Using WALS

Prospects of quantitative approaches
for linguistic typology

Wow !

WALS is just great

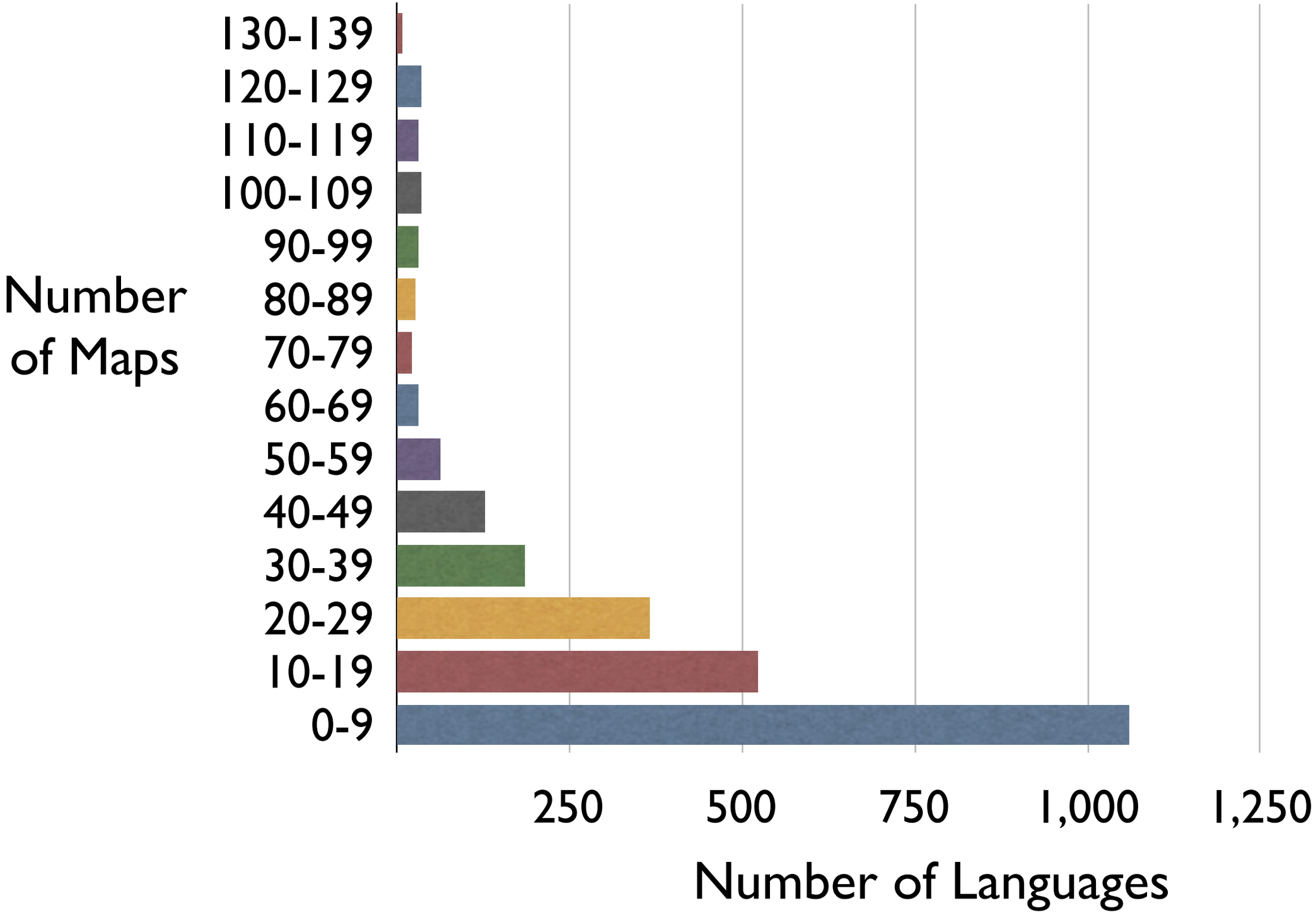
- Linguistic data is very expensive
  (estimated 6,000,000 EUR)
- but now we have so much data
- and so well organized !

Problems

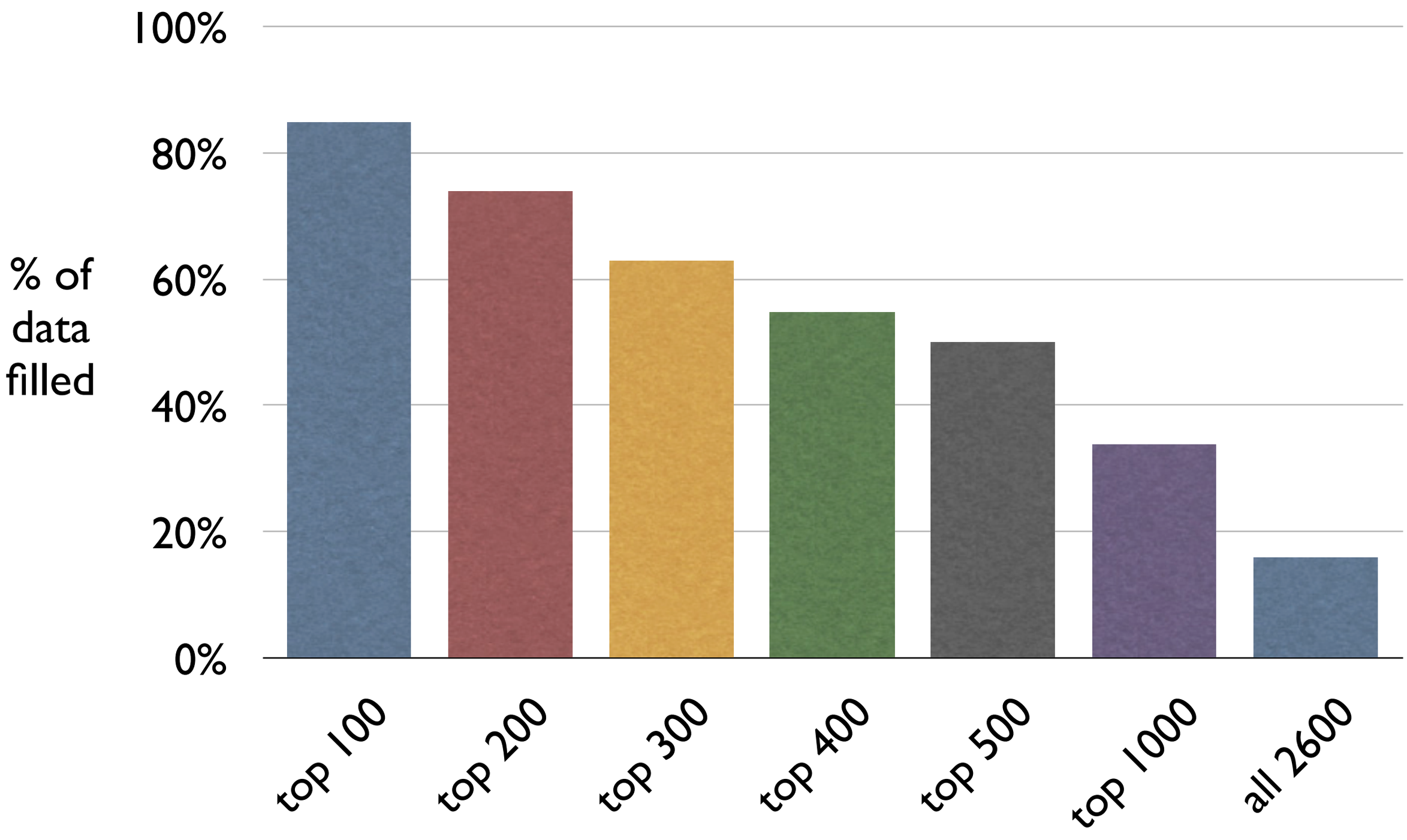
So much data ...

- 2,600 Languages
- 140 characteristics
- almost 60,000 datapoints
- wait a minute: 60,000/2,600*140 = 0.165
- the datatable is only 16.5 % filled

Most languages occur in few maps



Choosing the best languages



Reliability

- Latvian was checked (by B. Wälchli)
- 109 coding point in WALS
- 2 ‘technical’ errors (= 1.8 %)
- 5 ‘interpretative’ errors (= 4.6 %)

Coding problems

- Some maps combine independent dimensions: they have to be recoded
- Unwanted categories marking ‘leftovers’ have to be recoded
- Many maps have (hidden) definitional dependencies to other maps
- There is much structure in the data that is not coded

Exploration

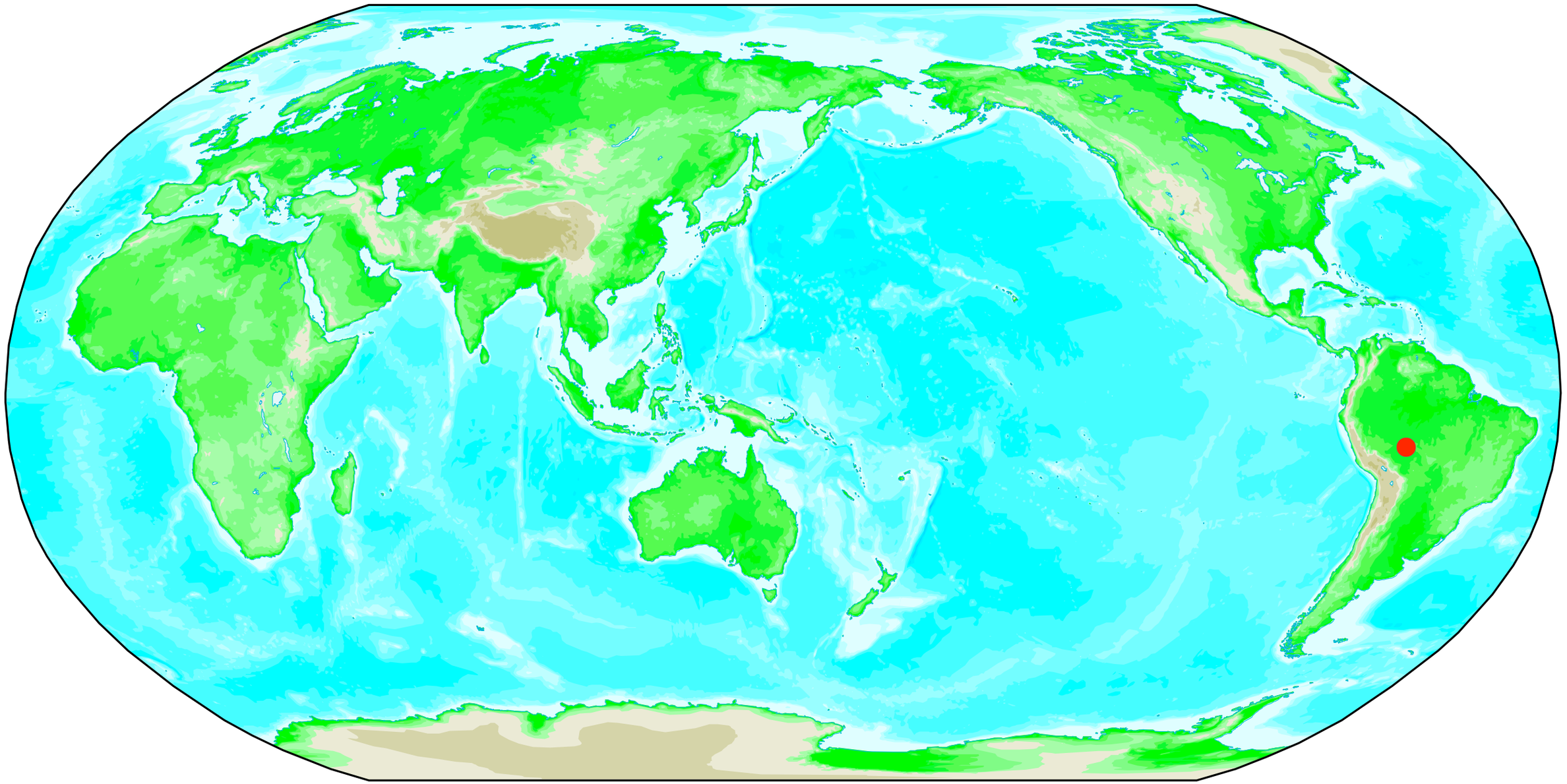
Distribution of **rare** characteristics

And the winners are:

In the category:

‘Most Unusual Individual Language’

