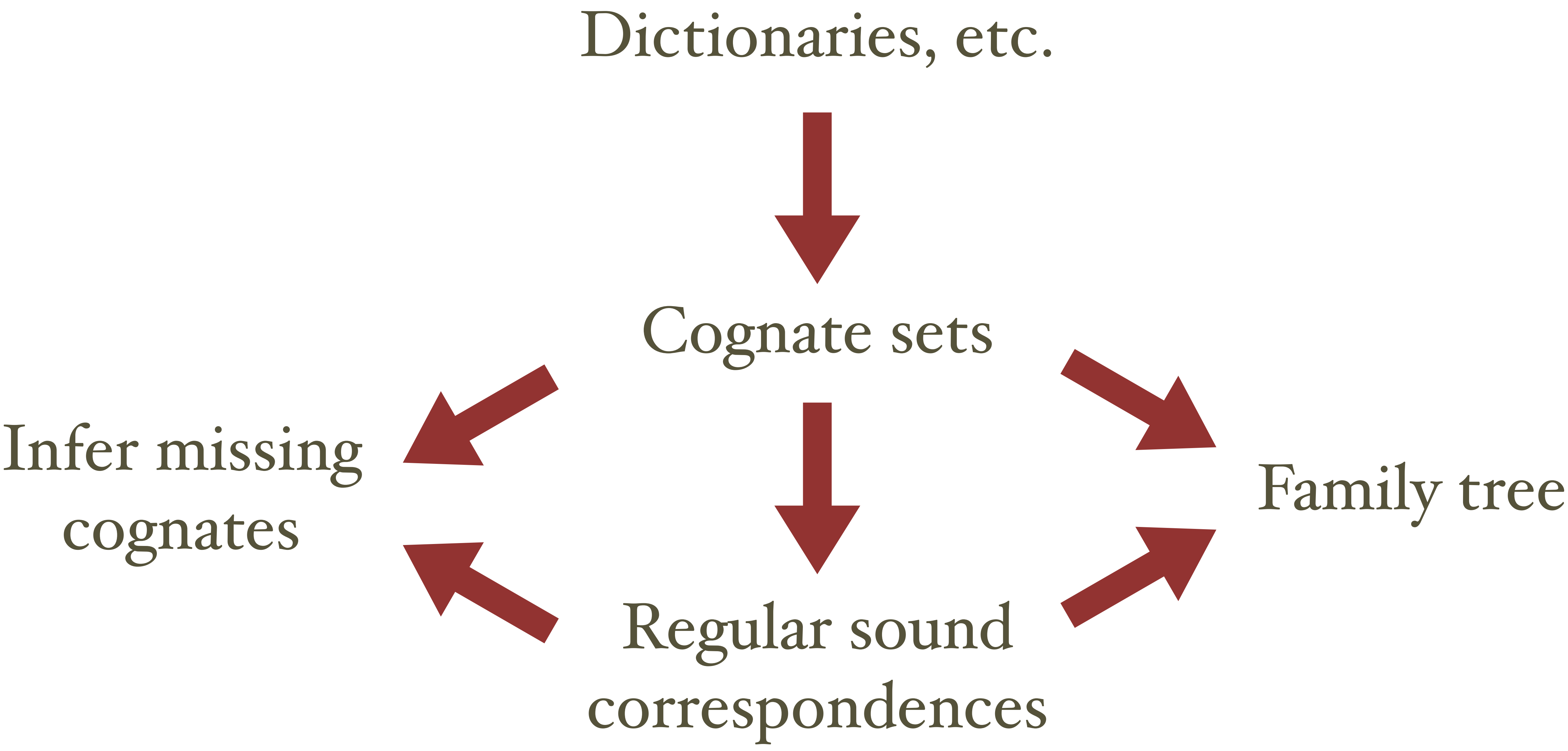
- Improve the integrations of *R* with the WALS-programm
- Going from exploring the data to testing hypotheses
- Taking the feature-perspective: which rare features cluster?

Part 2

Quantitative approaches to historical relatedness

- Range of recent papers, using methods from biological phylogenetic reconstruction to infer linguistic family trees
    - But: only final part of historical-comparative method is taken up
- Almost all on higher groupings of Indo-European
    - But: one should first check validity by applying the methods to agreed upon classifications

Inferring tree is just one of the many possibilities of quantitative approaches



Dictionaries, etc.

