Using WALS

Prospects of quantitative approaches for linguistic typology

Michael Cysouw

Max Planck Institute
for Evolutionary Anthropology

WALS – World Atlas of Language Structures

WALS

The World Atlas of Language Structures

edited by M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie

The Interactive Reference Tool

developed by Hans-Jörg Bibiko

OXFORD UNIVERSITY PRESS

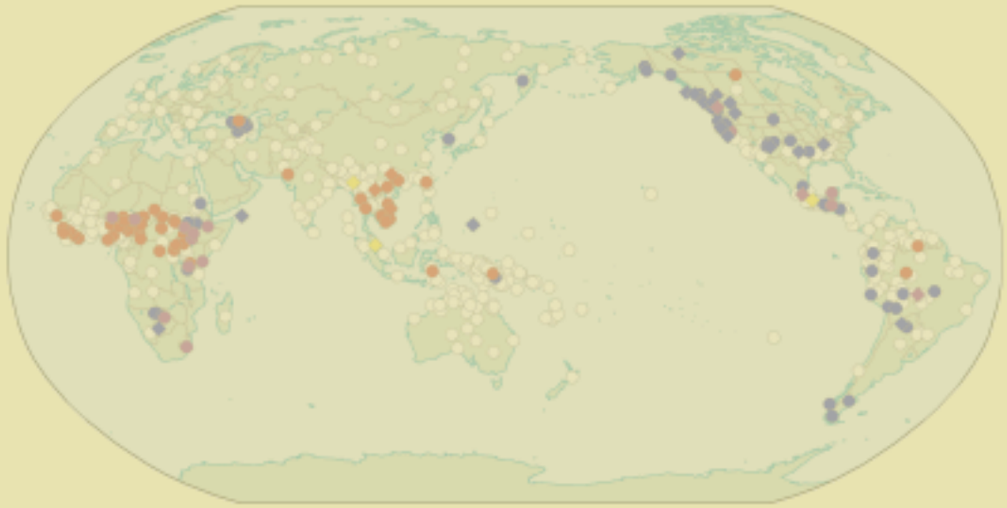
FEATURE VIEWER

LANGUAGE VIEWER

COMPOSER

ABOUT THIS PROGRAM

GUIDED TOUR



Wow !

WALS is just great

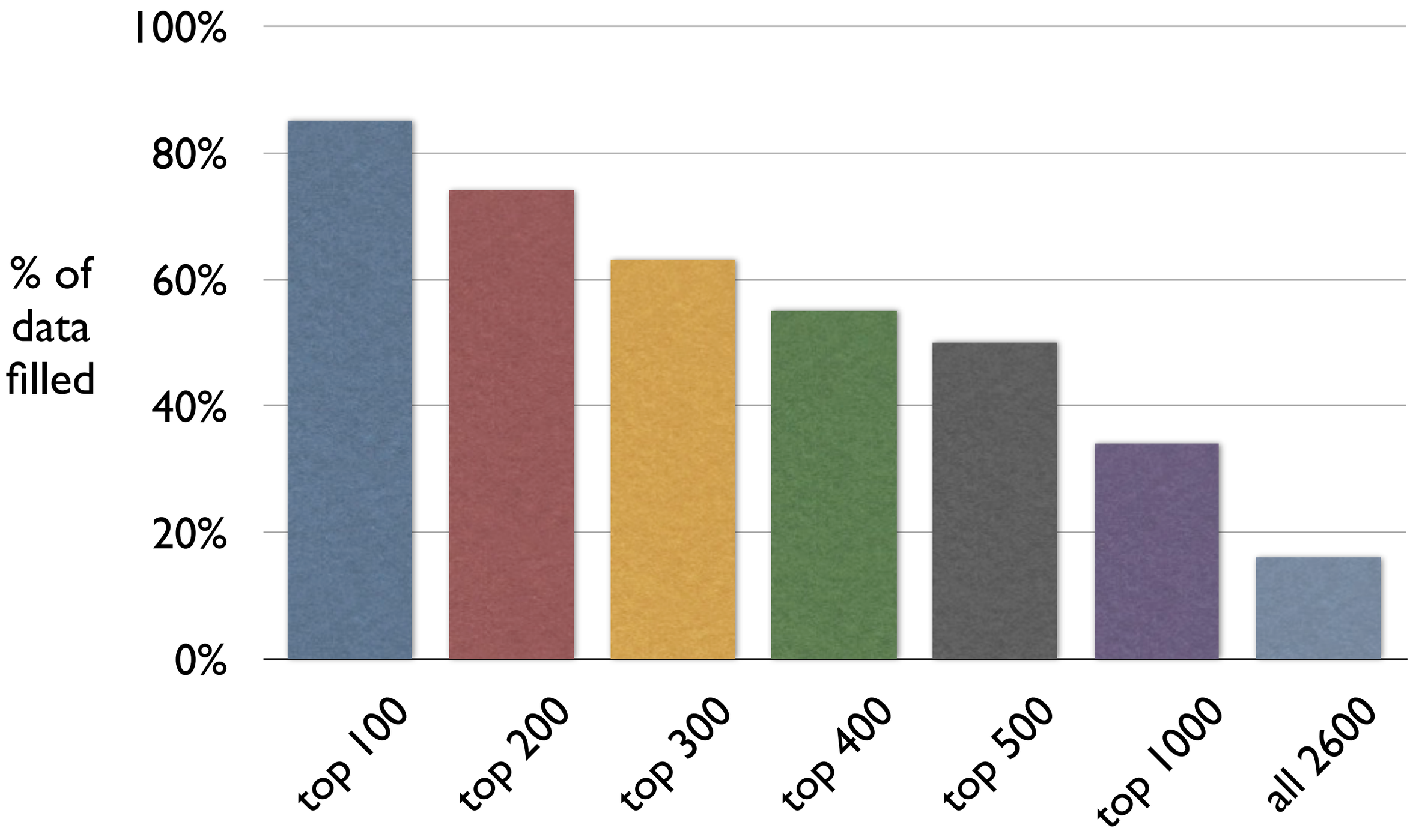
- Linguistic data is very expensive ...
  (data estimated 6,000,000 EUR)
- but now we have so much data
- from different linguistic domains
- and so well organized !