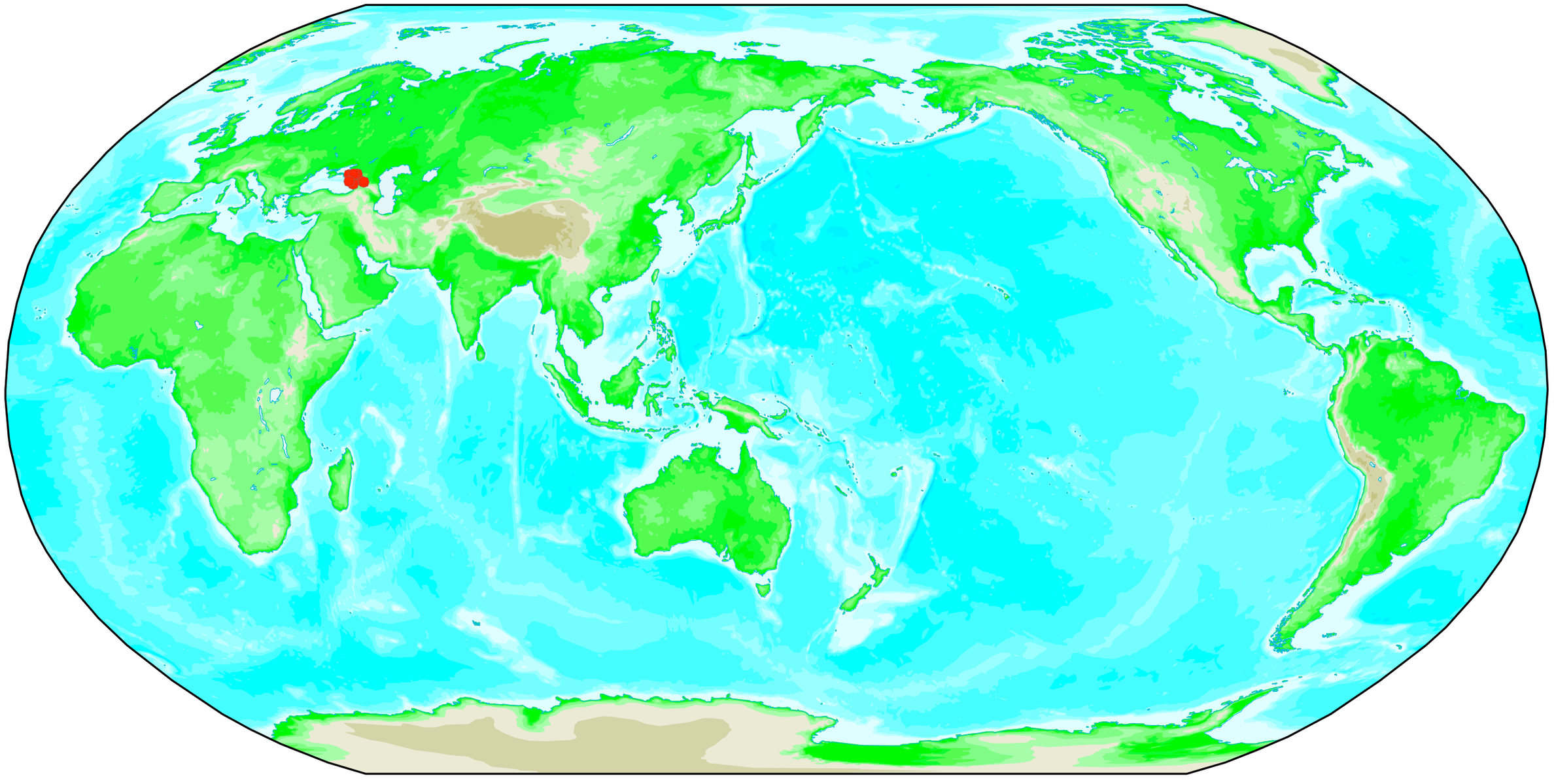
Wari’



In the category:

‘Most Unusual Genealogical Group’

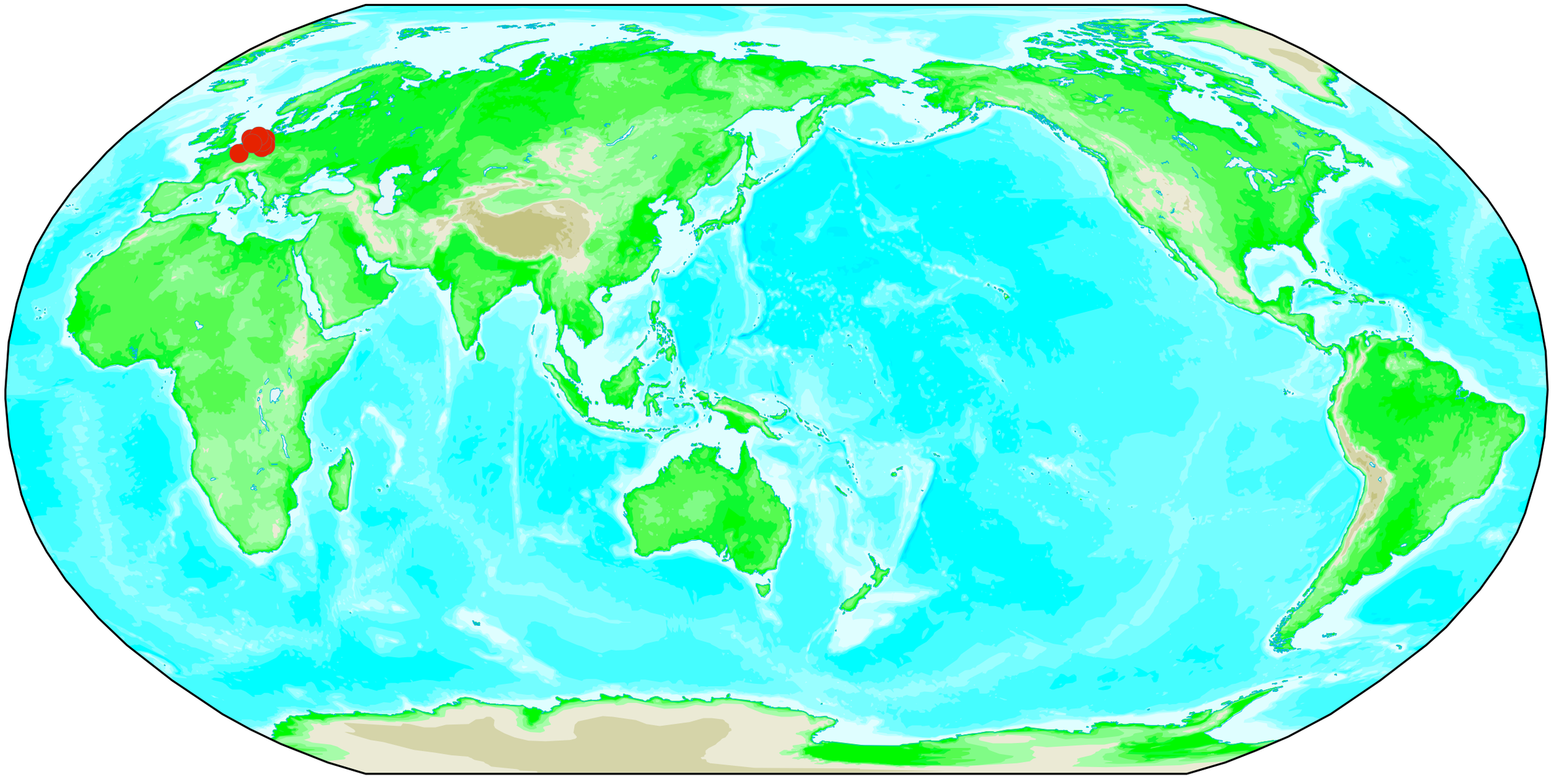
Northwest Caucasian



In the category:

‘Most Unusual Geographical Area’

