On the distribution of rare characteristics

Michael Cysouw

Max Planck Institute for Evolutionary Anthropology

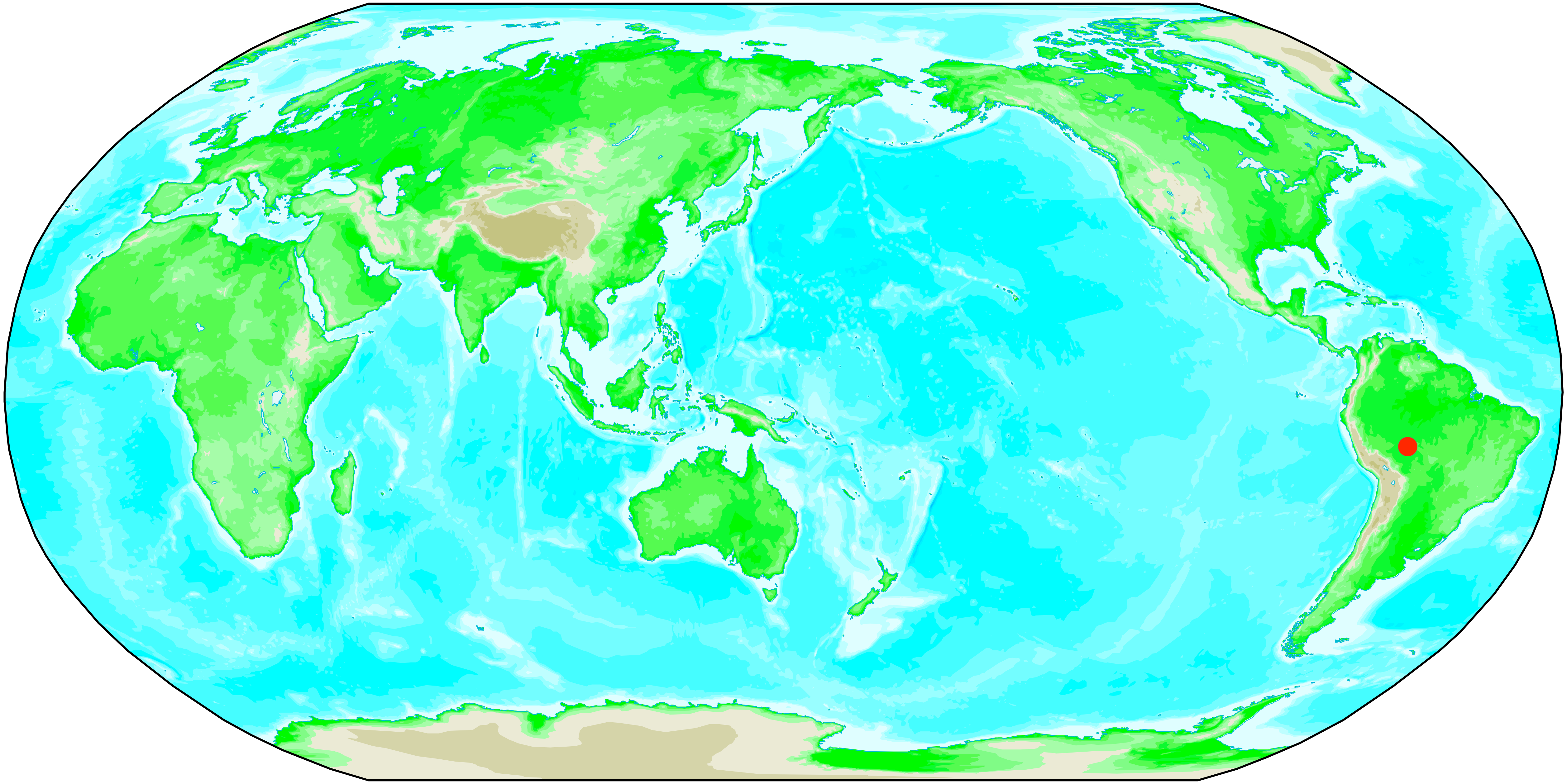
The basic idea

- Use the WALS data for ‘holistic’ typology
- Not look at the content of the features, but at their relative ubiquity
- Are there languages/families/areas that have more rare features than other?

And the winners are:

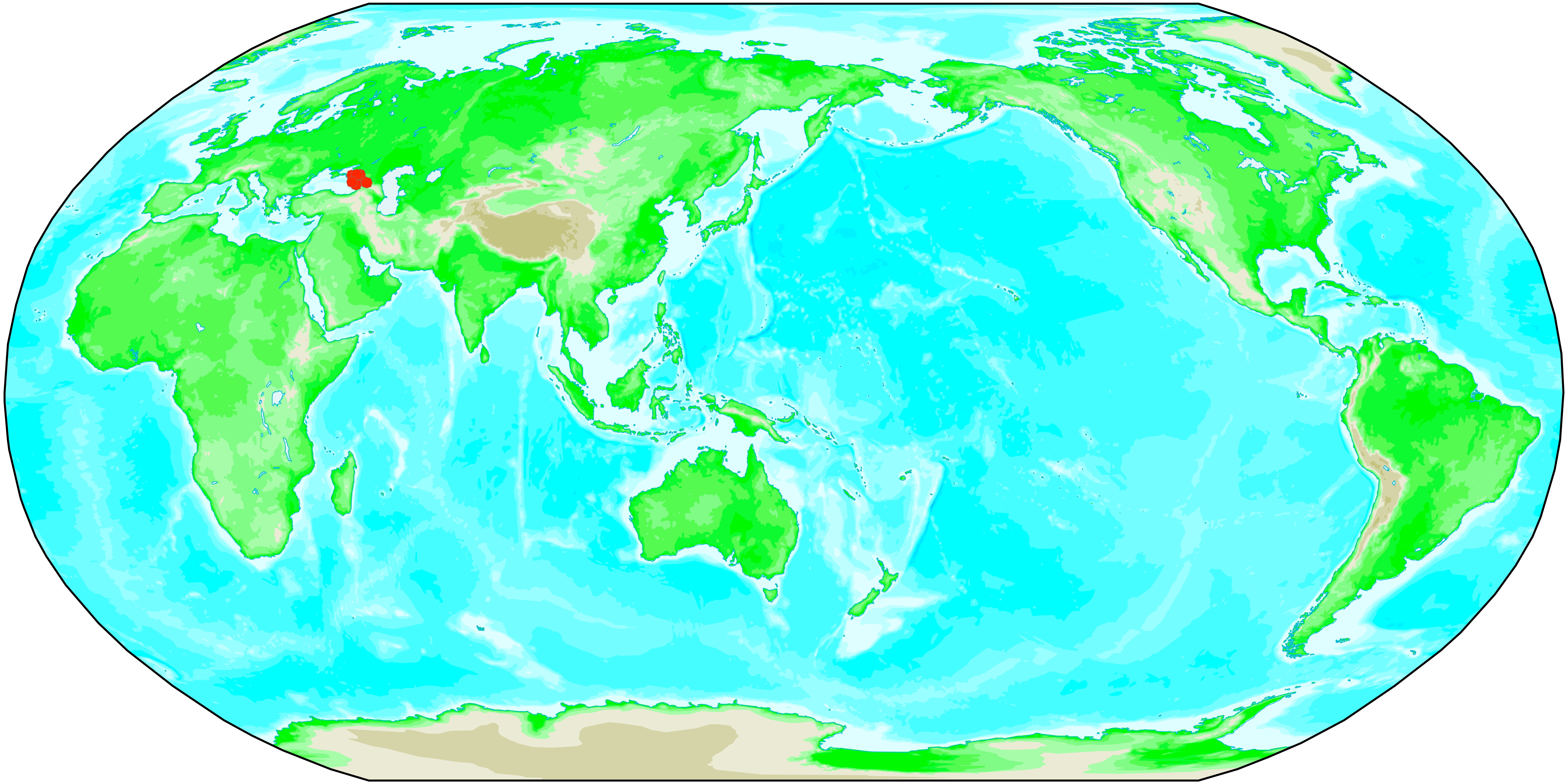
In the category
‘Rarest Individual Language’

Wari’