Cognate sets

Infer missing cognates

Regular sound correspondences

Family tree

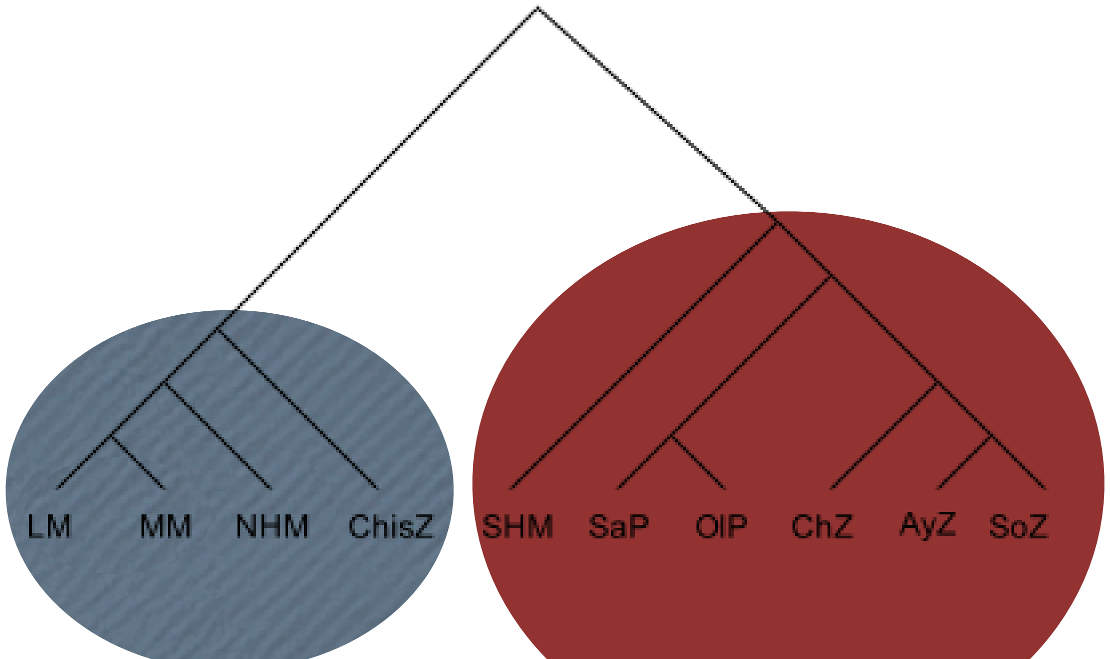
Testing Holm's approach

(together with Søren Wichmann and David Kamholz)

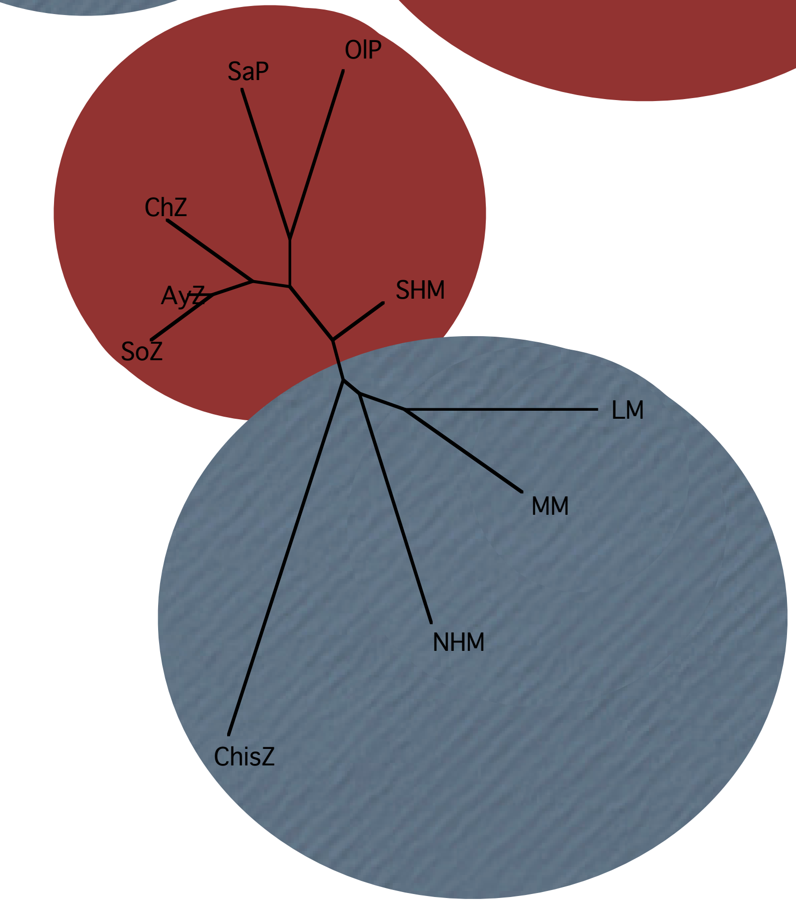
- Holm's idea: use etymological dictionary instead of Swadesh-style wordlists
- By counting the number of shared retentions for each pair of languages, he estimates the relative point of split between each pair (dissimilarity estimates)
- In simulations (by Kamholz) the approach seems to work
- We tested the method on Mixe-Zoque data