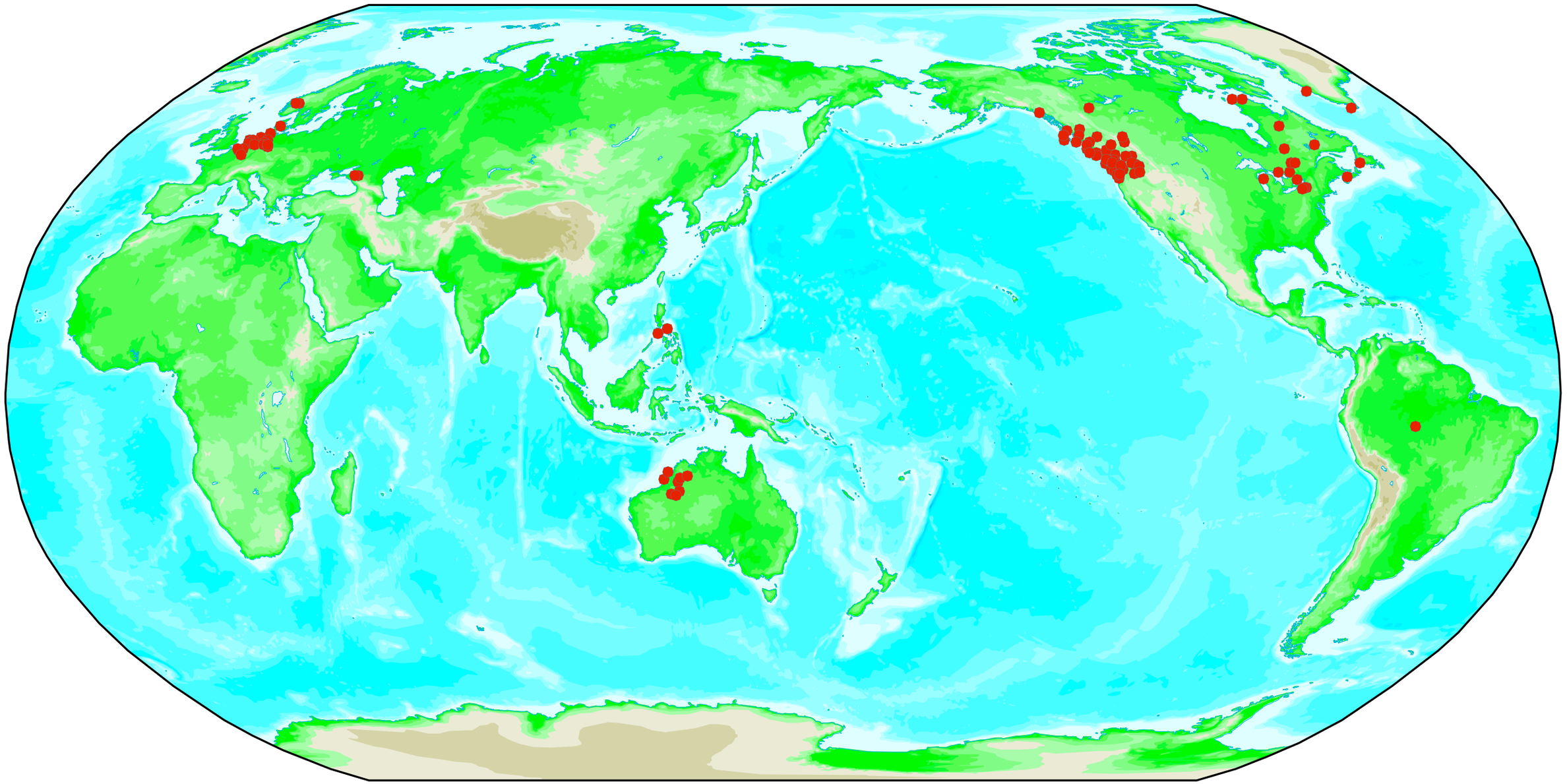
Northwest Continental Europe



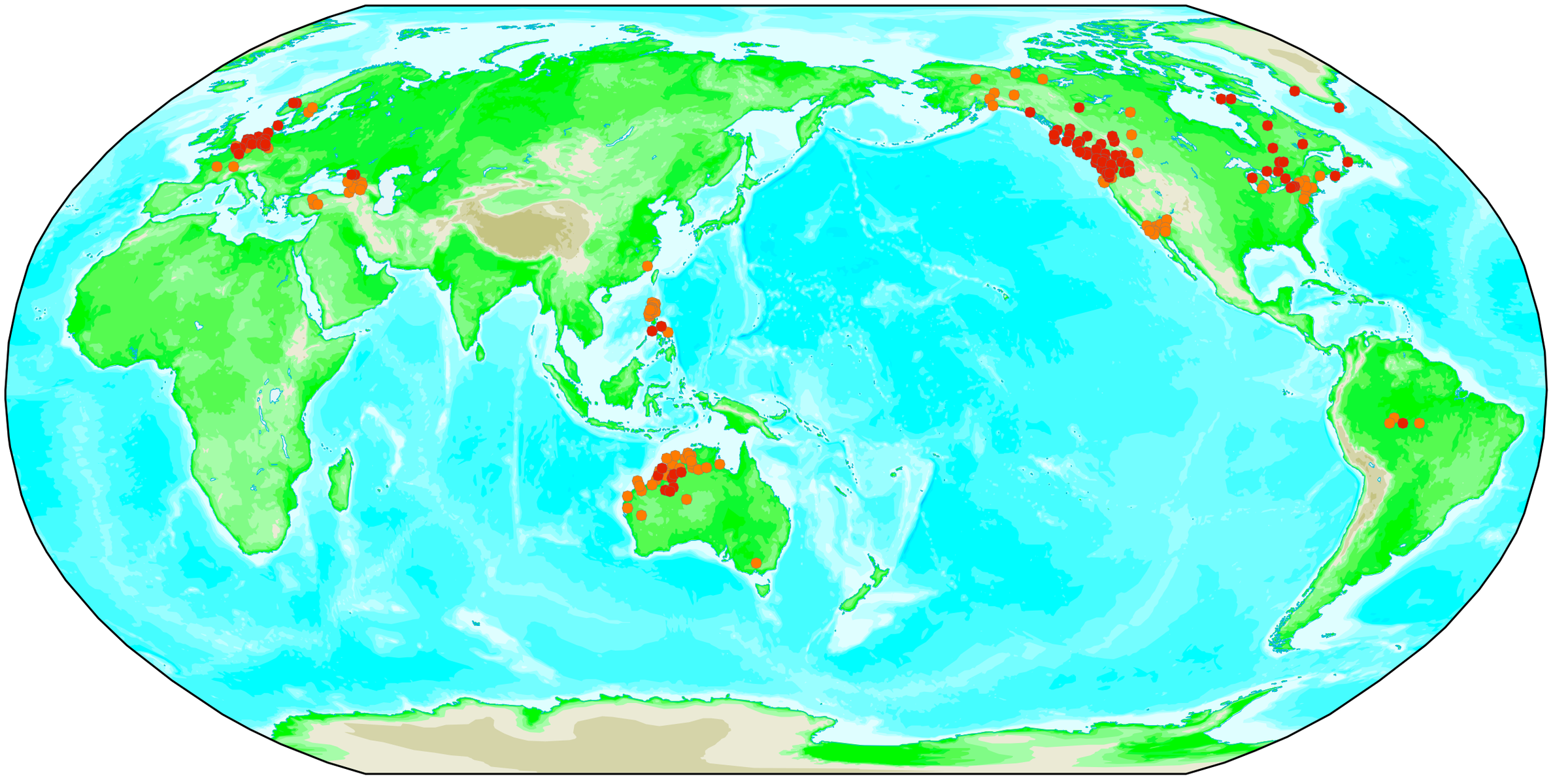
Areal groups

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

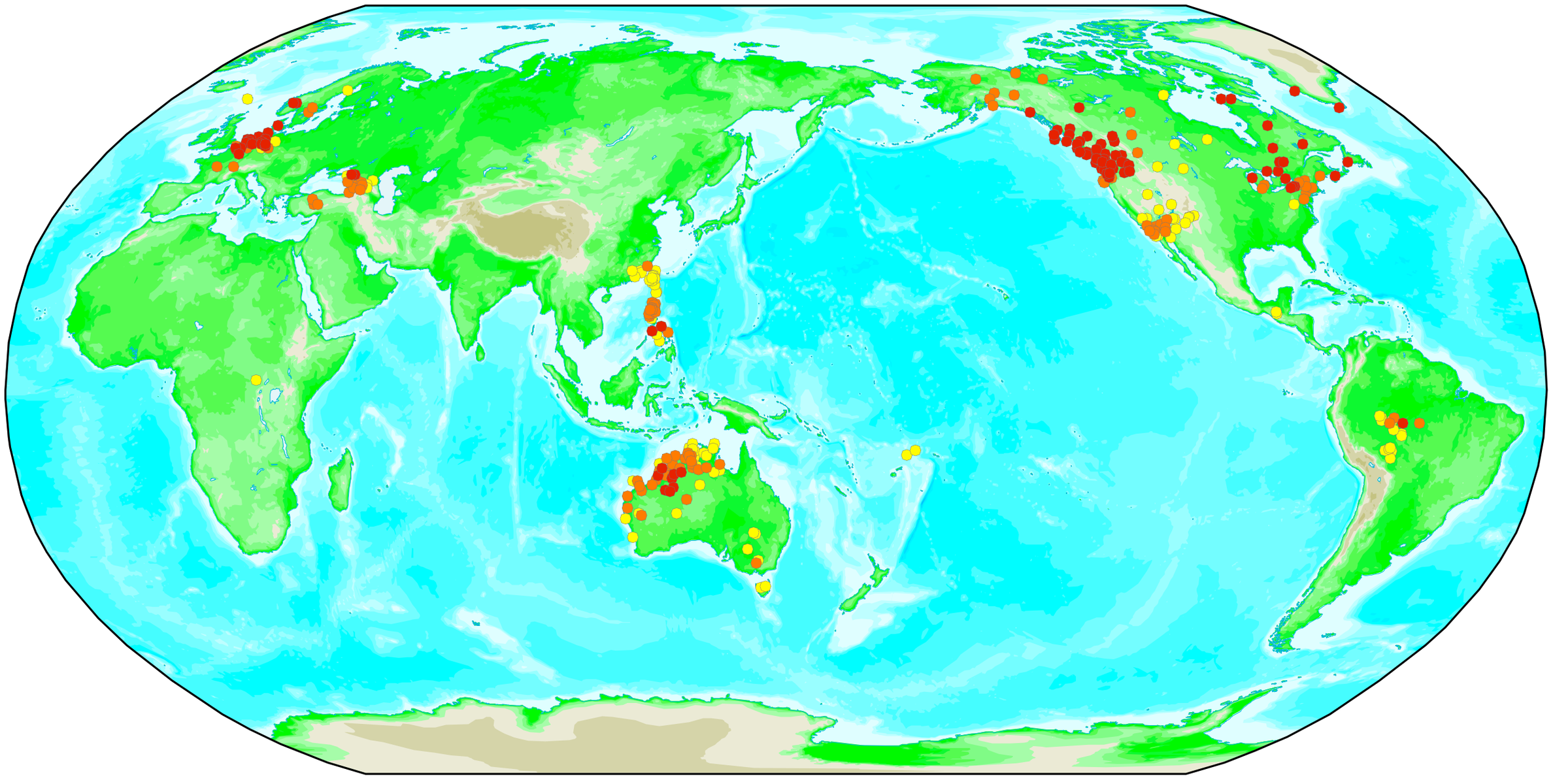
Top 100



Top 200

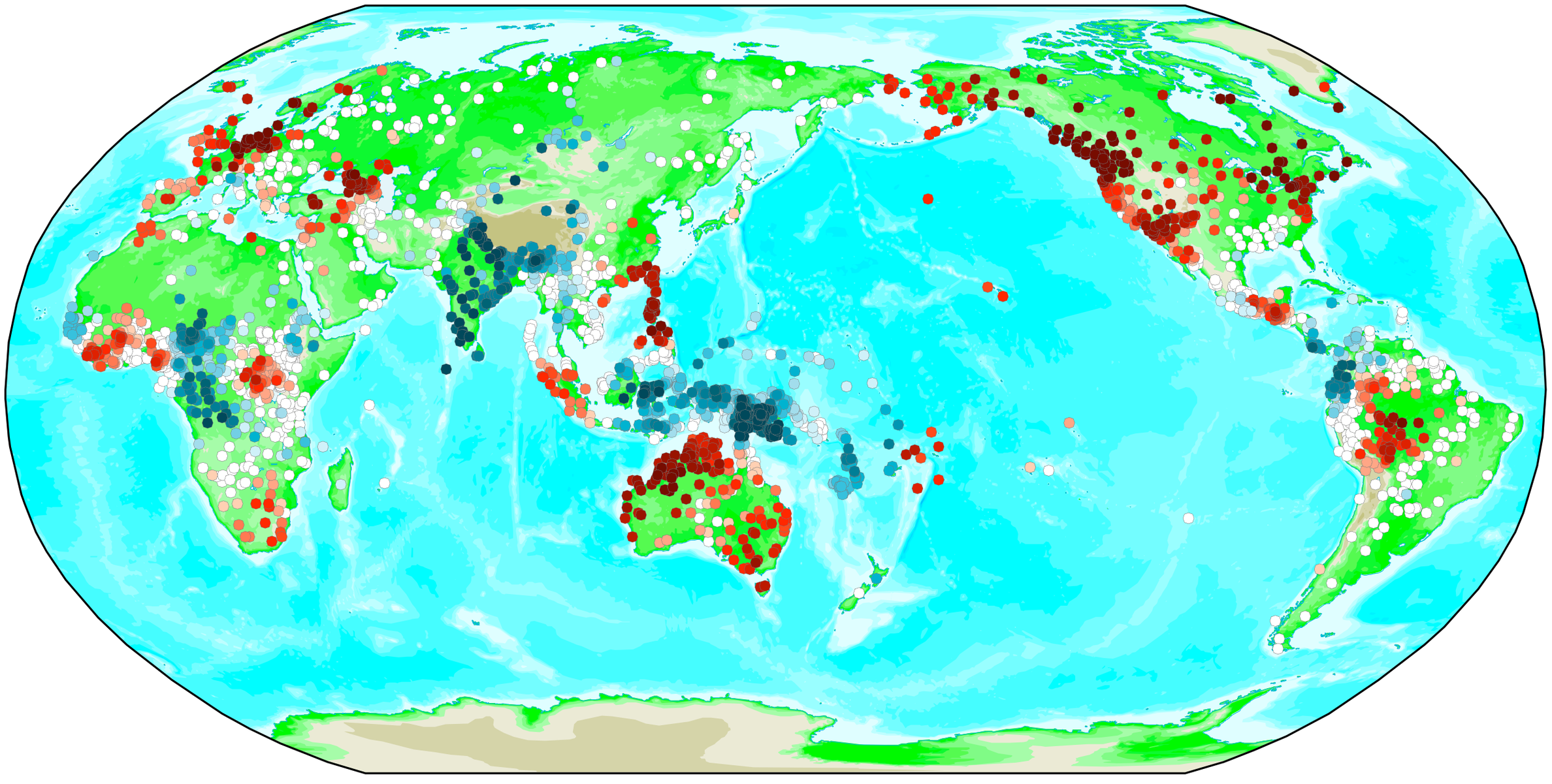


Top 300