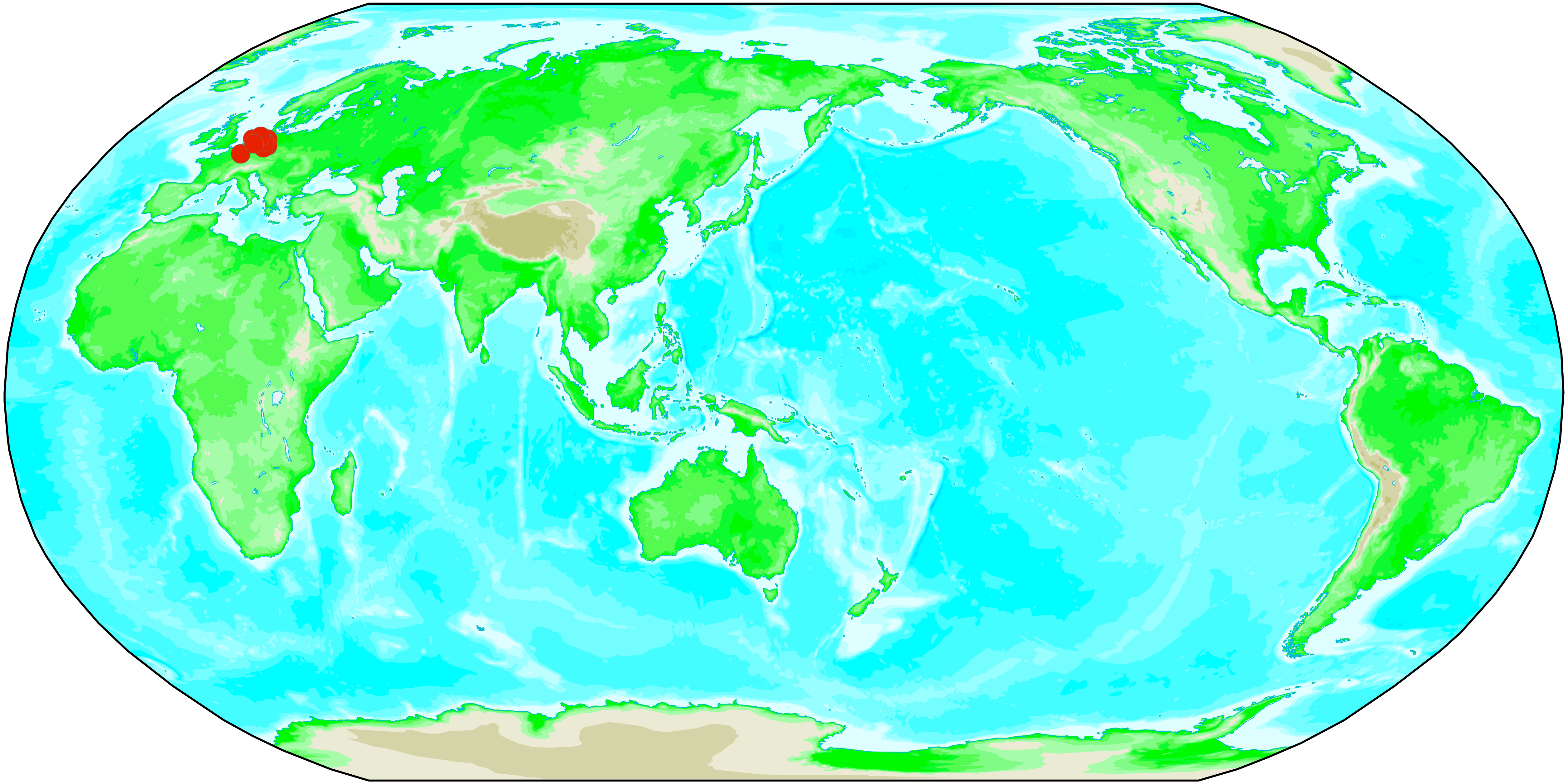
In the category
‘Rarest Genealogical Group’

Northwest Caucasian



In the category ‘Rarest Geographical Area’

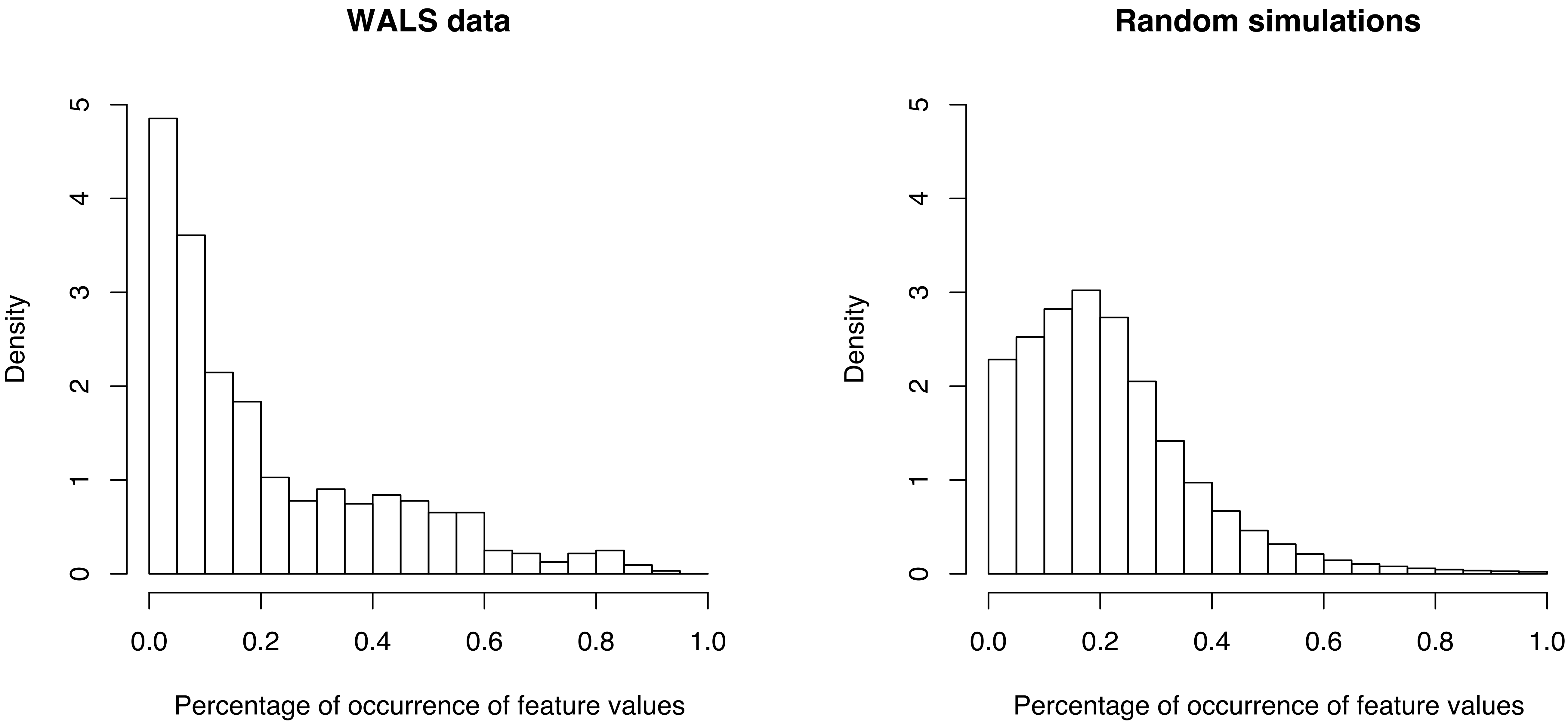
Northwest Continental Europe



Distribution of values

- 138 features, total 643 values
- Chance distribution complex because of varying number of values per feature
- Distribution is highly skewed - there are more than expected many:
    - extremely rare values
    - mid-range values

Distribution of values



WALS data

Random simulations

Rarity Index R_i

n = number of feature values

f_i = frequency of feature value i

f_{tot} = total number of languages included

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

Inverse Index

I used the inverse instead:

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

Because the mean of all R_i values is one:

$$\frac{\sum_{i=1}^{n}(R_i \cdot f_i)}{f_{tot}} = 1$$

Rarity Index for a language

- Each language gets a list of *Rarity Indices*, one for each feature-value in the WALS