Interpreting the estimates

Following Holm’s interpretation:

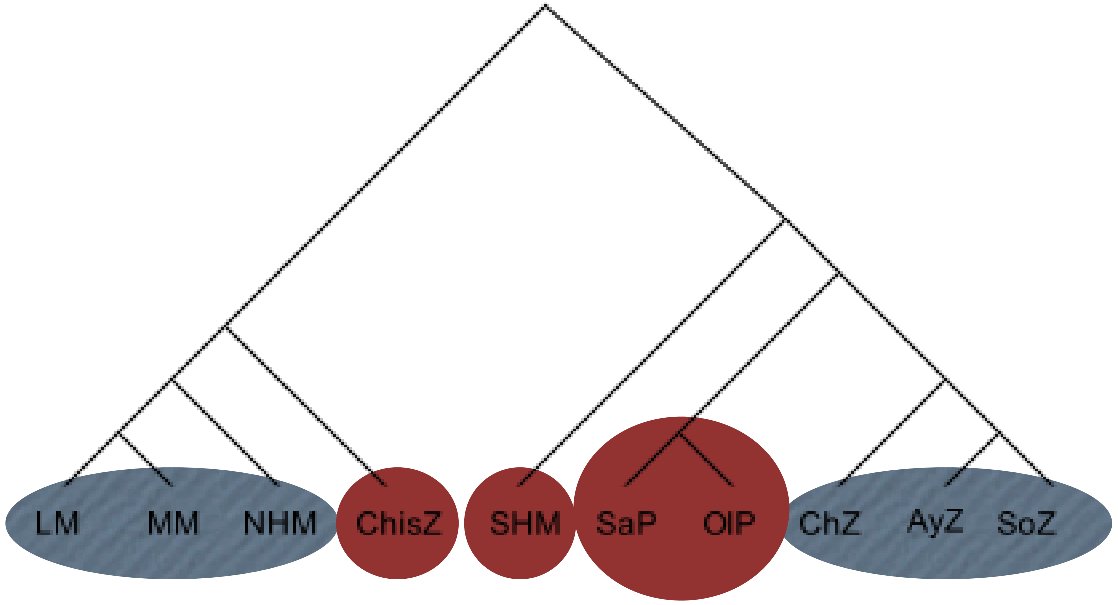


Using ADDTREE on the estimates:

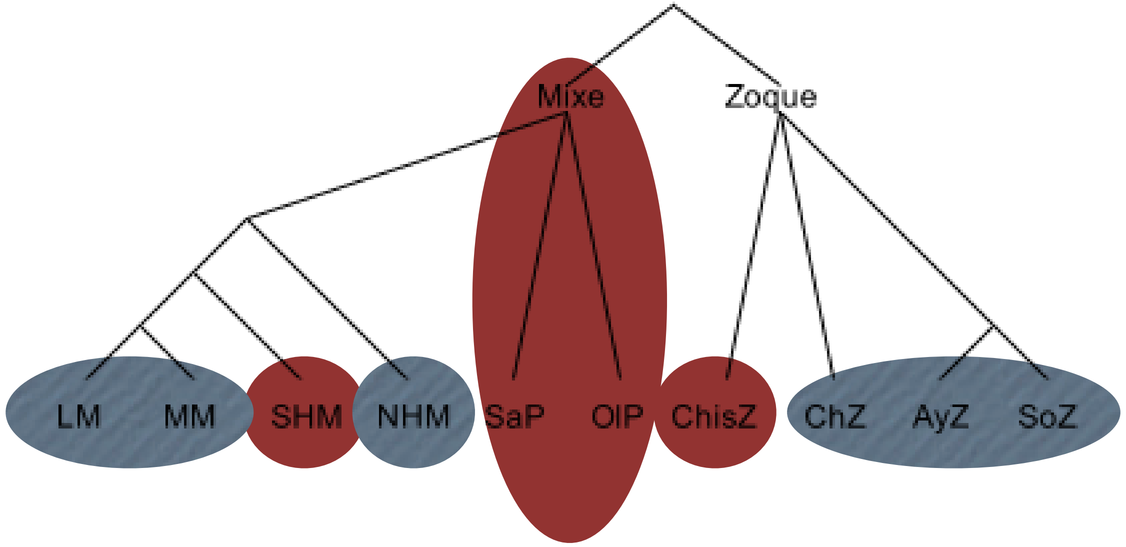


How did it work out?

Holm's method:

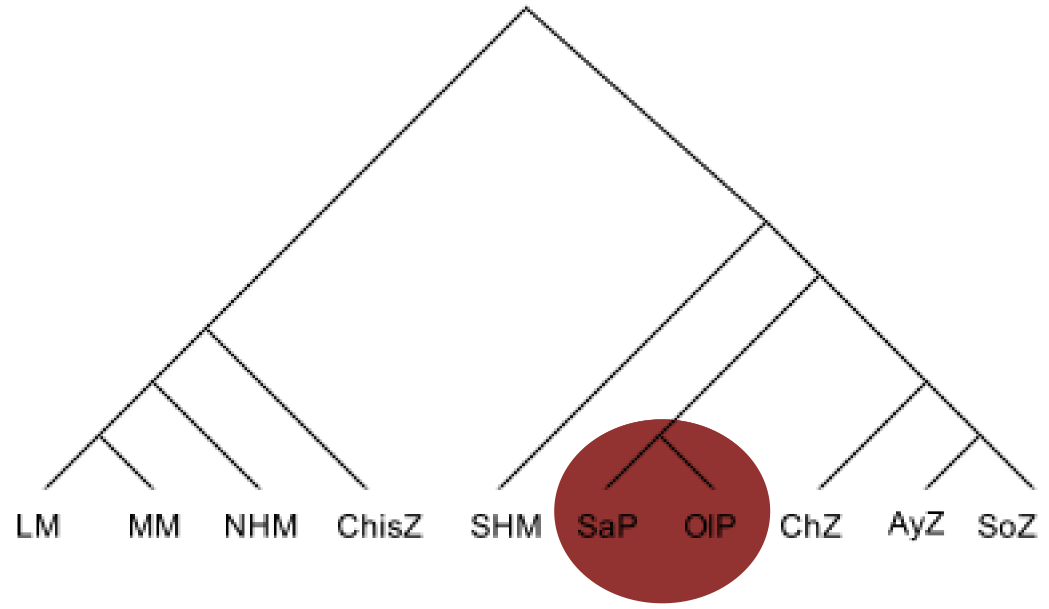


Wichmann (1995)

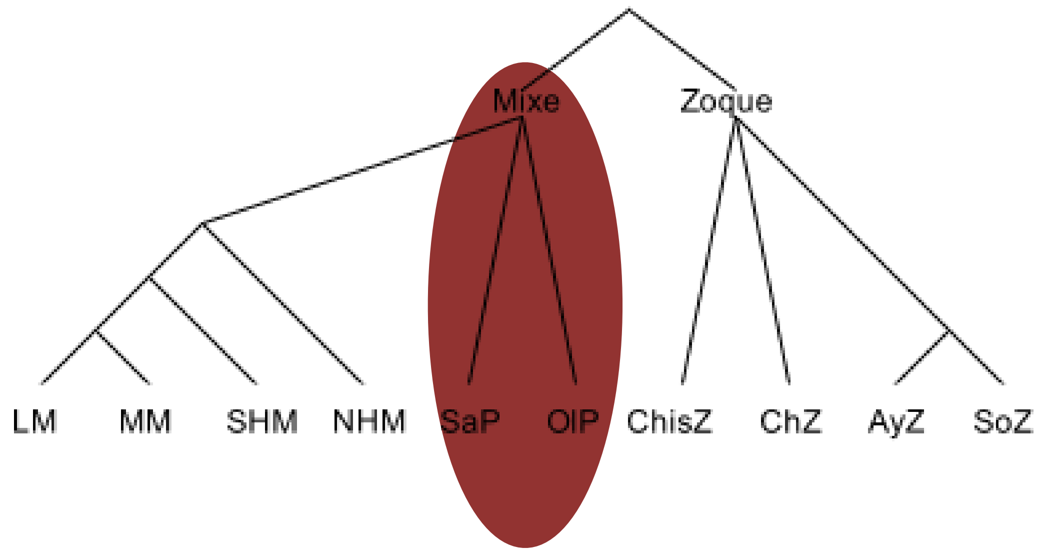


The Popolucan errors

Holm's method:



Wichmann (1995)

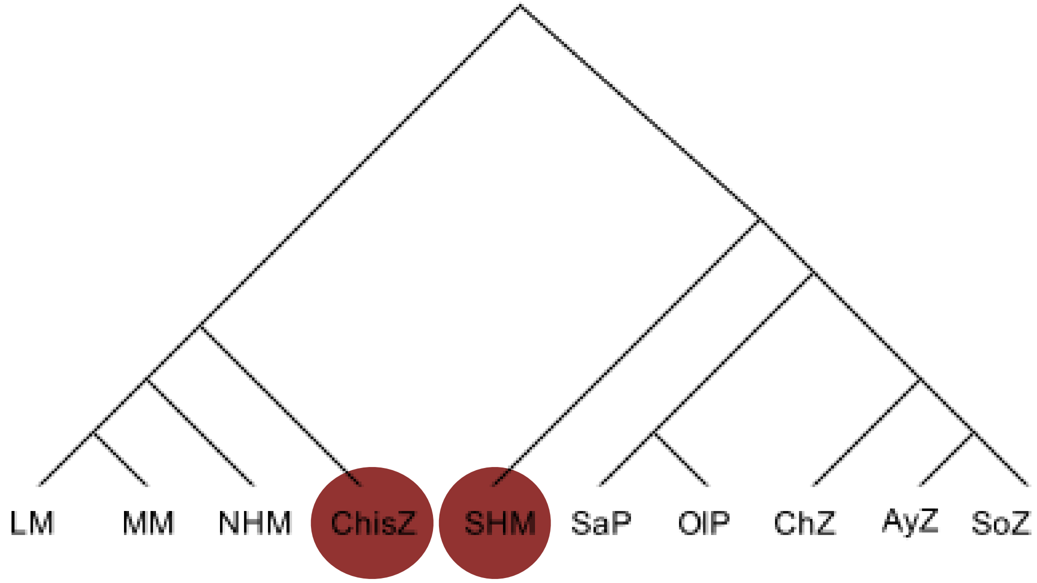


The Popolucan errors

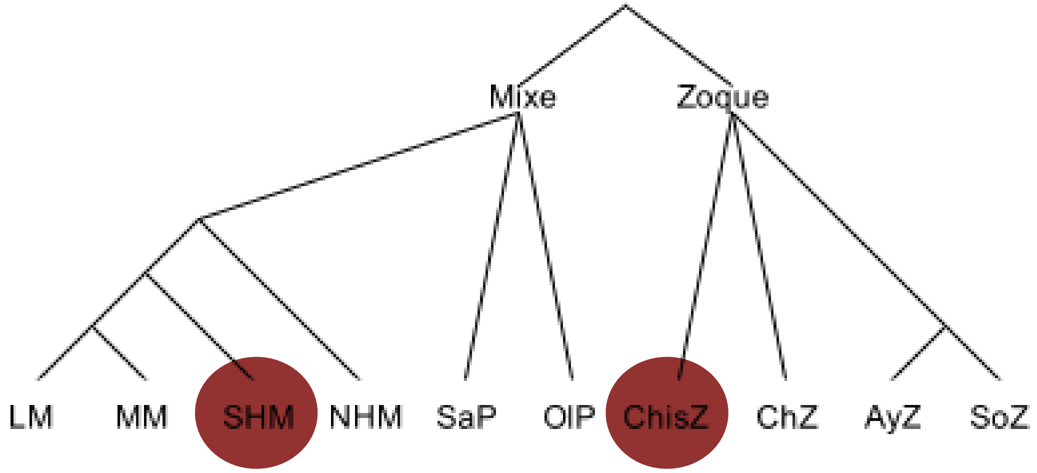
- Error 1: they are grouped together, because of many shared retentions
    - But: there are no shared innovations!
- Error 2: they are grouped with Zoque instead of with Mixe
    - Circularity problem: reconstruction depends on tree, and Holm makes tree out of reconstruction