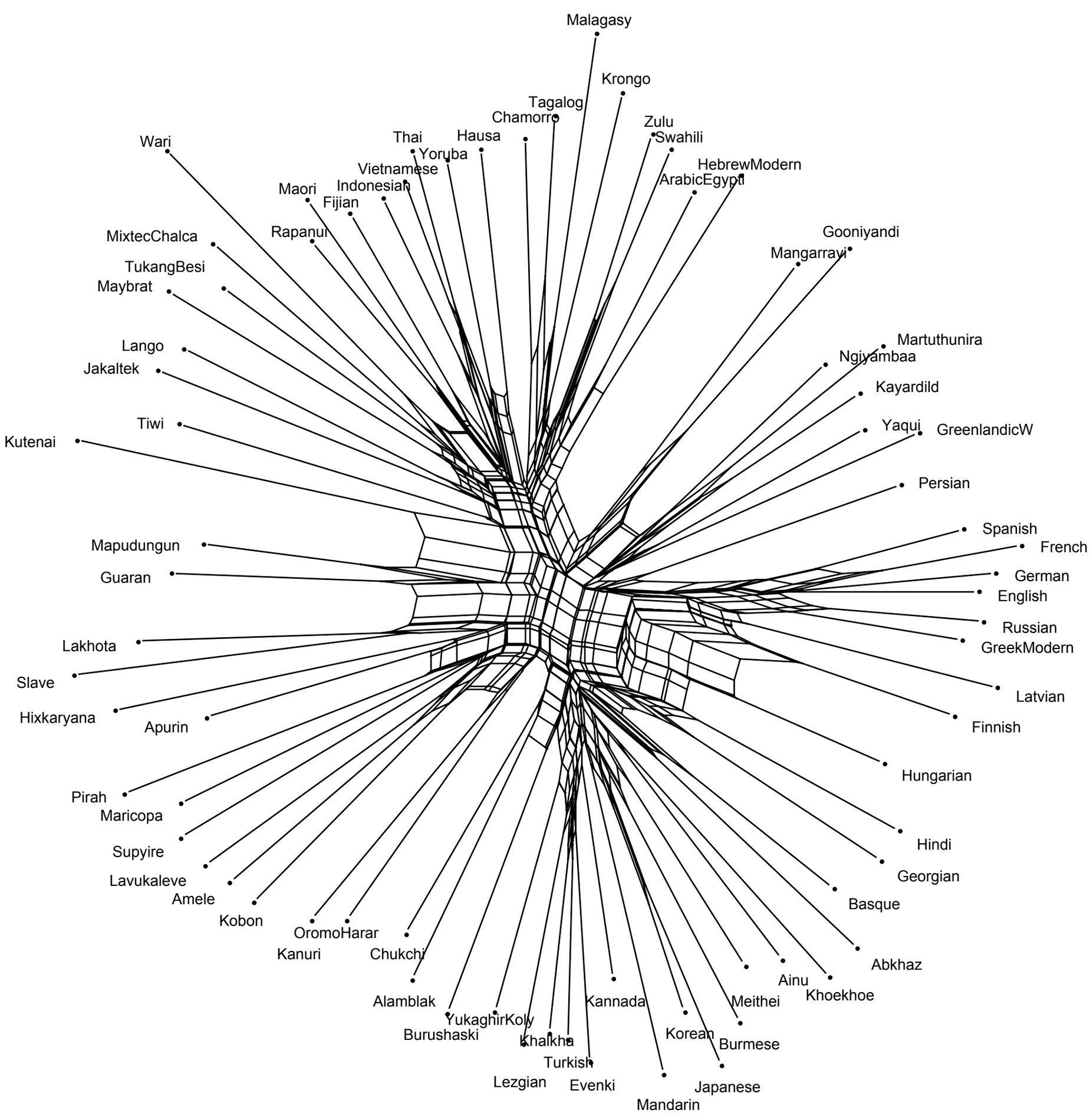


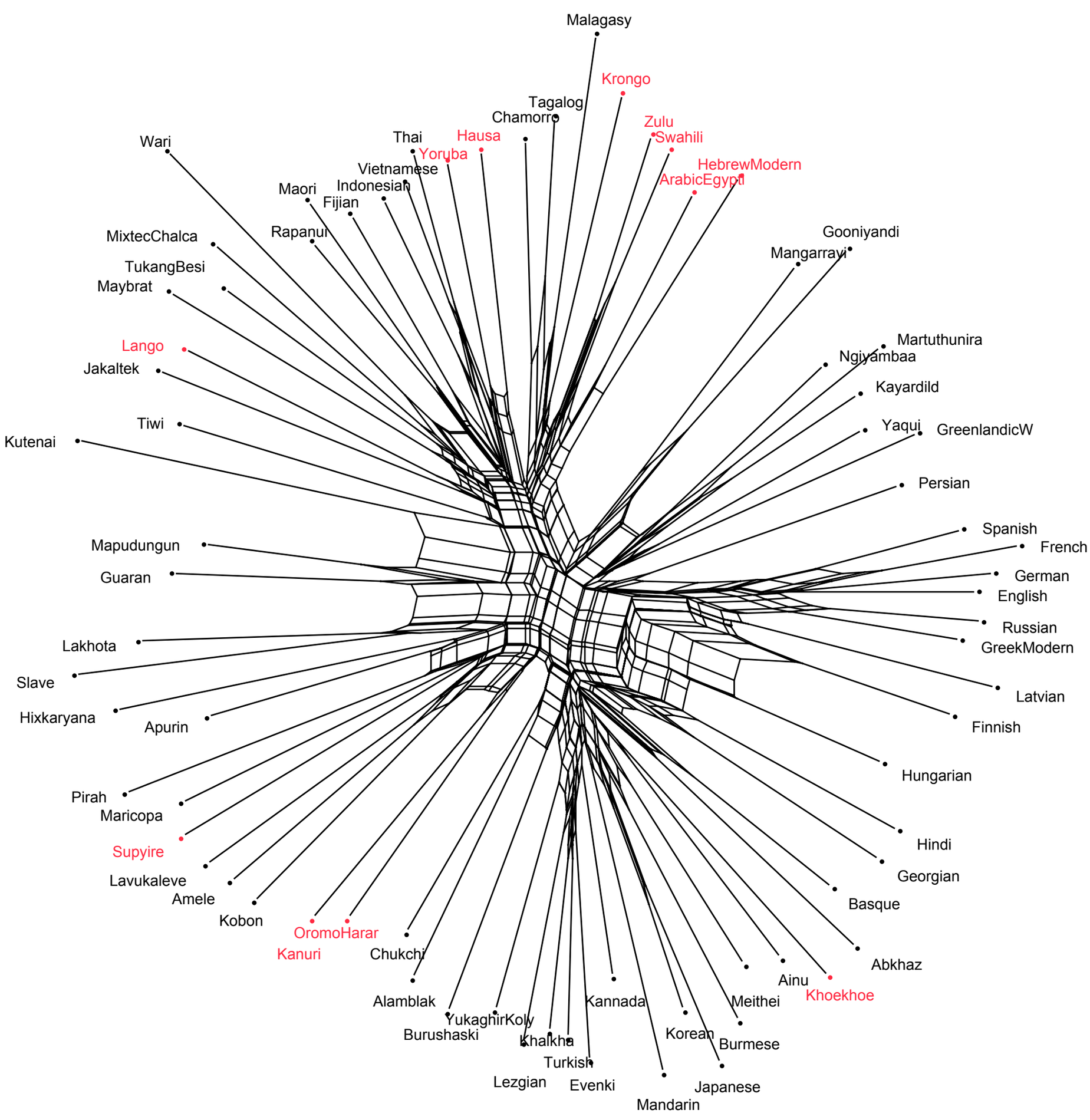
All languages

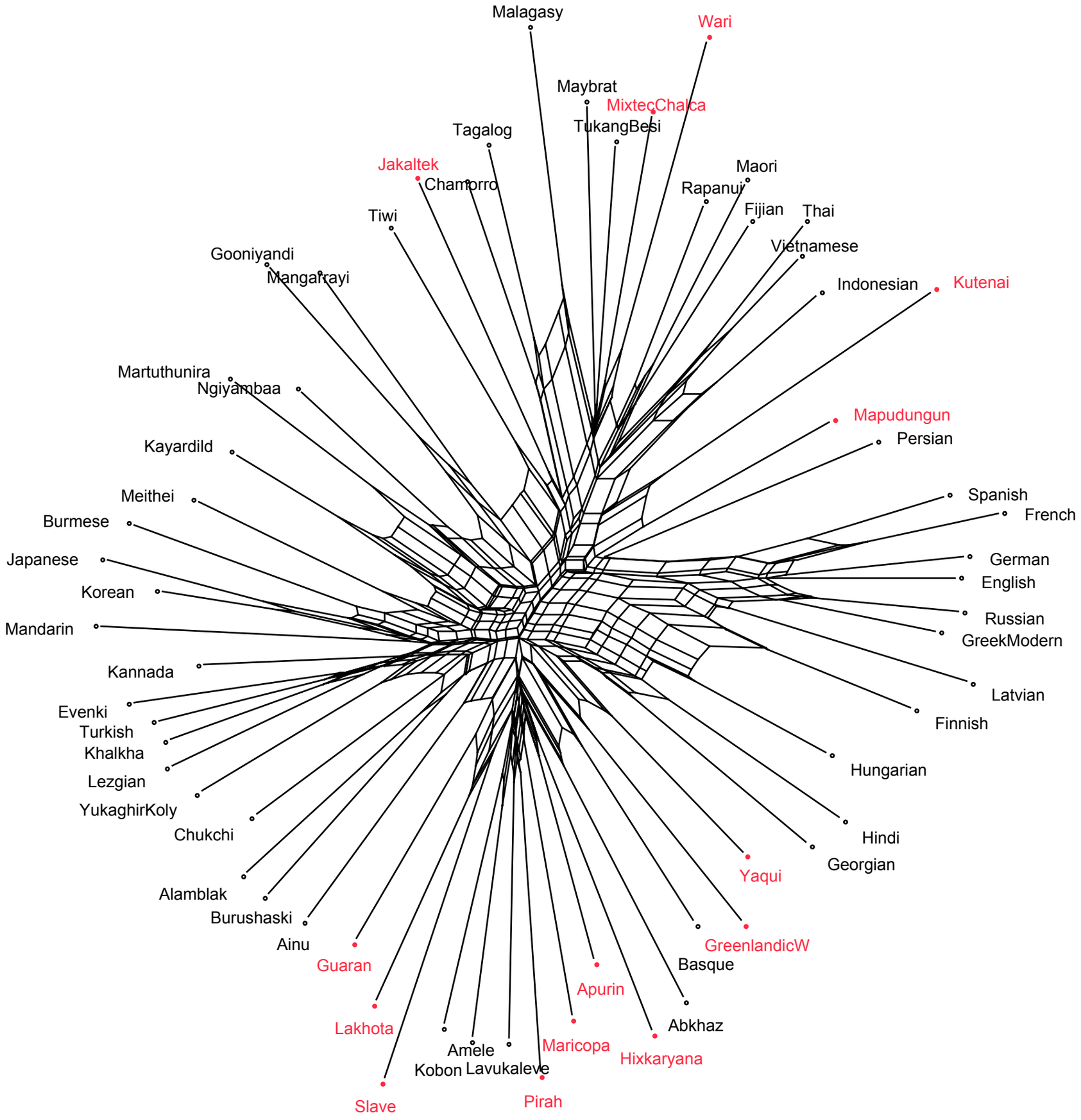
(red = rare, blue = common)

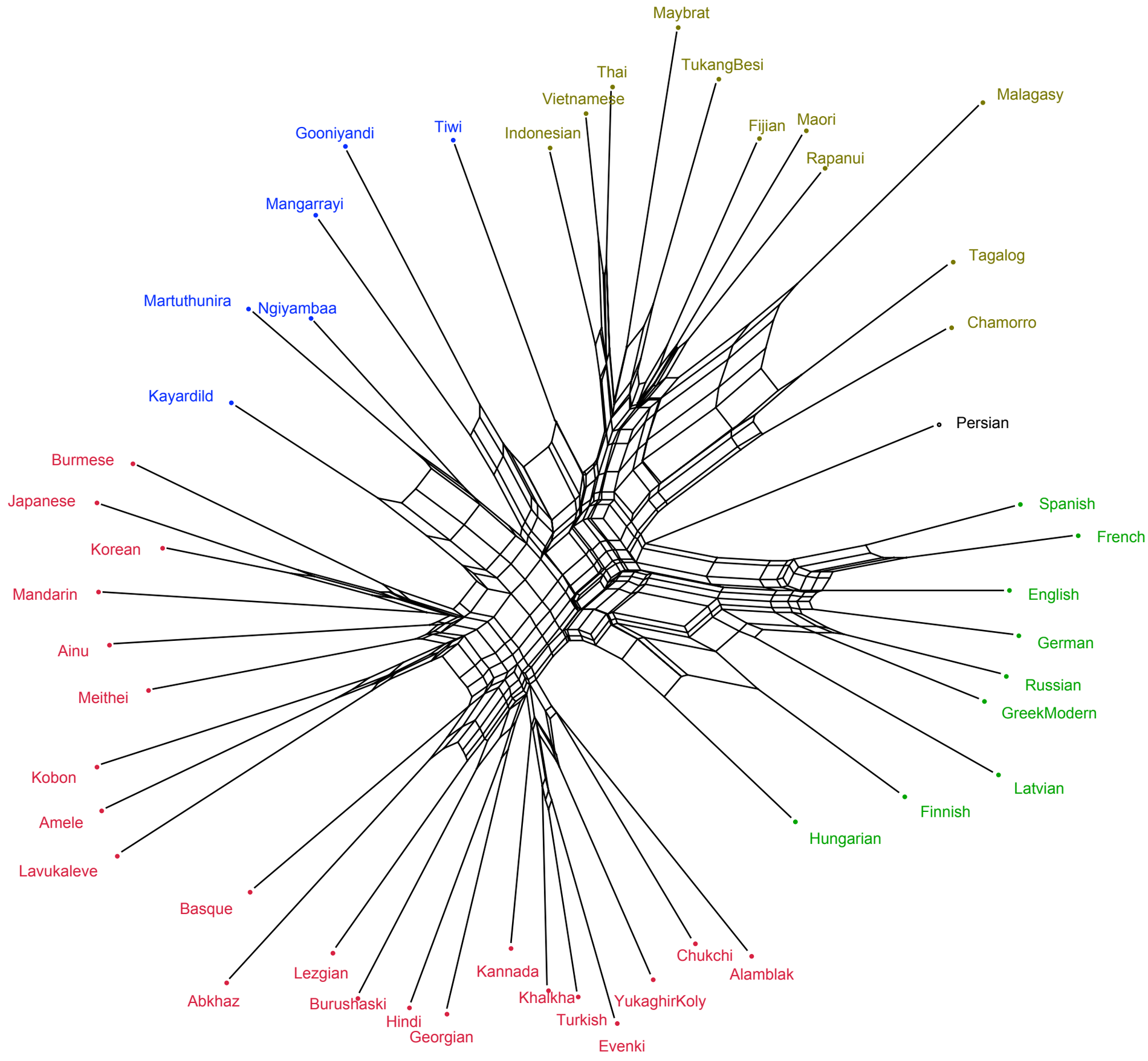


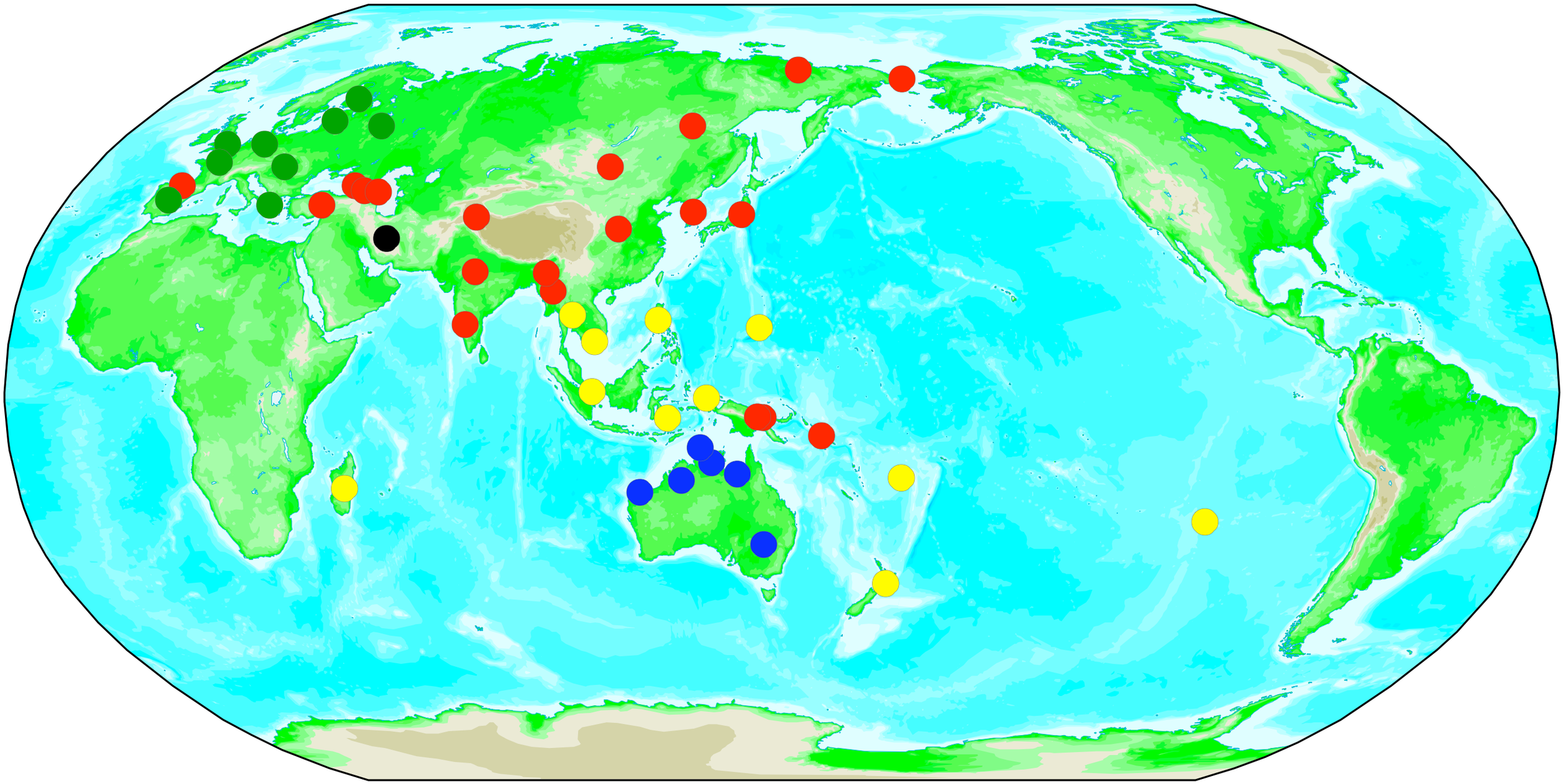
How **tree-like** is the WALS-data?