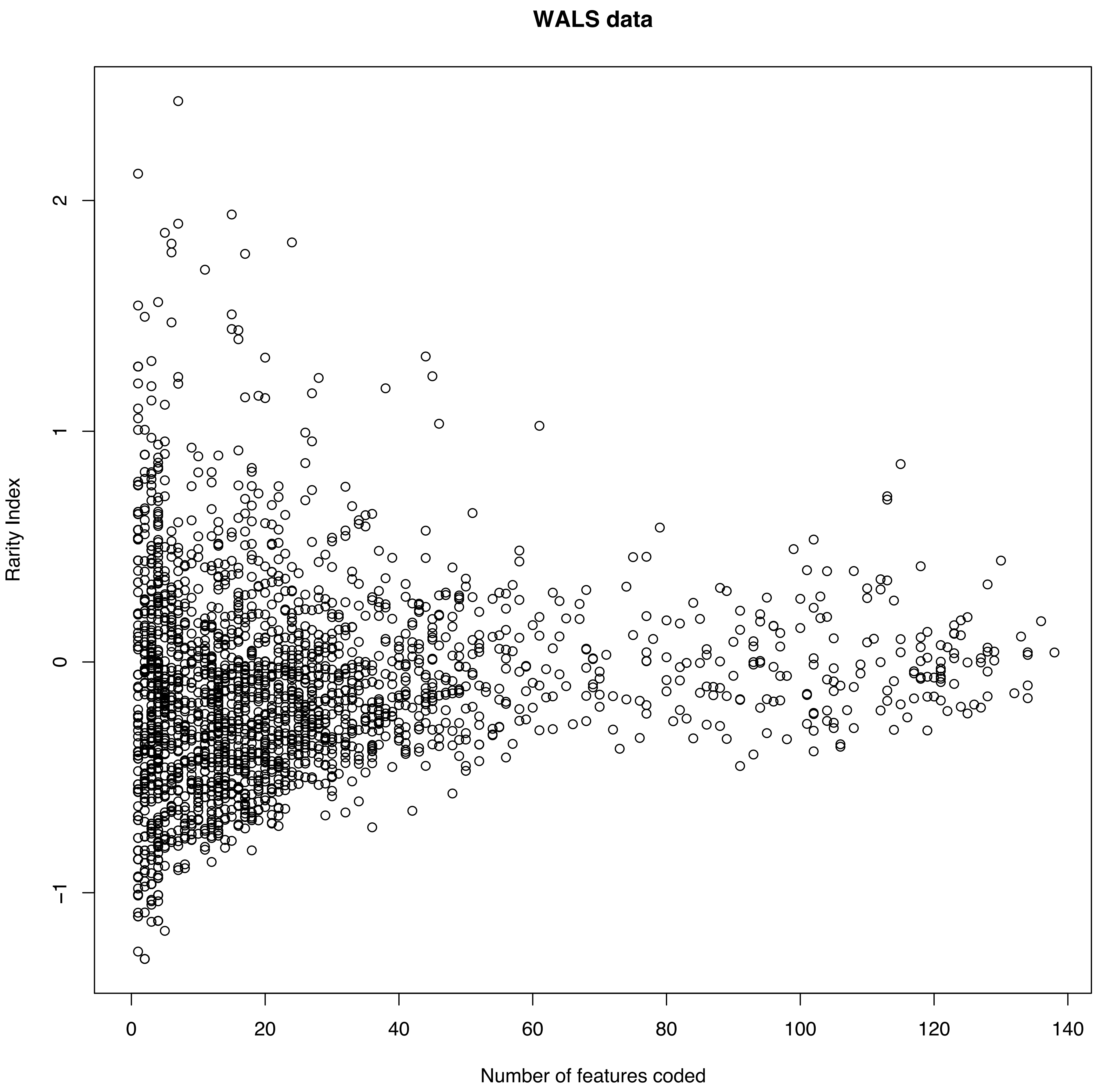
- How should a list of *Rarity Indices* be interpreted?

- I tried various approaches (mean, median, various kinds of weighted means, with/without normalisation of the indices)

- They all correlated highly with each other, so I hold on to the simple mean

Rarity Index for a language

- What is the ‘most rare’ language?
- Simply taking the highest *Mean Rarity* is no good...
- If only few features are coded in the WALS, there will be strong random effects
- Plot *Mean Rarity* by number of features coded:

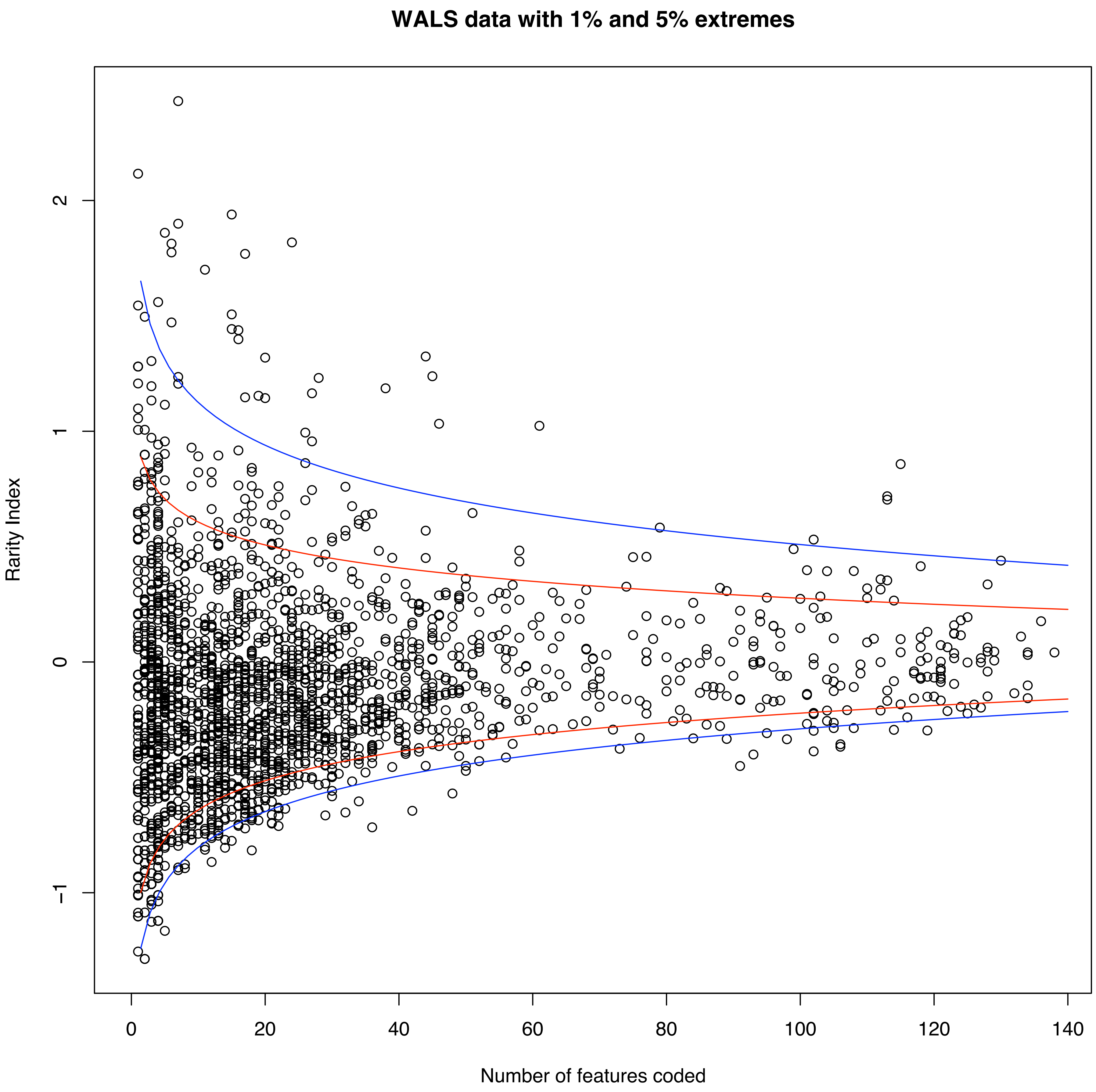

WALS data
Rarity Index
Number of features coded
−1
0
1
2
0
20
40
60
80
100
120
140

Rarity Index for a language

- Assess distribution by randomisation

- Keeping the actual numbers of the feature-values constant, I constructed random languages for each number of features