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 !

Choosing the best languages



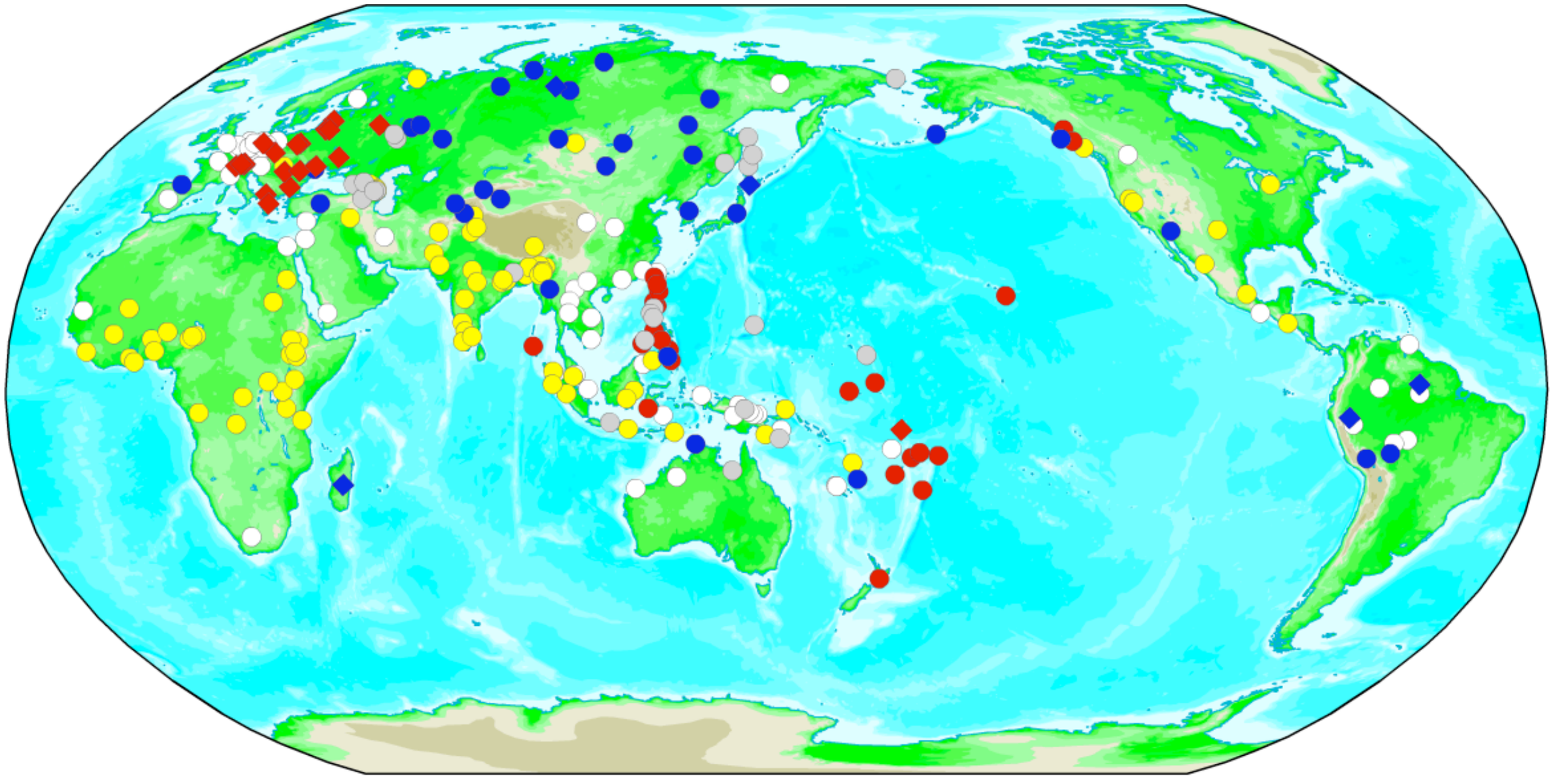
Reliability

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

Coding Problems

- Types consisting of dissimilar languages
- Independent features combined in one map
- Definitional dependencies between maps
- No relative similarities available