Malagasy
Krongo
Tagalog
Chamorro
Zulu
Swahili
Thai
Hausa
Yoruba
Wari
Vietnamese
HebrewModern
ArabicEgypti
Maori
Indonesian
Fijian
Rapanui
Gooniyandi
MixtecChalca
Mangarrayi
TukangBesi
Maybrat
Martuthunira
Ngiyambaa
Lango
Jakaltek
Kayardild
Yaqui
GreenlandicW
Tiwi
Kutenai
Persian
Spanish
Mapudungun
French
Guaran
German
English
Russian
Lakhota
GreekModern
Slave
Latvian
Hixkaryana
Apurin
Finnish
Hungarian
Pirah
Maricopa
Hindi
Supyire
Georgian
Lavukaleve
Basque
Amele
Kobon
Abkhaz
OromoHarar
Kanuri
Chukchi
Ainu
Khoekhoe
Alamblak
Kannada
Meithei
YukaghirKoly
Burushaski
Korean
Khalkha
Burmese
Turkish
Lezgian
Evenki
Japanese
Mandarin

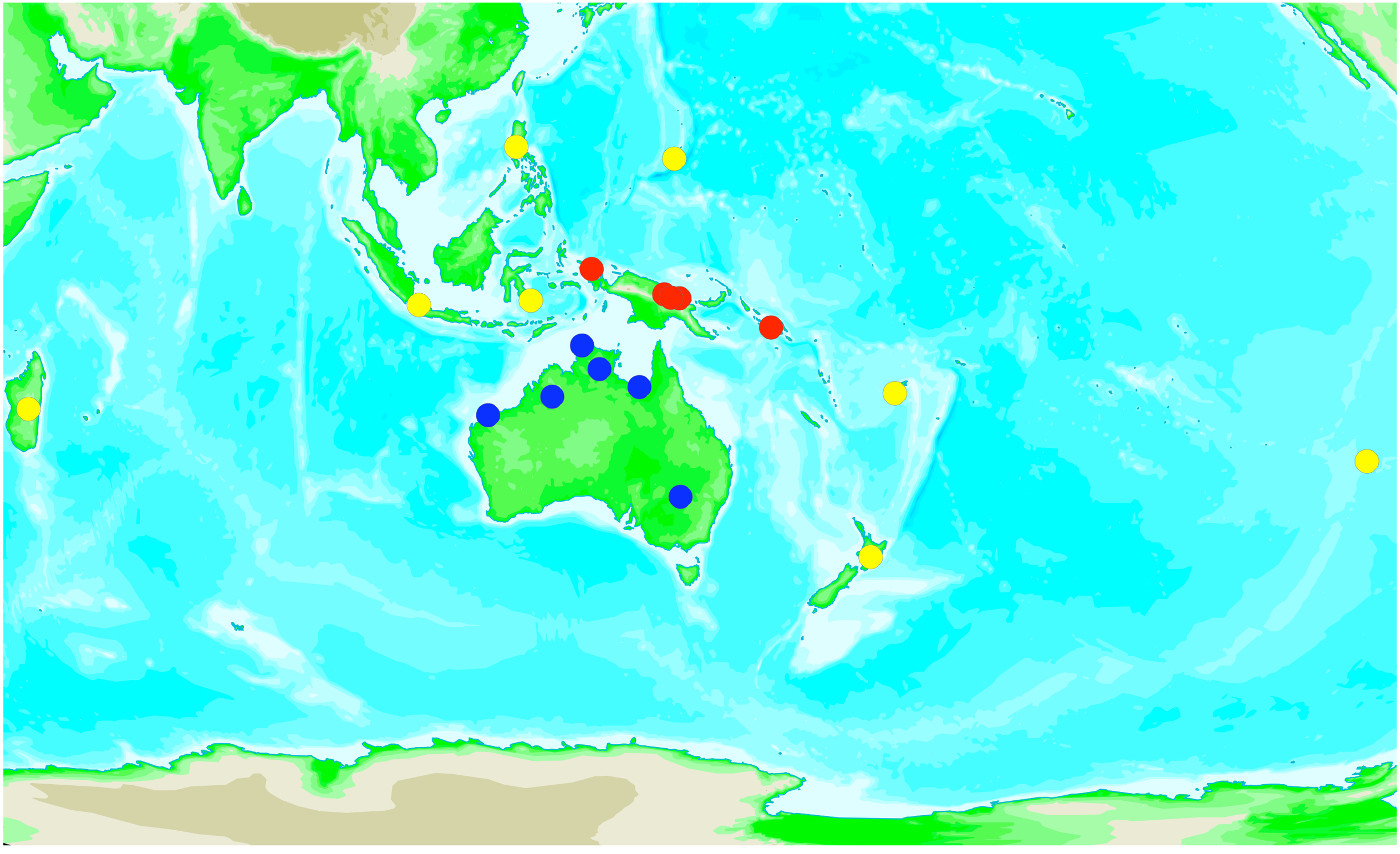

Malagasy
Krongo
Tagalog
Chamorro
Zulu
Swahili
Thai
Hausa
Yoruba
Wari
Vietnamese
HebrewModern
ArabicEgypt
Maori
Indonesian
Fijian
Rapanui
Gooniyandi
MixtecChalca
Mangarrayi
TukangBesi
Maybrat
Martuthunira
Lango
Ngiyambaa
Jakaltek
Kayardild
Tiwi
Yaqui
GreenlandicW
Kutenai
Persian
Spanish
Mapudungun
French
Guaran
German
English
Russian
Lakhota
GreekModern
Slave
Latvian
Hixkaryana
Finnish
Apurin
Hungarian
Pirah
Hindi
Maricopa
Georgian
Supyire
Basque
Lavukaleve
Abkhaz
Amele
Ainu
Kobon
Khoekhoe
OromoHarar
Kannada
Kanuri
Chukchi
Meithei
Alamblak
Korean
YukaghirKoly
Burushaski
Burmese
Khalkha
Turkish
Lezgian
Evenki
Japanese
Mandarin


Malagasy
Wari
Maybrat
MixtecChalca
TukangBesi
Tagalog
Jakaltek
Chamorro
Maori
Rapanui
Fijian
Thai
Tiwi
Vietnamese
Gooniyandi
Mangarrayi
Indonesian
Kutenai
Martuthunira
Ngiyambaa
Mapudungun
Persian
Kayardild
Meithei
Spanish
Burmese
French
Japanese
German
Korean
English
Russian
Mandarin
GreekModern
Kannada
Latvian
Finnish
Evenki
Turkish
Khalkha
Hungarian
Lezgian
YukaghirKoly
Hindi
Chukchi
Georgian
Yaqui
Alamblak
Burushaski
GreenlandicW
Ainu
Guaran
Basque
Apurin
Lakhota
Abkhaz
Amele
Maricopa
Kobon
Lavukaleve
Hixkaryana
Slave
Pirah

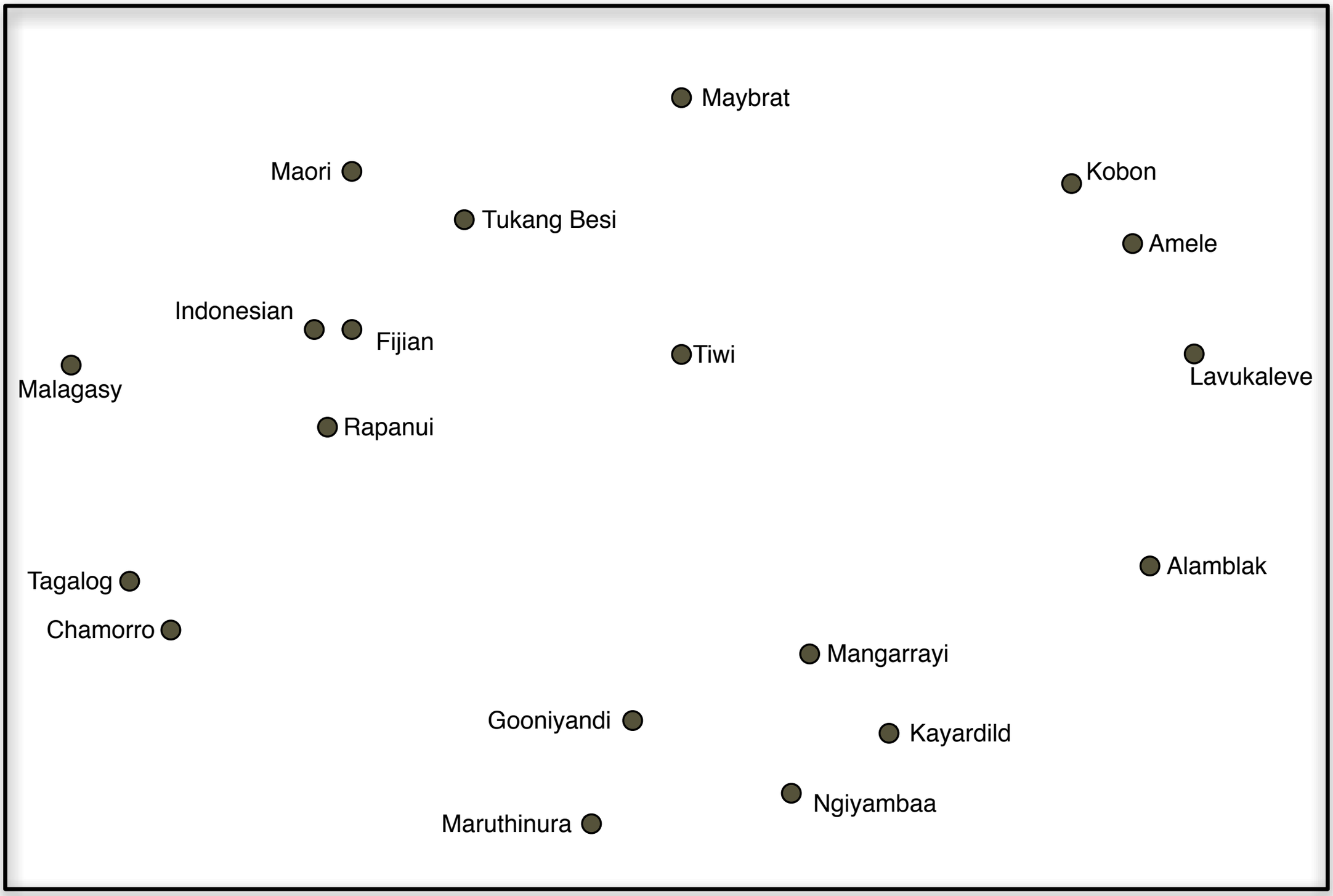

Maybrat
TukangBesi
Thai
Vietnamese
Malagasy
Gooniyandi
Tiwi
Indonesian
Fijian
Maori
Rapanui
Mangarrayi
Tagalog
Martuthunira
Ngiyambaa
Chamorro
Kayardild
Persian
Burmese
Spanish
Japanese
French
Korean
English
Mandarin
German
Ainu
Russian
Meithei
GreekModern
Kobon
Latvian
Amele
Finnish
Lavukaleve
Hungarian
Basque
Chukchi
Kannada
Lezgian
Alamblak
Abkhaz
Khalkha
YukaghirKoly
Burushaski
Hindi
Turkish
Georgian
Evenki



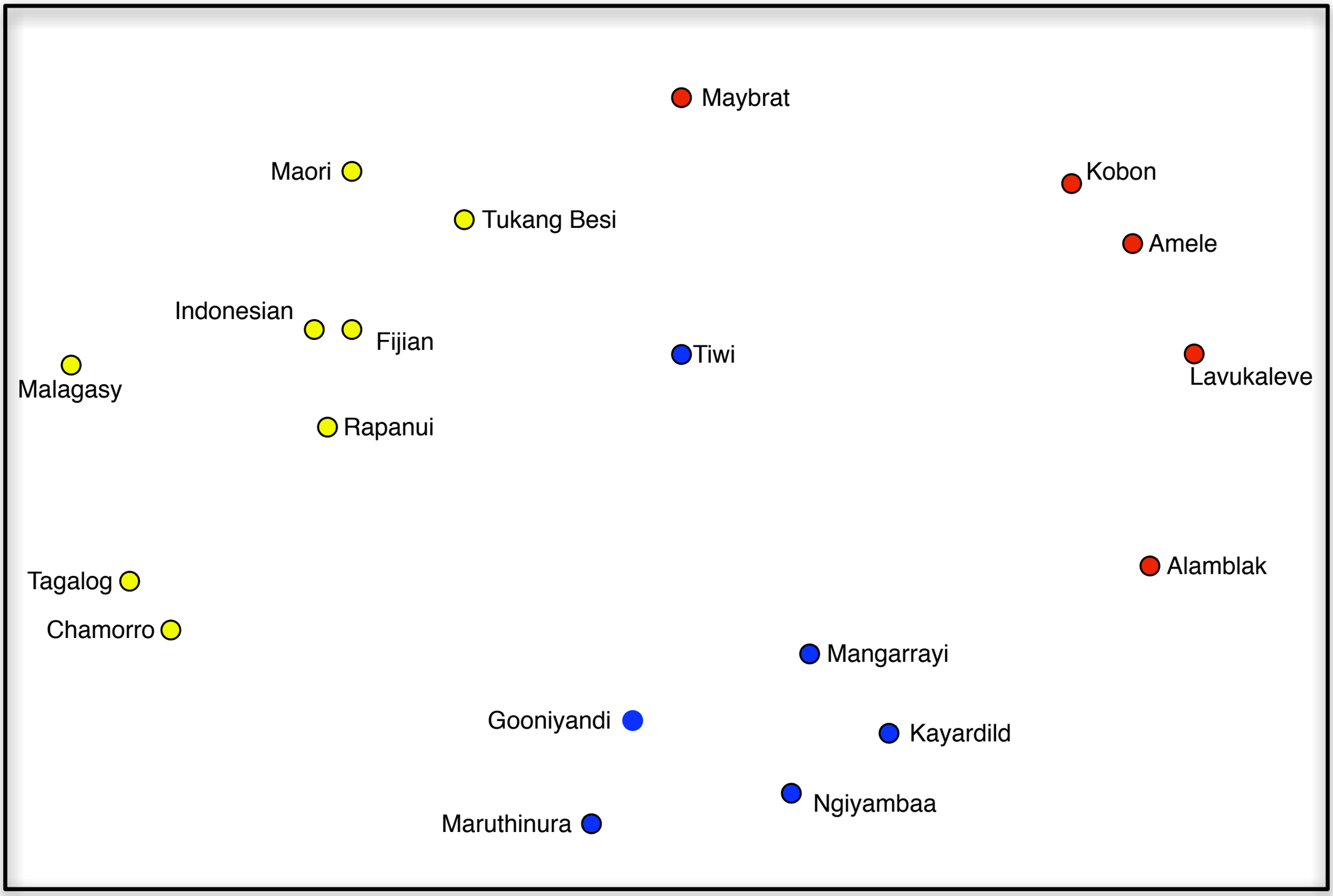
A closer look at geography: the case of **Oceania**