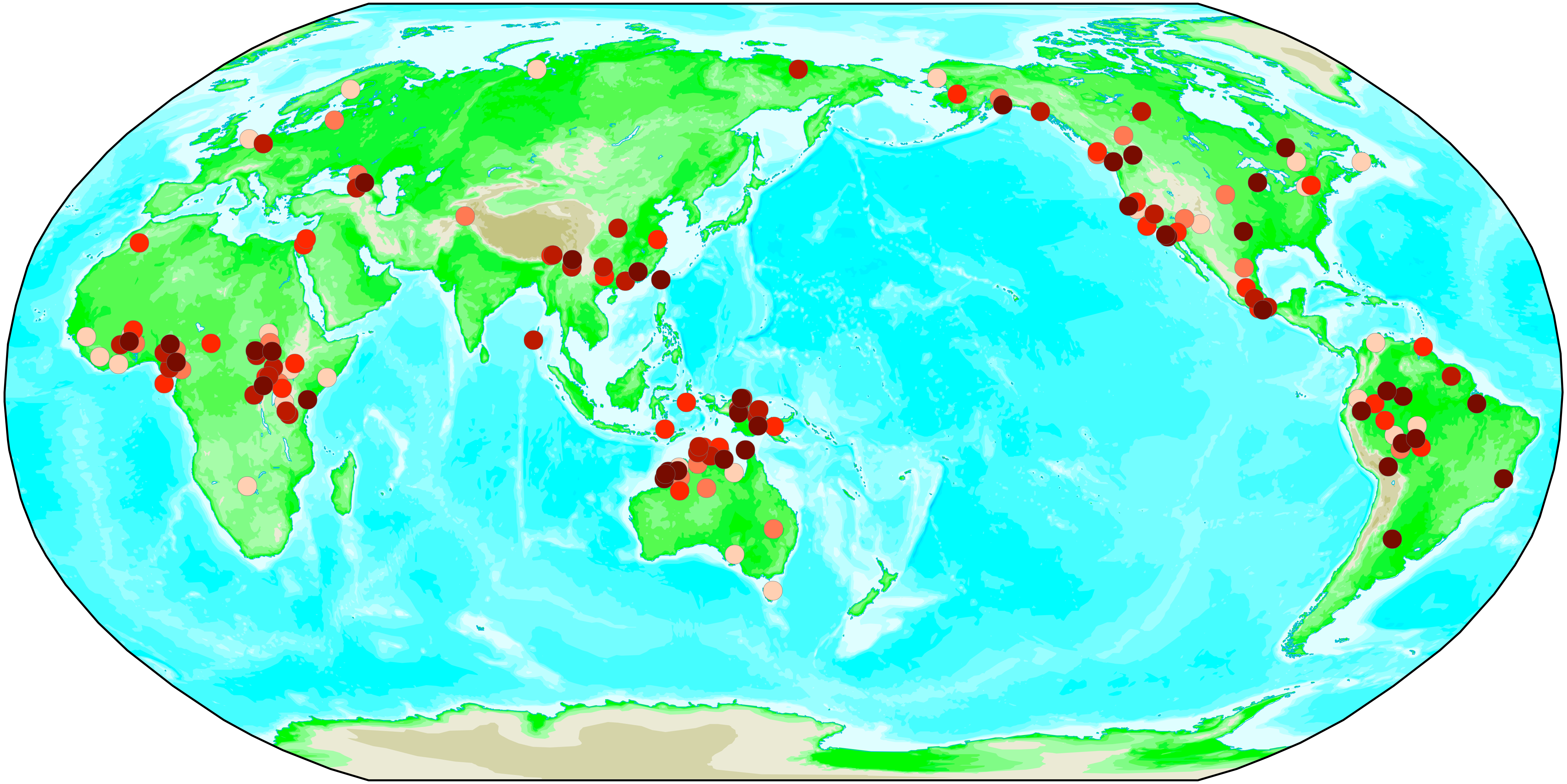
- This gives a value for the internal extremes (i.e. how extreme is a language within the given set of WALS languages)


WALS data with 1% and 5% extremes
Rarity Index
Number of features coded
−1
0
1
2
0
20
40
60
80
100
120
140

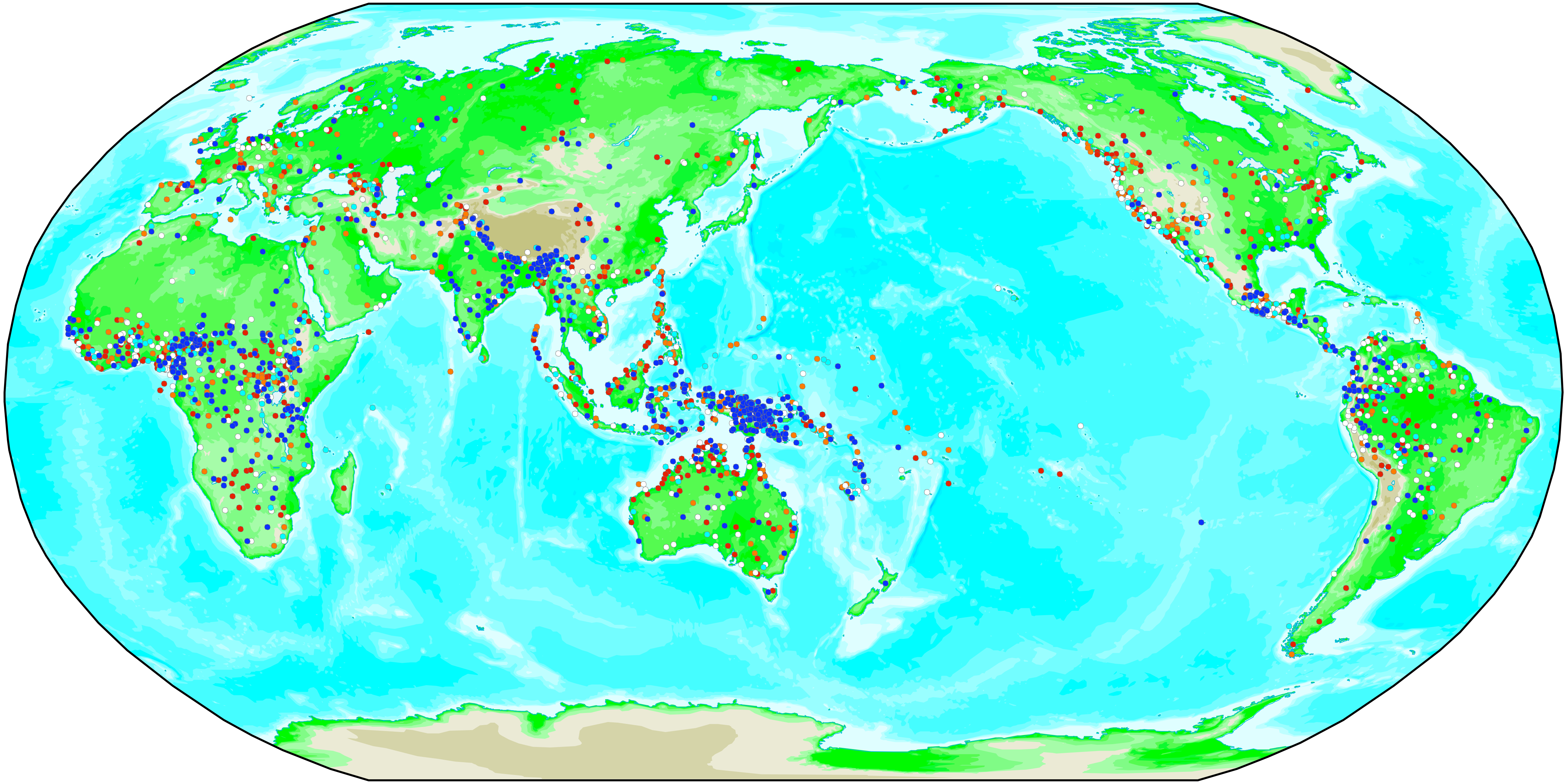
Highest *Mean Rarity*

Language	Family	Genus	Features Coded	Mean Rarity	%
Wari’	Chapacura-Wanhan	Chapacura-Wanhan	115	2.36	99.9
Dinka	Nilo-Saharan	Nilotic	45	3.45	99.9
Tiipay (Jamul)	Hokan	Yuman	44	3.76	99.9
Nuer	Nilo-Saharan	Nilotic	28	3.42	99.9
Karó (Arára)	Tupian	Tupi-Guaranì	24	6.16	99.9
Winnebago	Siouan	Siouan	7	11.37	99.9