The other two errors

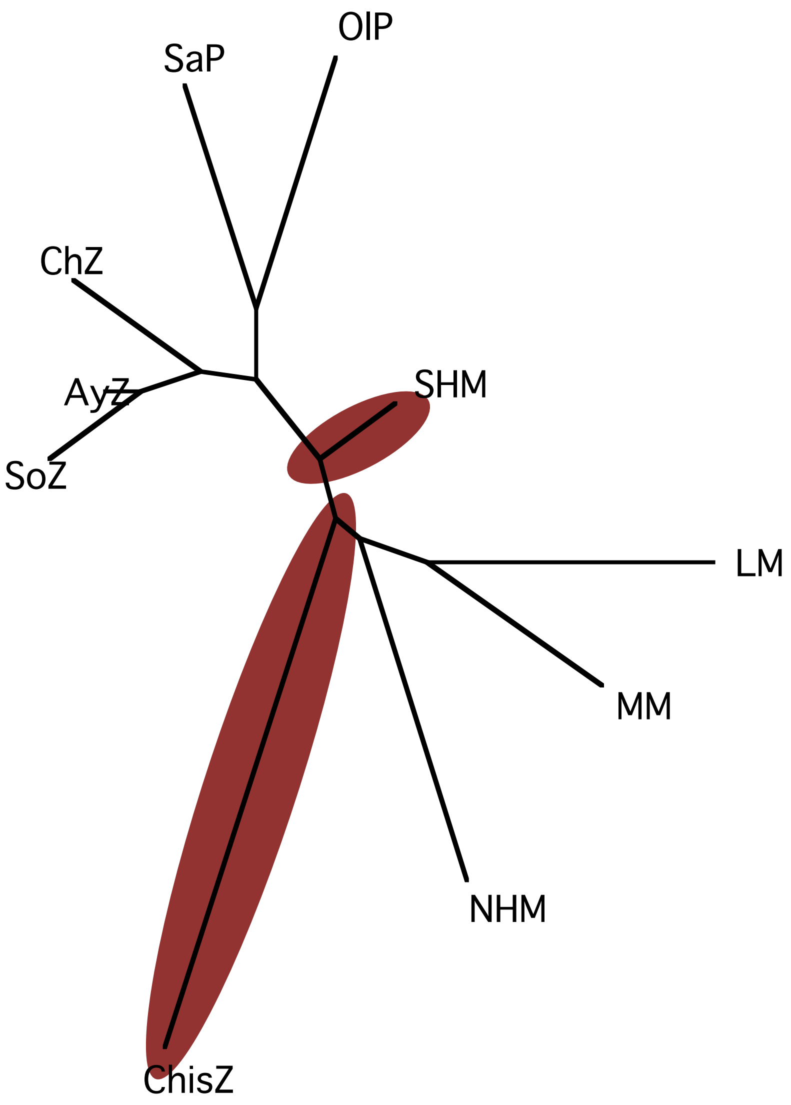
Holm's method:



Wichmann (1995)