MDS of typological distances



MDS of typological distances

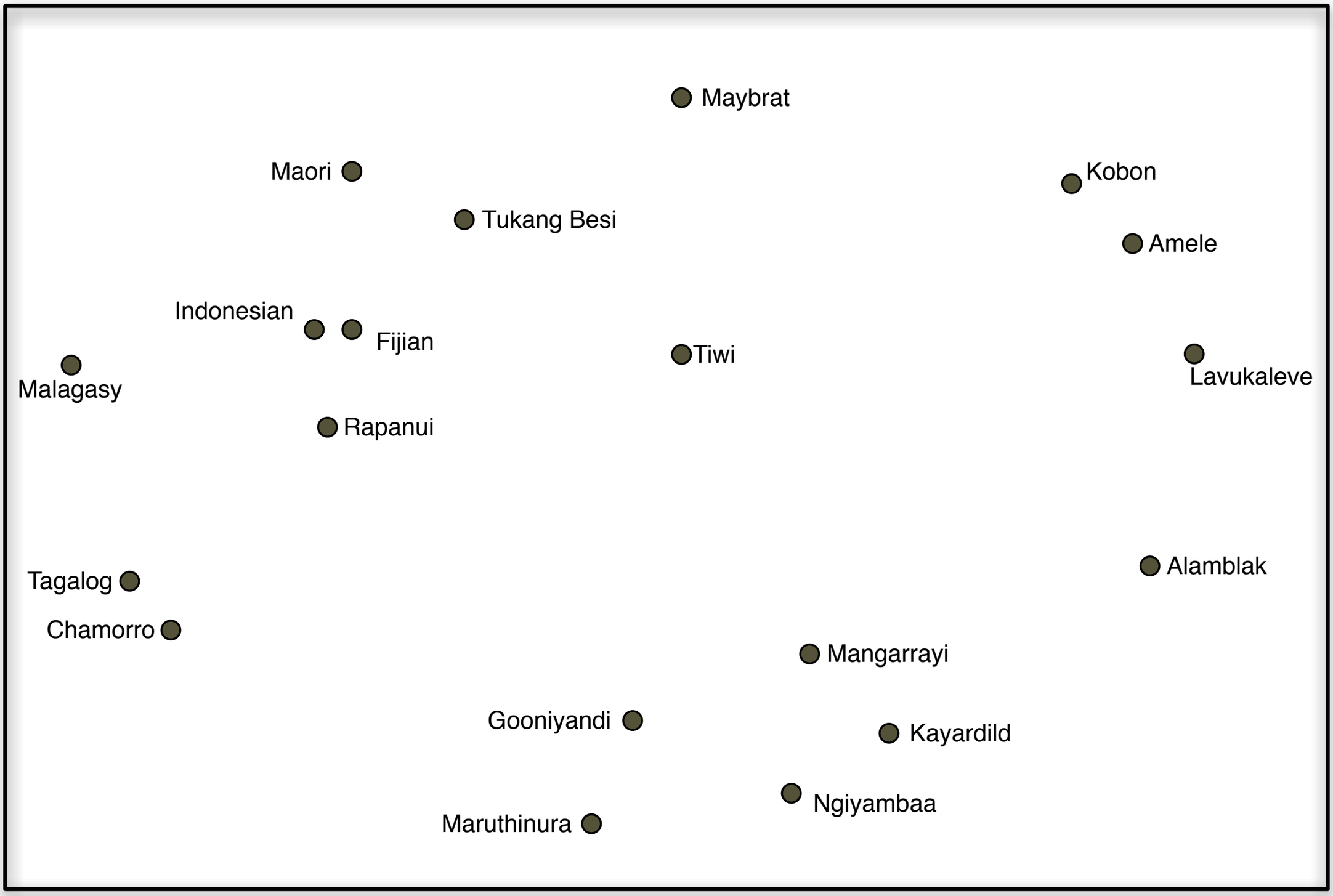


Maybrat
Kobon
Tukang Besi
Amele
Tiwi
Lavukaleve
Alamblak
Mangarrayi
Gooniyandi
Kayardild
Ngiyambaa
Maruthinura

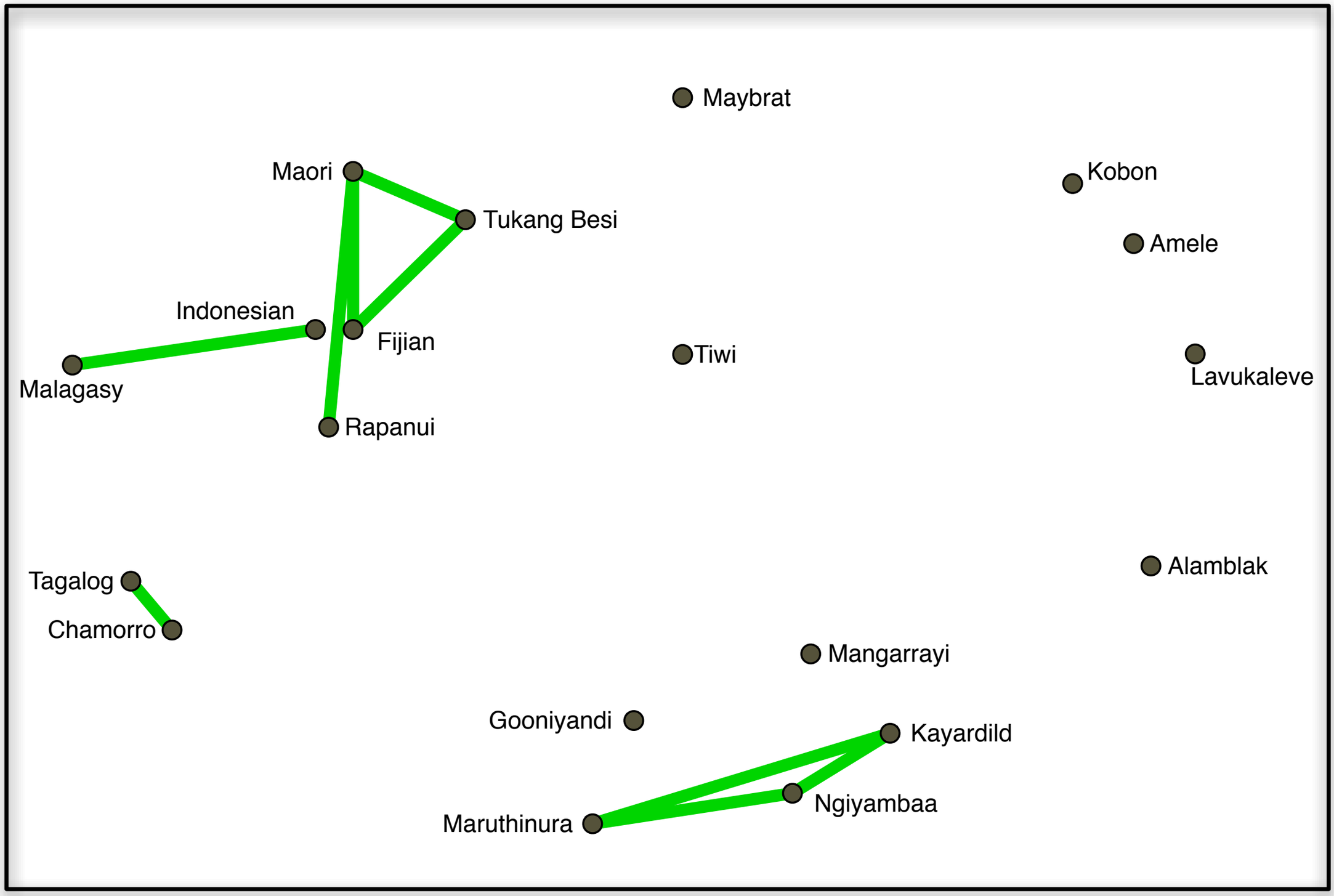
Exploring the language-geography relations

- remove the biggest distances (distortion)
- take the extremes of typology/geography
- very low values: linguistically (too) similar
- very high values: linguistically (too) diverse

MDS of typological distances

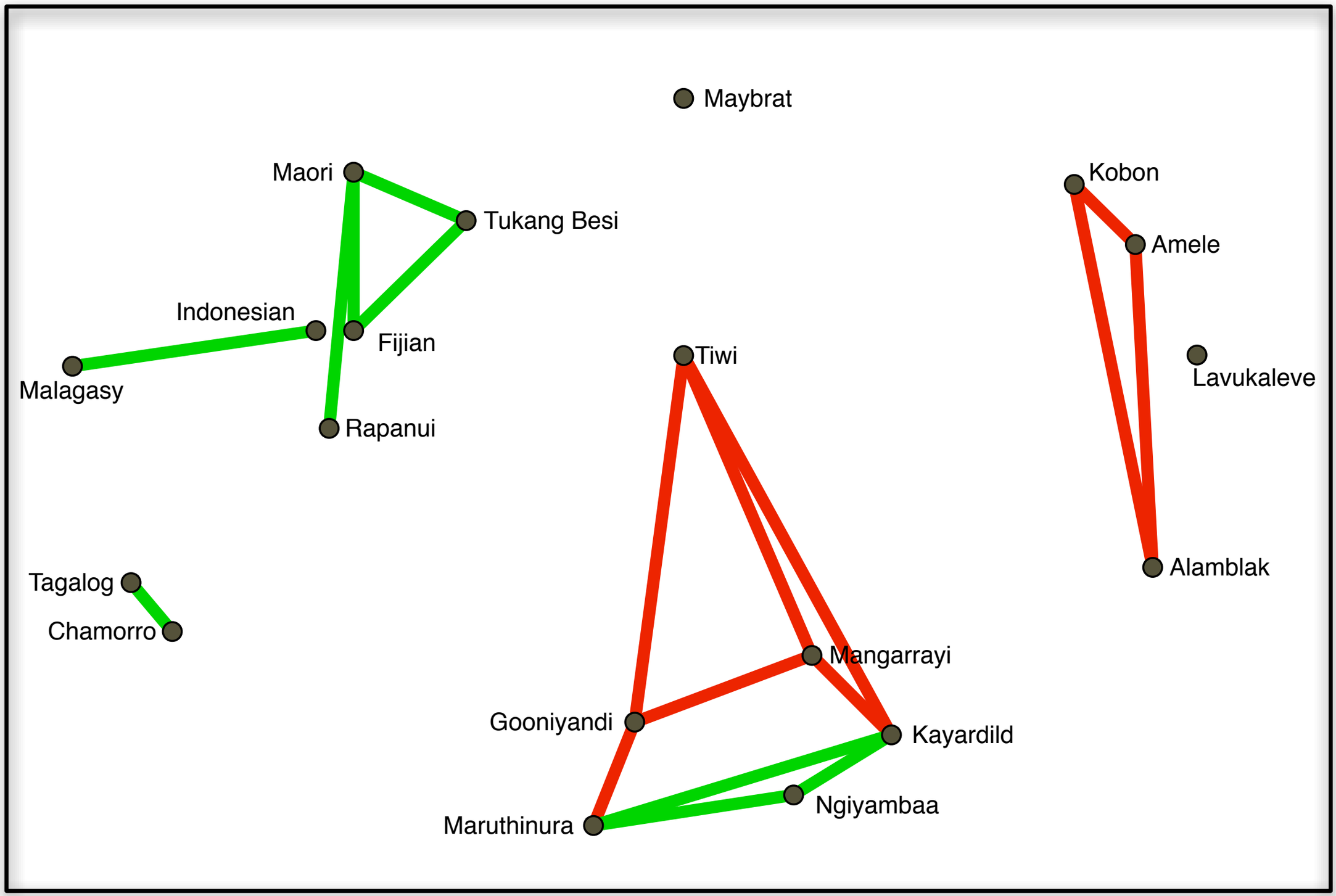


MDS of typological distances



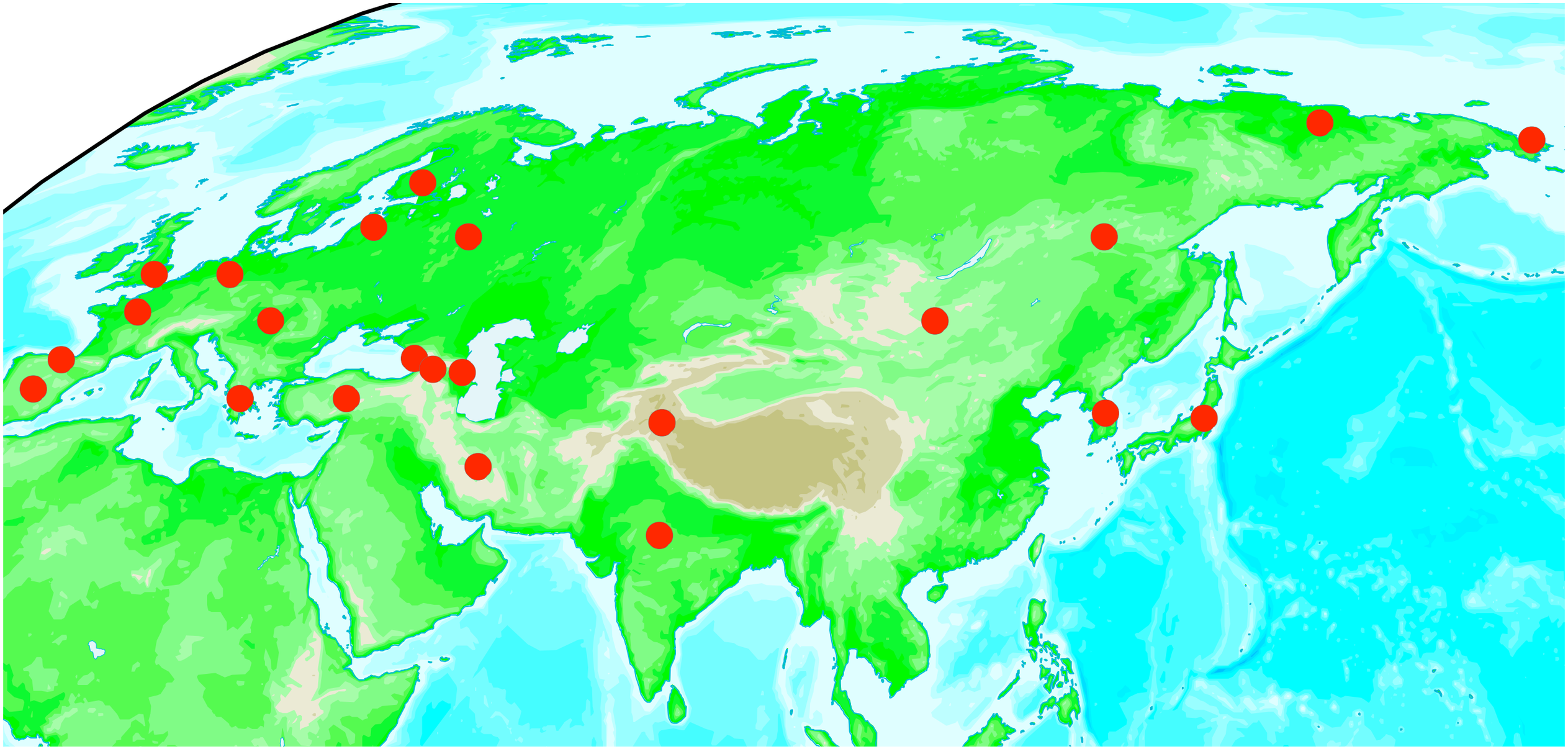
— Pairs of linguistically (too) similar languages