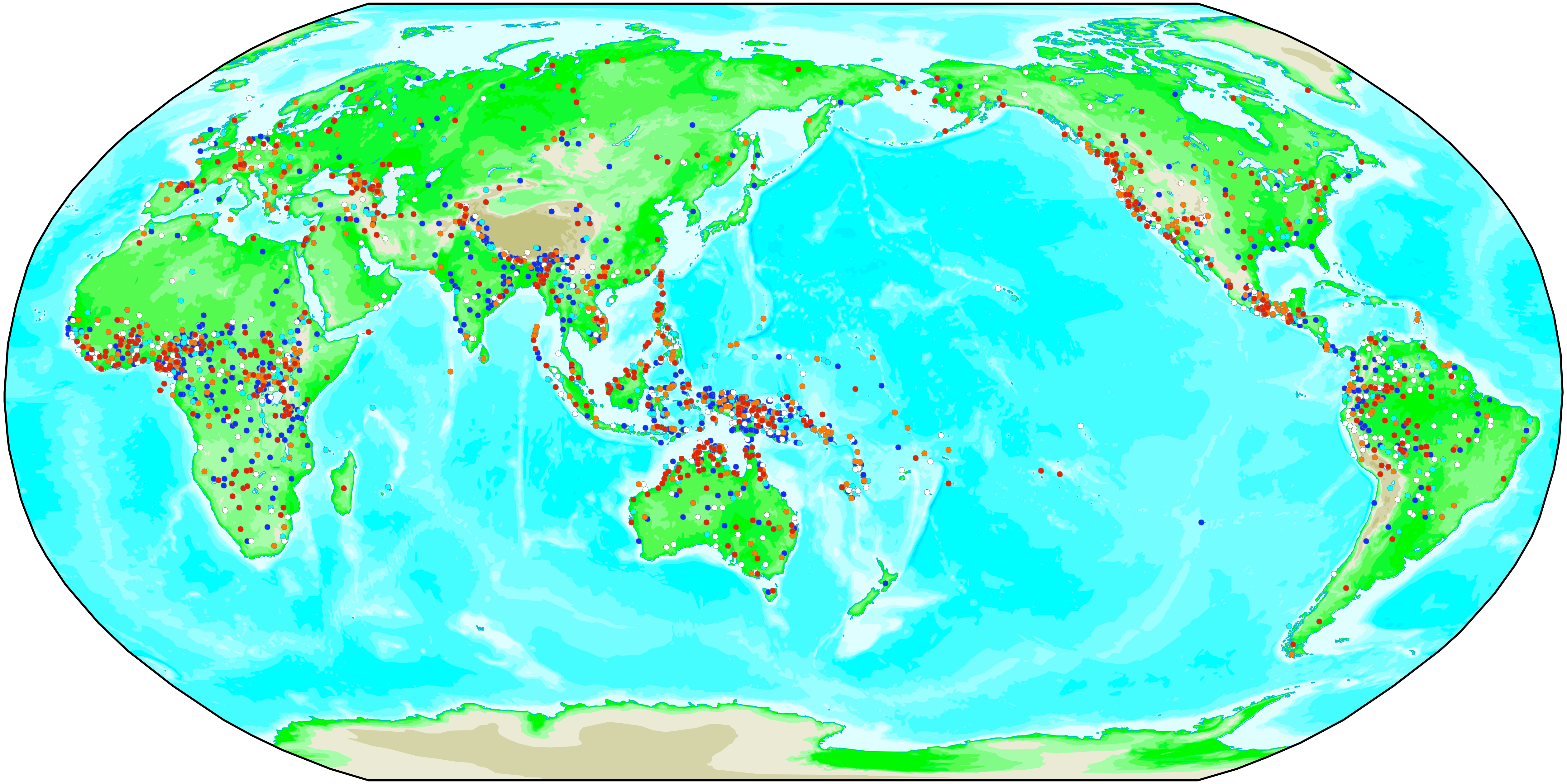
Top 5% by *Mean Rarity*



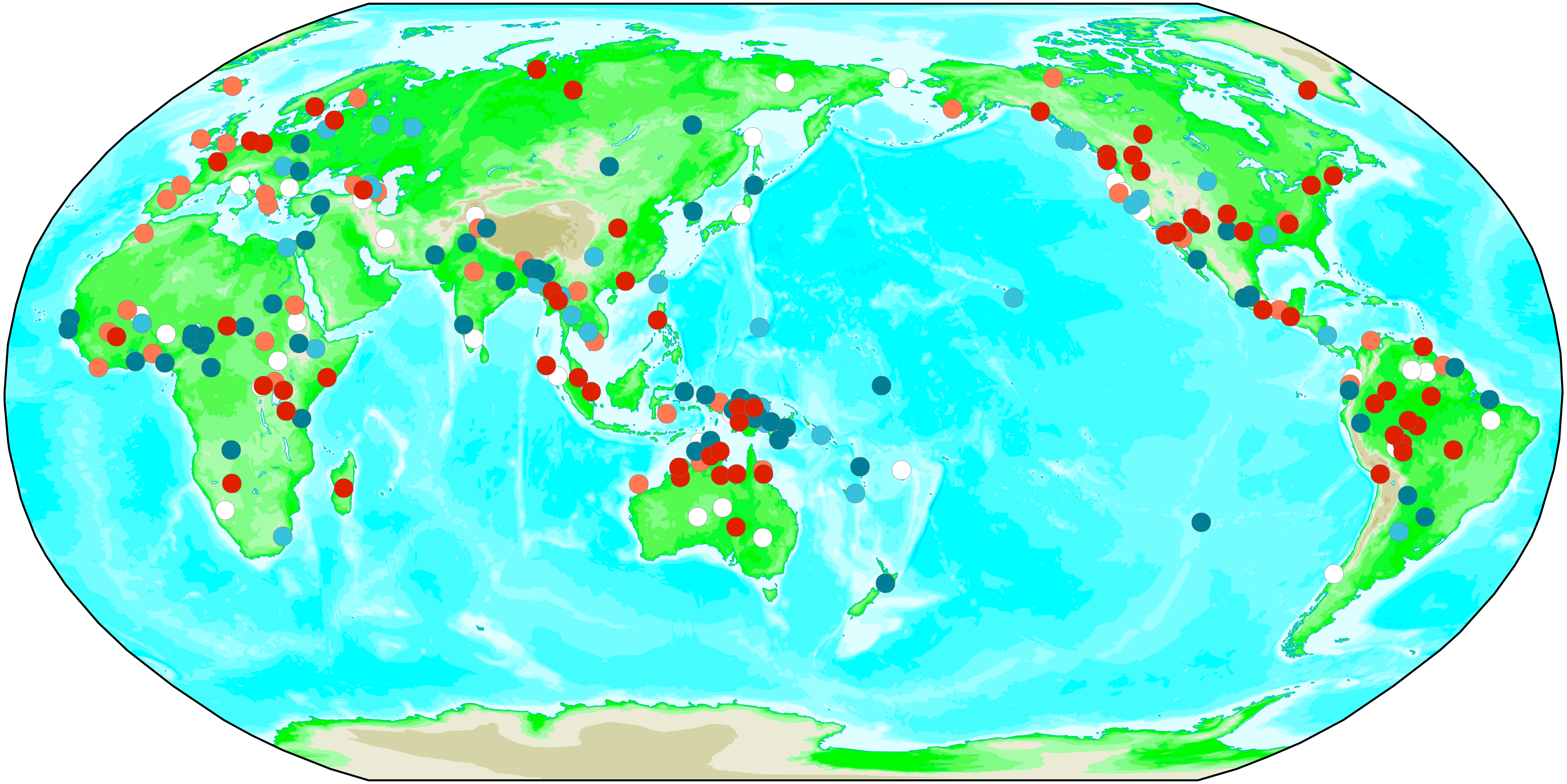
Mean Rarity of all languages
(red = rare, blue = common)



Mean Rarity of all languages, printed in reverse colour order (red = rare, blue = common)