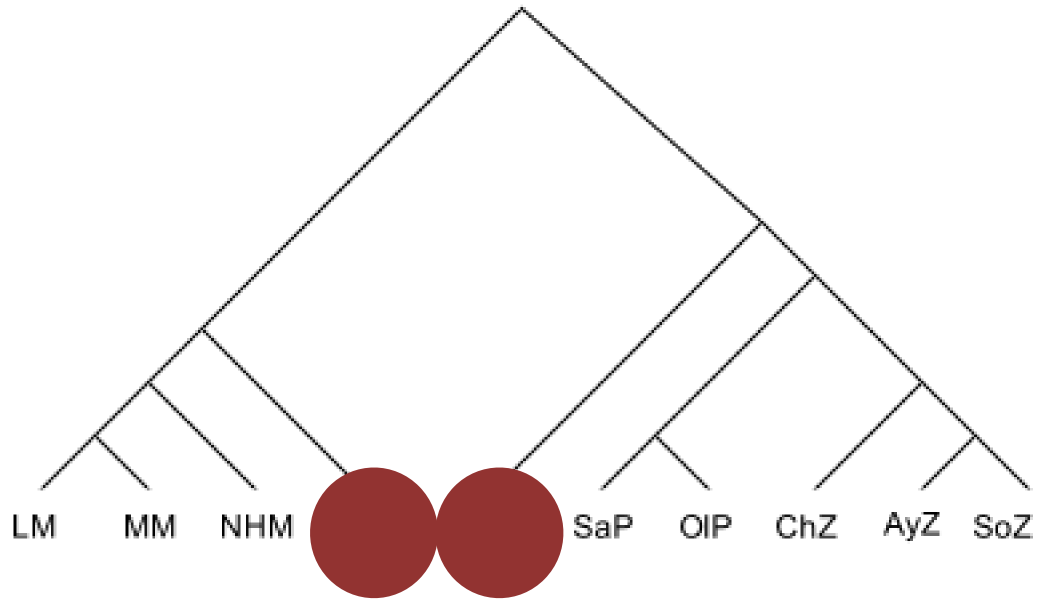


Difficult to place in the tree

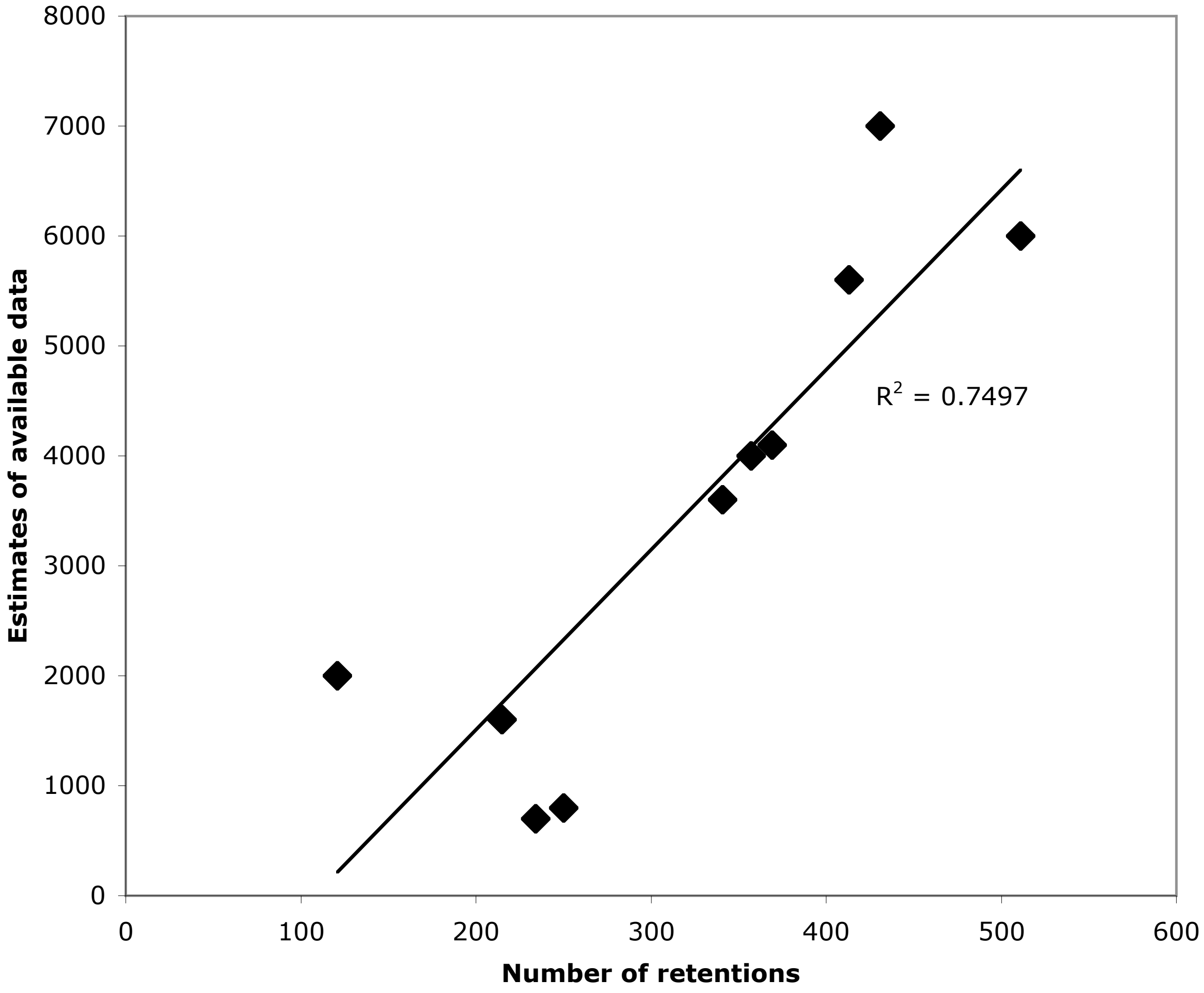


Estimates of available knowledge about Mixe-Zoque

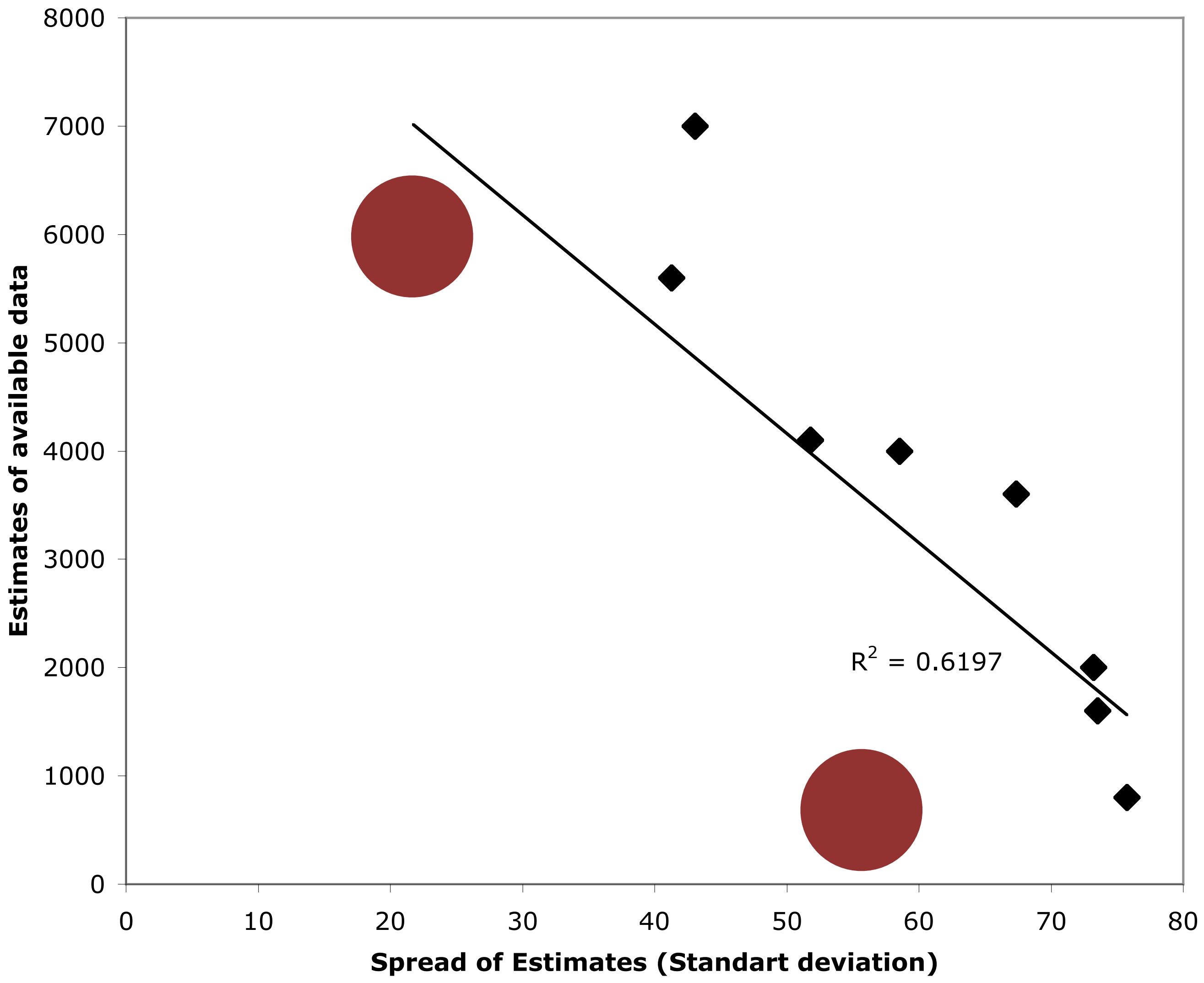
Language	Number of dictionary entries
LM	7000
[redacted]	[redacted]
NHM	5600
MM	4100
OlP	4000
SaP	3600
AyZ	2000
ChZ	1600
SoZ	800
[redacted]	[redacted]



Number of retentions depends on available knowledge



Spread of estimates depends on available knowledge



Summary of problems

- Absence of shared innovations is not counted
- The data that enter in the analysis (i.e. reconstructed etyma) partly depend on the outcome (i.e. the tree)
- Unbalanced amount of available data distorts the estimates

The End