MDS of typological distances

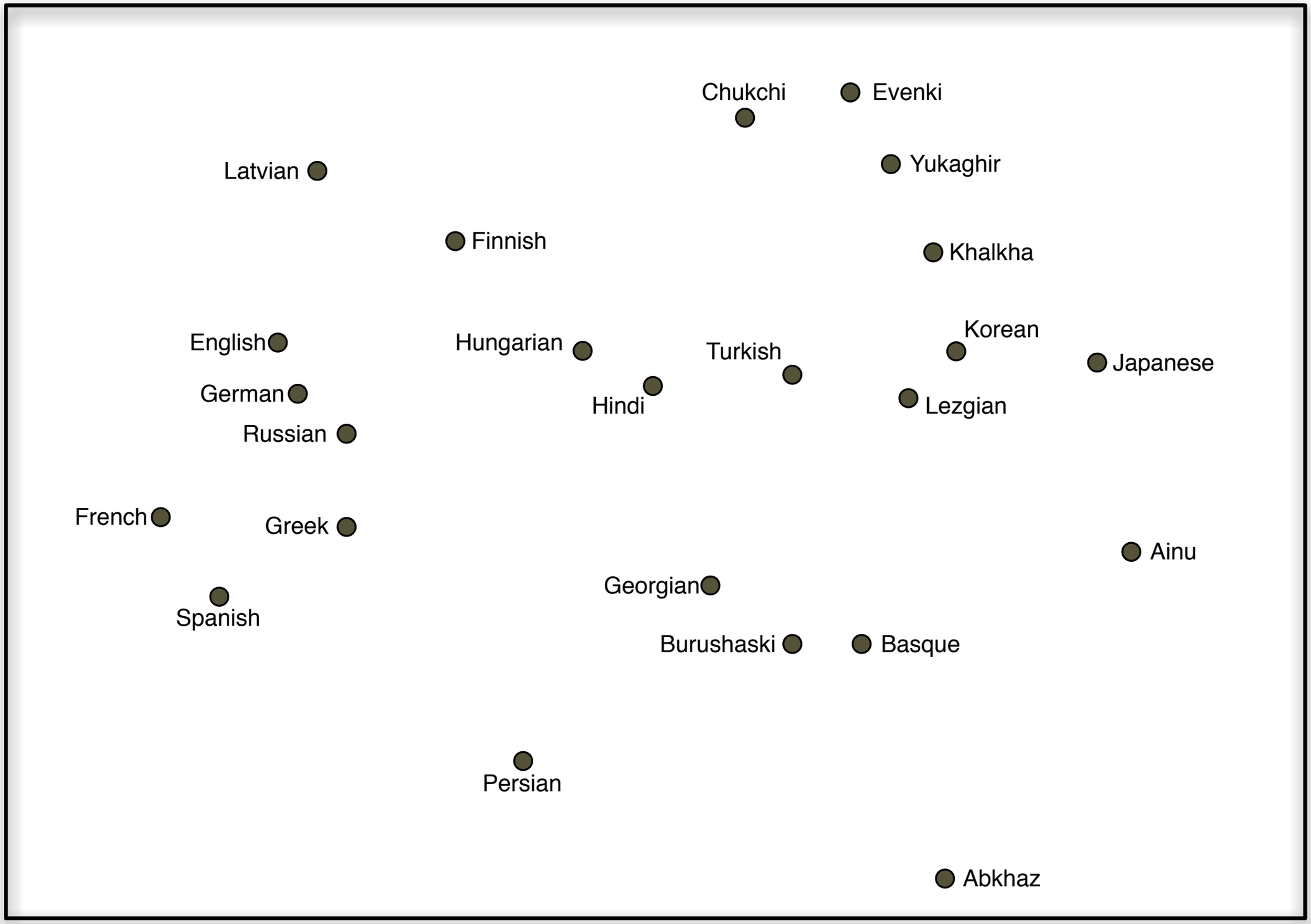


– Do not interpret them as a group

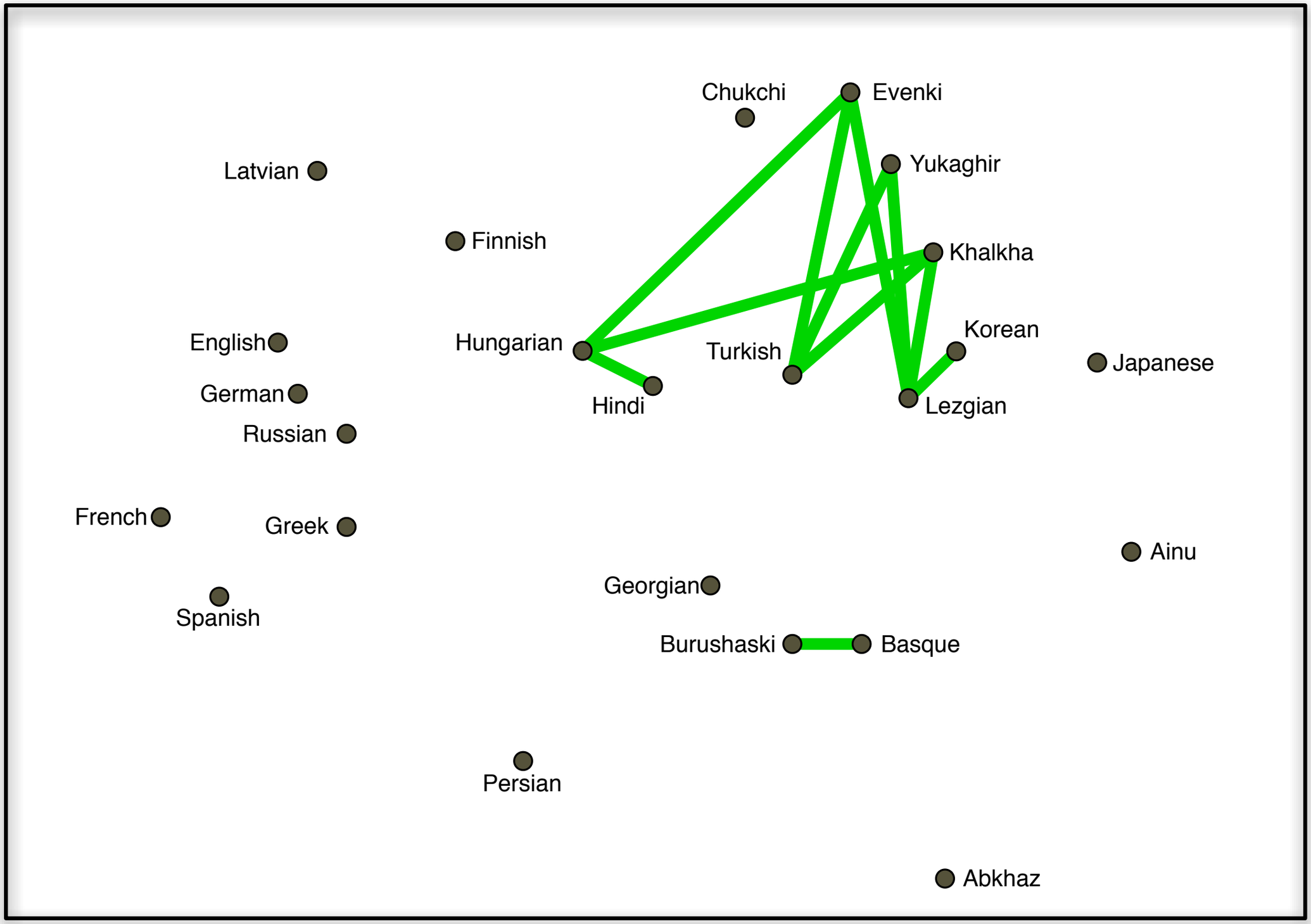
A closer look at geography: the case of **Eurasia**