All languages with more than 60 features coded for (red = rare, blue = common)



Rarity Index for a group

- How to evaluate the relative rarity of a group of languages?
- Simply taking the mean of all languages is too crude
- I decided to take a weighted mean of the extremity percentages
- The result is a mean of various percentages
- This is difficult to interpret, but it can be used for ranking

Rarity Index for a group

n = number of languages in a group

L_i = number of features coded for language *i*

$\%R_i$ = Relative position of mean of all Rarity Indices for language *i*

Weighted mean of $\%R_i$ by number of features coded:

$$\frac{\sum_{i=1}^{n} \log(L_i) \cdot (\%R)_i}{\sum_{i=1}^{n} \log(L_i)}$$

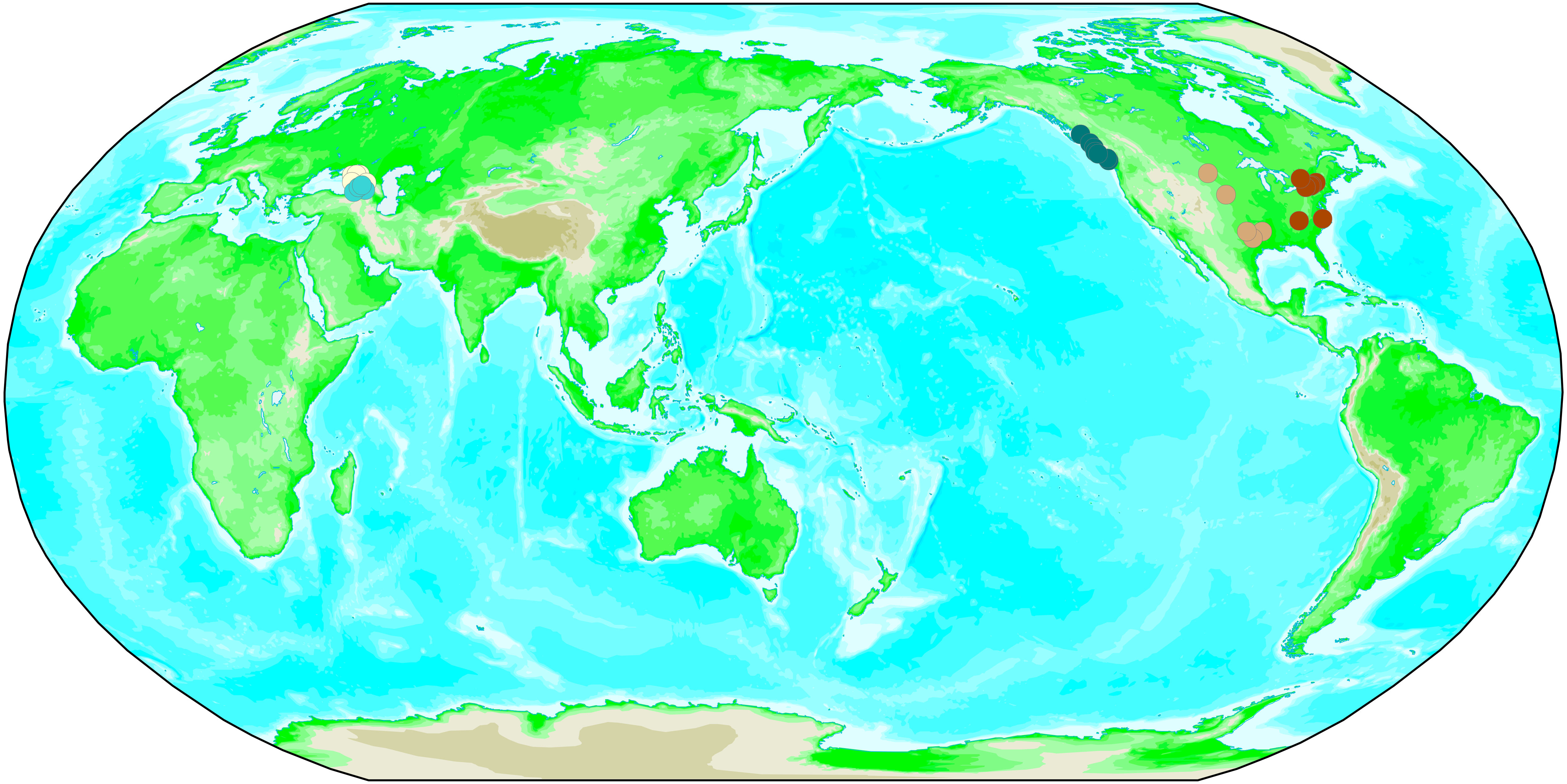
Genealogical groups

- Compute *Group Indices* for all Families and Genera as coded in the WALS
- Only groups with more than three members are shown, to be sure to get a group measure, and not an effect of an individual language

Top 5 Families

Family	Languages	%
NorthwestCaucasian	7	87.8
Kartvelian	4	83.7
Caddoan	5	82.2
Wakashan	7	80.2
Iroquoian	8	76.3