MDS of typological distances

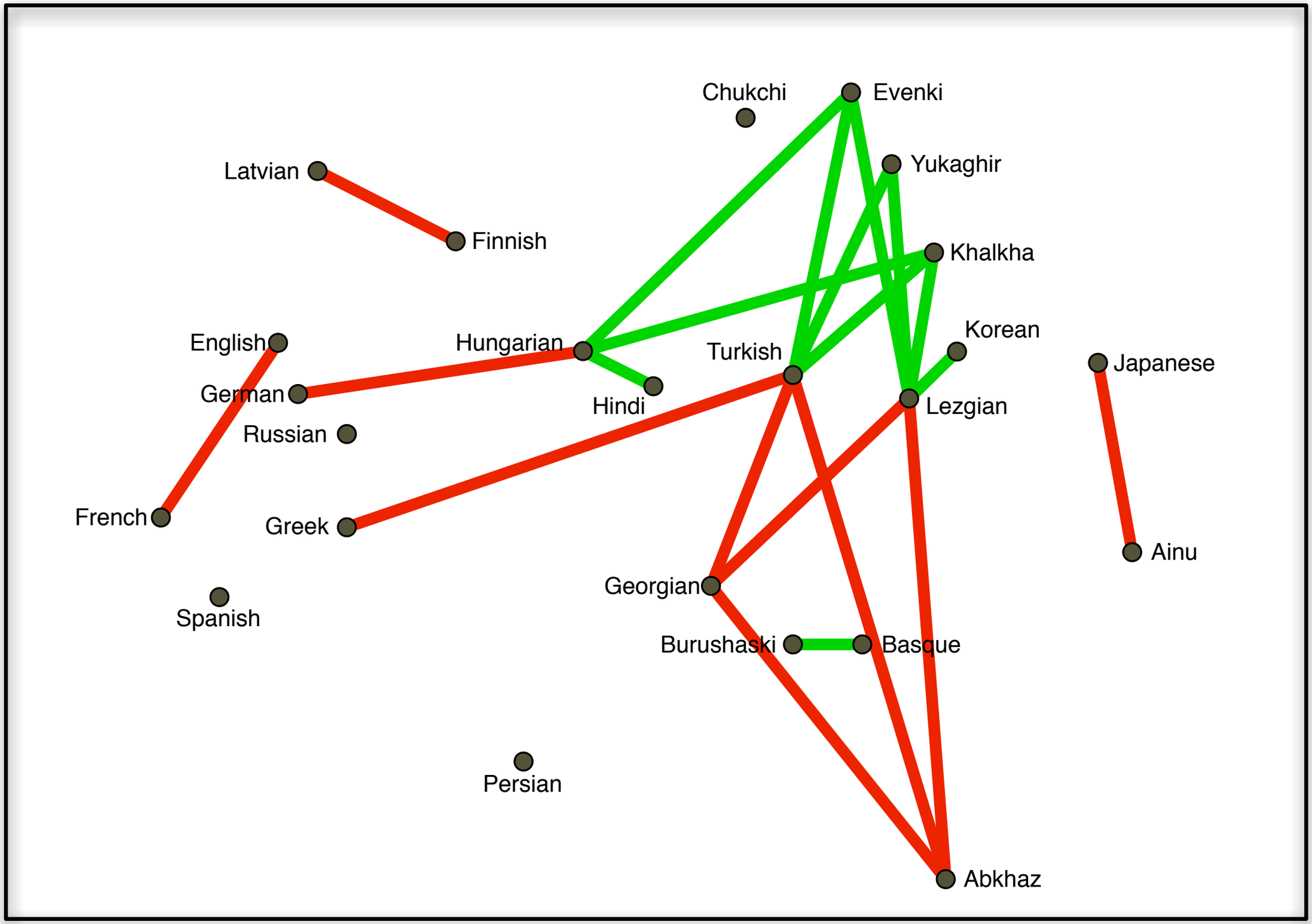


MDS of typological distances



— Pairs of linguistically (too) similar languages

MDS of typological distances



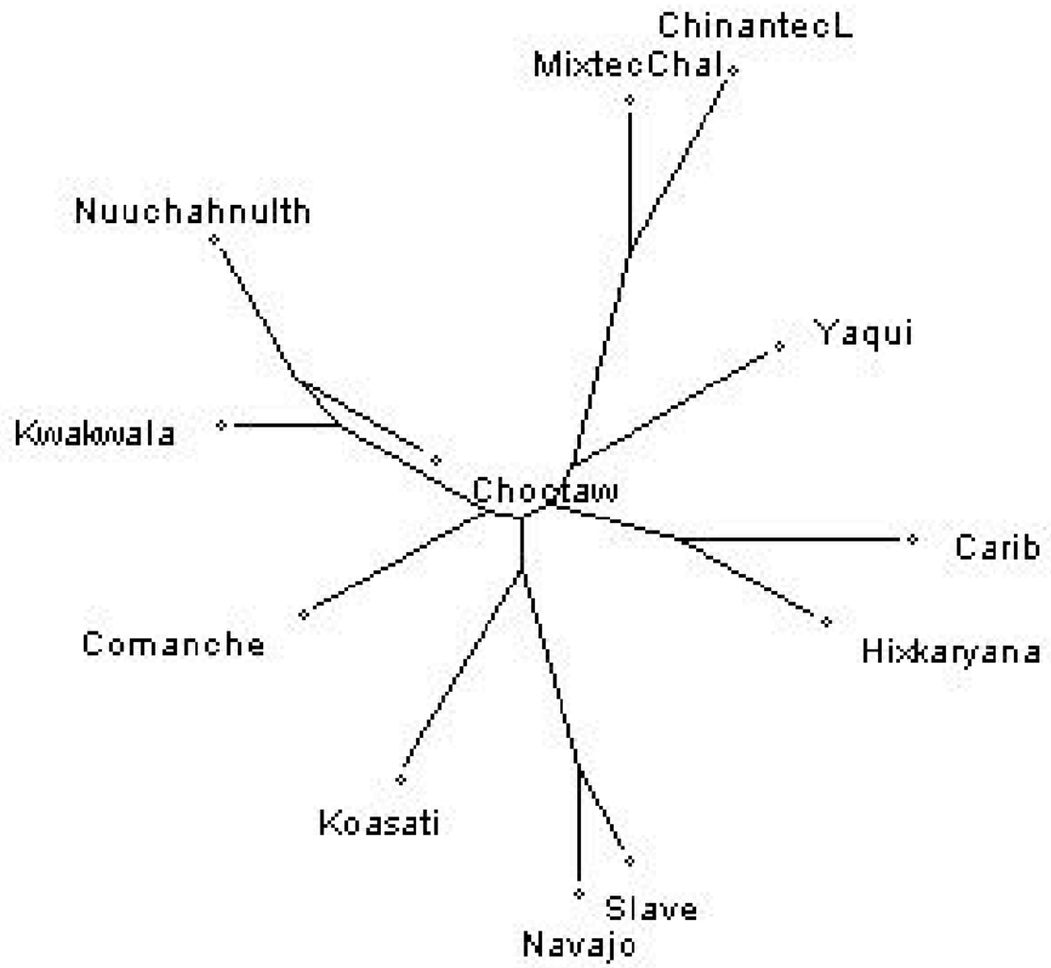
– Do not interpret them as a group

Selection of suitable characteristics

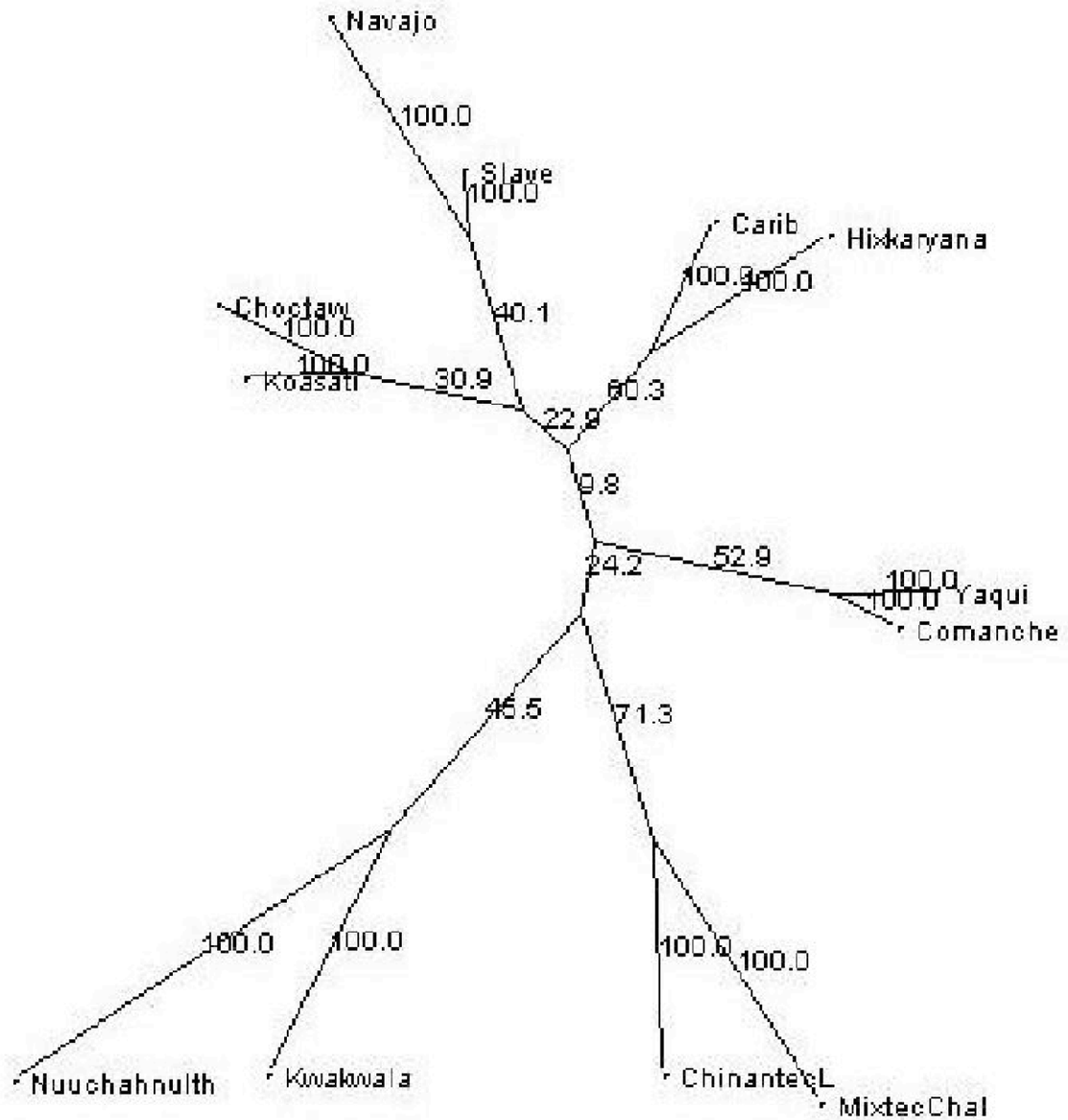
Usable for phylogeny

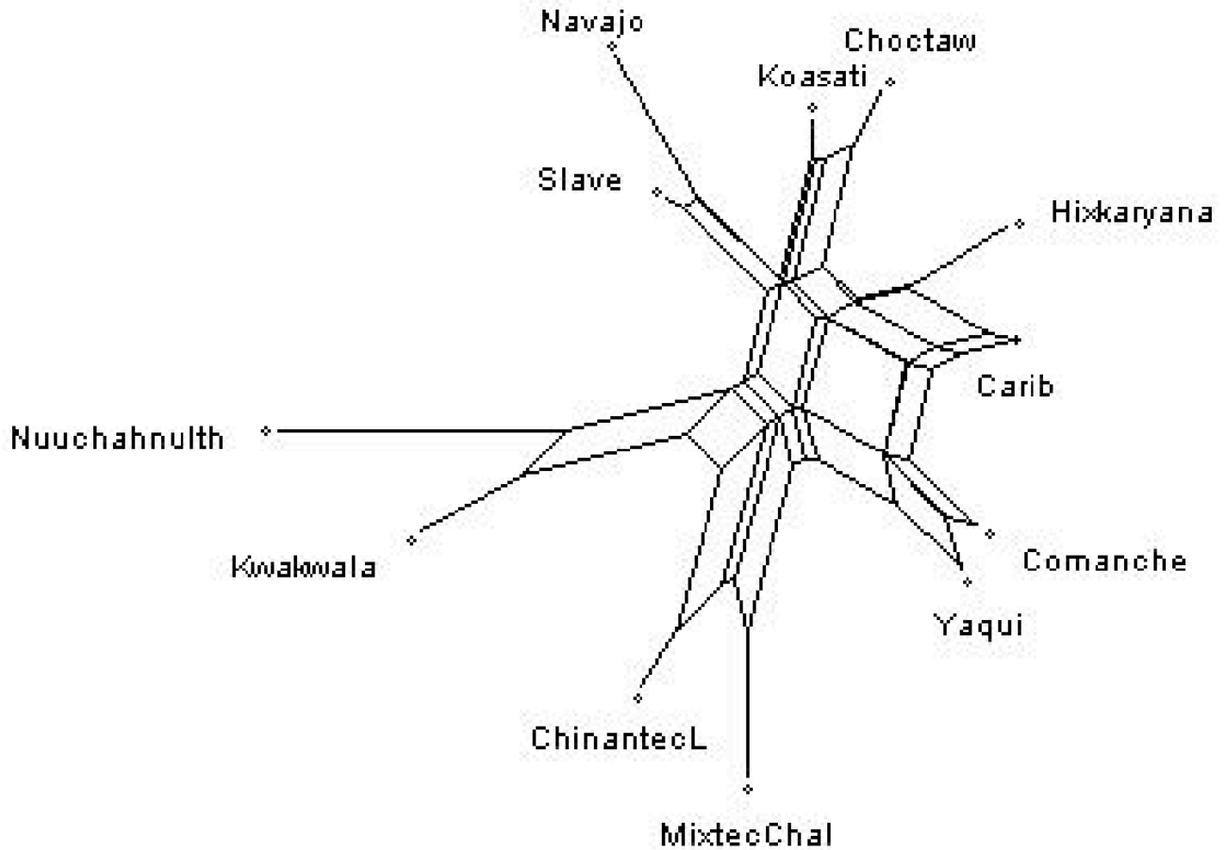
- Various approaches comparing each feature to the complete WALS dataset (A. Dress)
- Consistency/Retentions Indices (R. Gray)
- Consistency-measure within lower-level linguistic subgrouping (S. Wichmann)
- Energy-based consistency on a known partial tree (M. Albu)

Using the WALS data

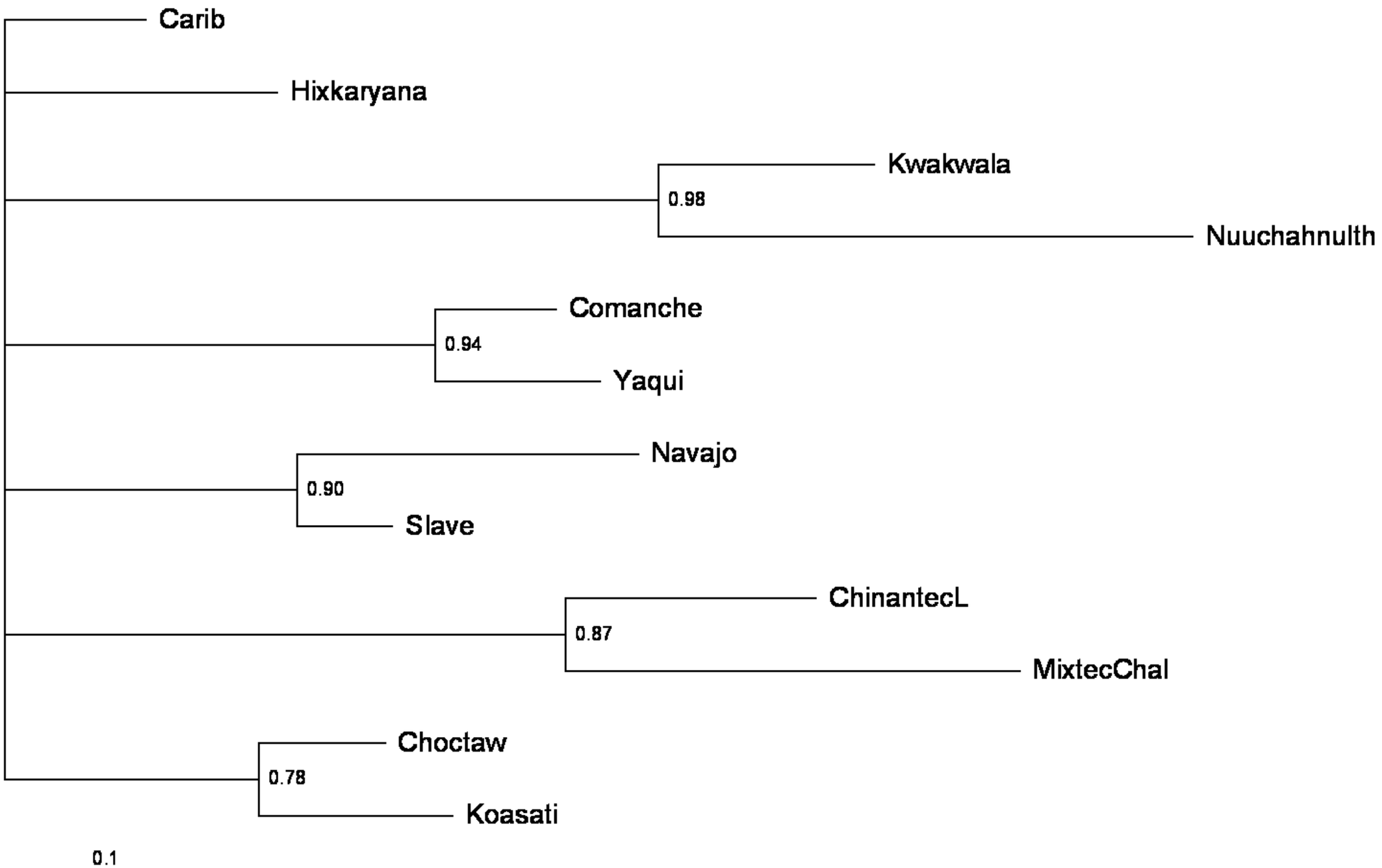


Using only the ‘best’ data is better





Bayesian-approach works best



WALS forever !

Party Downstairs