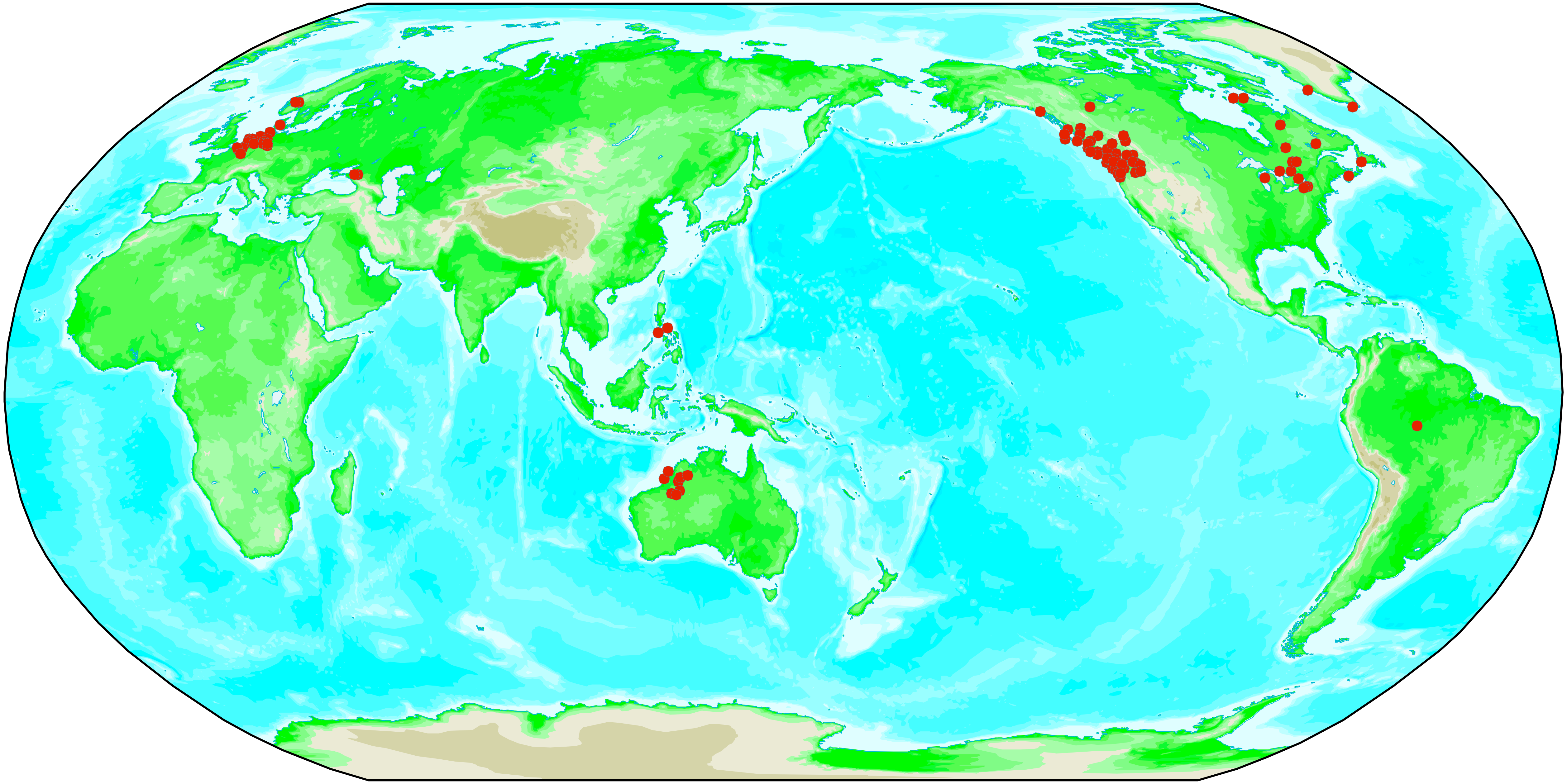
Top 5 Families



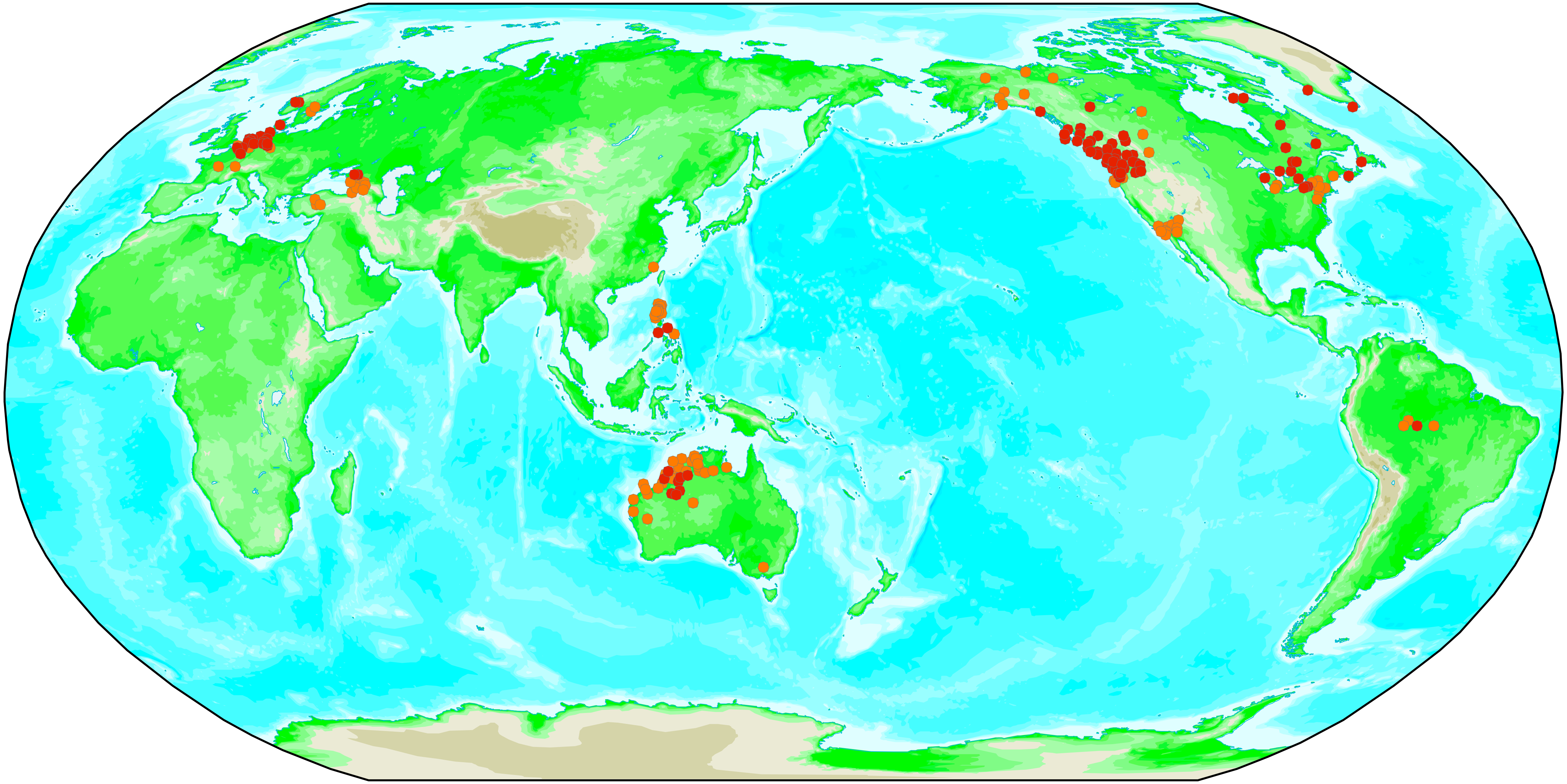
Areal groups

- For each language, take the 30 geographically nearest languages
- Compute *Group Indices* for the surrounding area of each language
- Such a measure should be definition be areally consistent, but it can indicate geographical centers of ‘rarity’

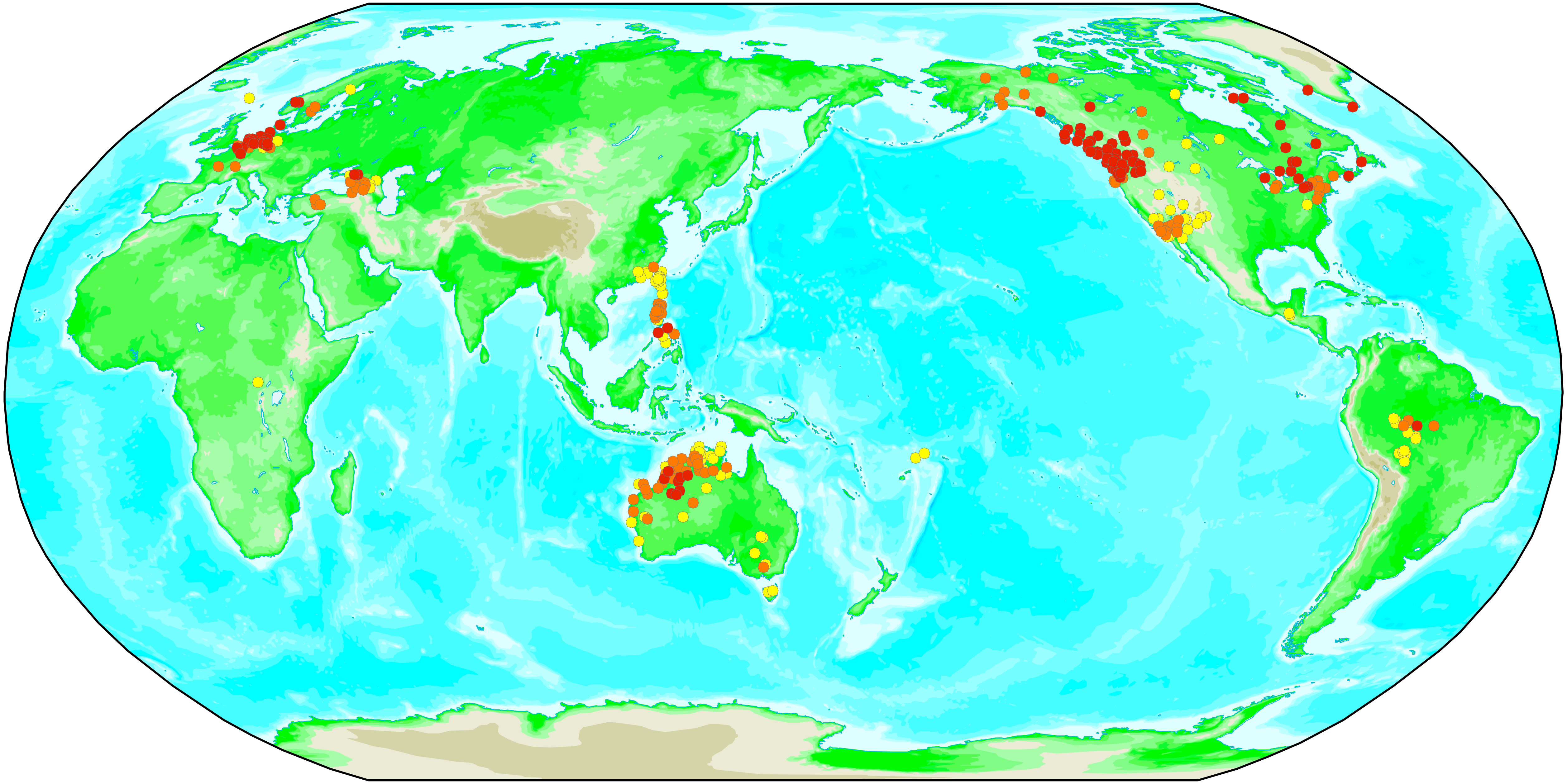
Top 100



Top 200

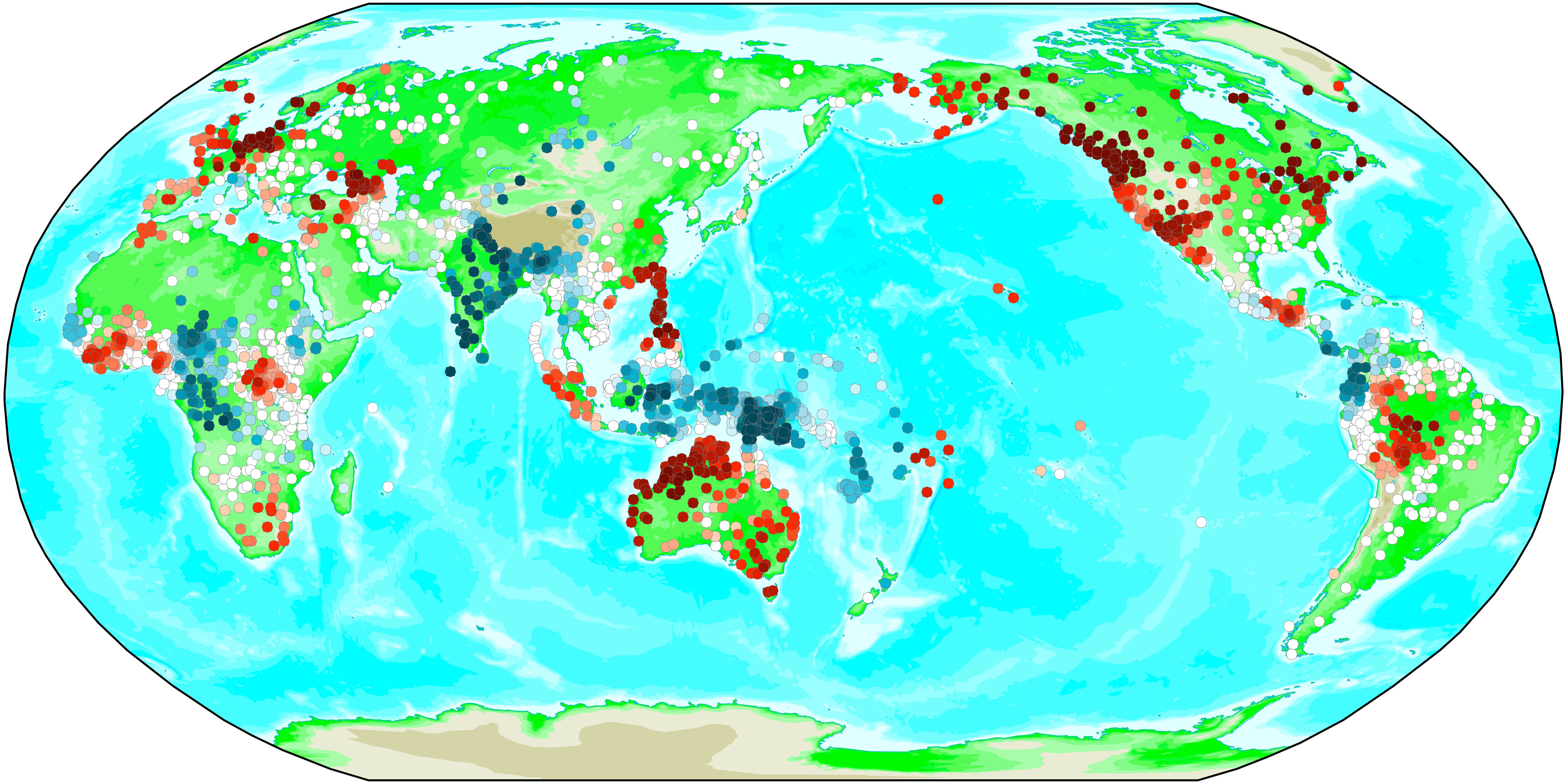


Top 300



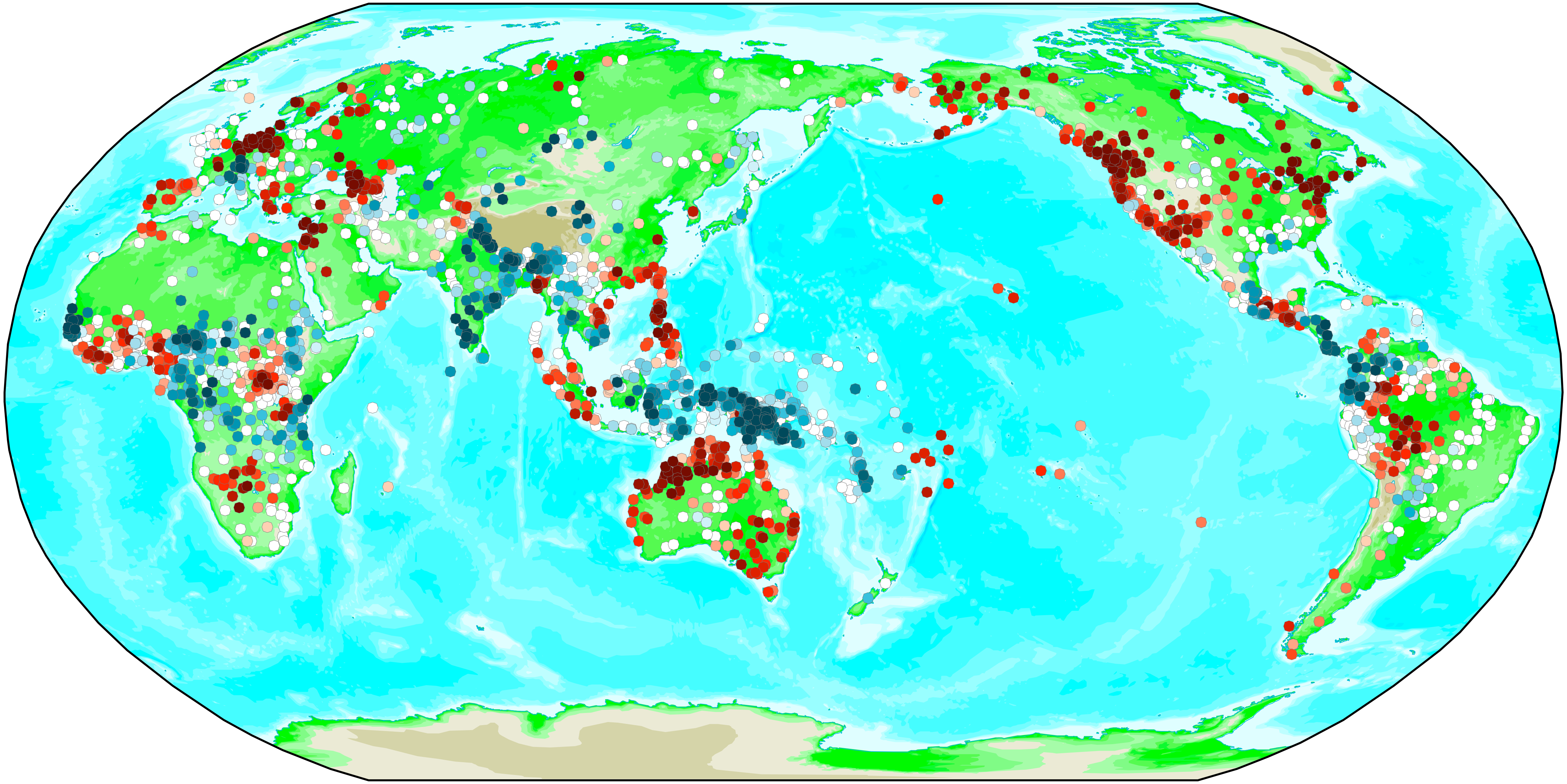
All languages

(red = rare, blue = common)



All languages

(only taking nearest 10 languages)



The End