Map 54: Distributive numerals



1. No distributive numerals [62]
2. Marked by reduplication [84]
3. Marked by prefix [23]
4. Marked by suffix [32]
5. Marked by preceding word [21]
6. Marked by following word [5]
7. Marked by mixed or other strategies [23]

Solution:

recode all as different types

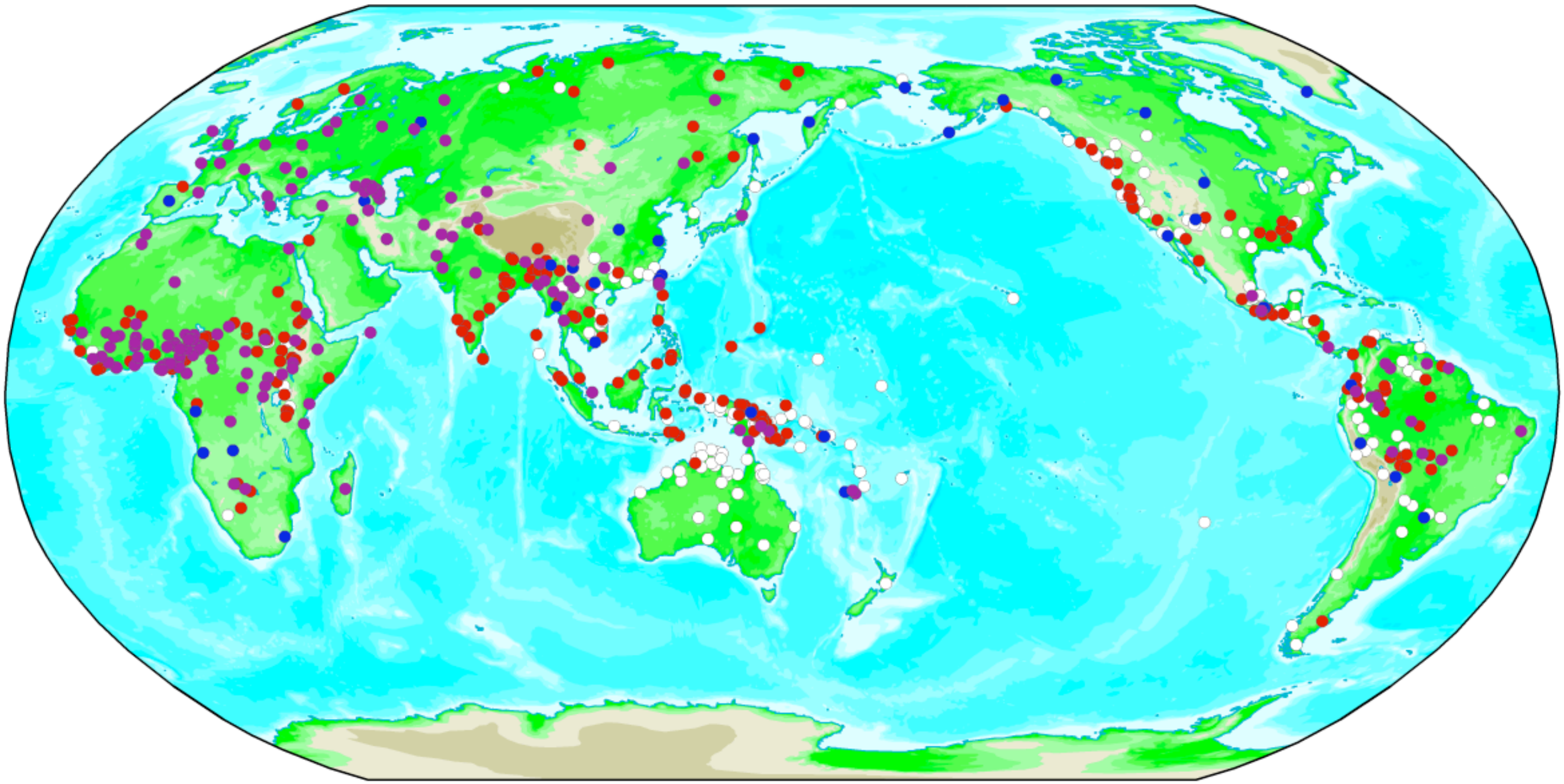
Coding Problems

- Types consisting of dissimilar languages
- Independent features combined in one map
- Definitional dependencies between maps
- No relative similarities available

Coding Problems

- Types consisting of dissimilar languages
- Independent features combined in one map
- Definitional dependencies between maps
- No relative similarities available

Map 4: Voicing in Plosives and Fricatives



1. No voicing constrast [181]
2. In plosives alone [189]
3. In fricatives alone [38]
4. In both plosives and fricatives [158]

Solution:
split, and disregard the original

Coding Problems

- Types consisting of dissimilar languages
- Independent features combined in one map
- Definitional dependencies between maps
- No relative similarities available

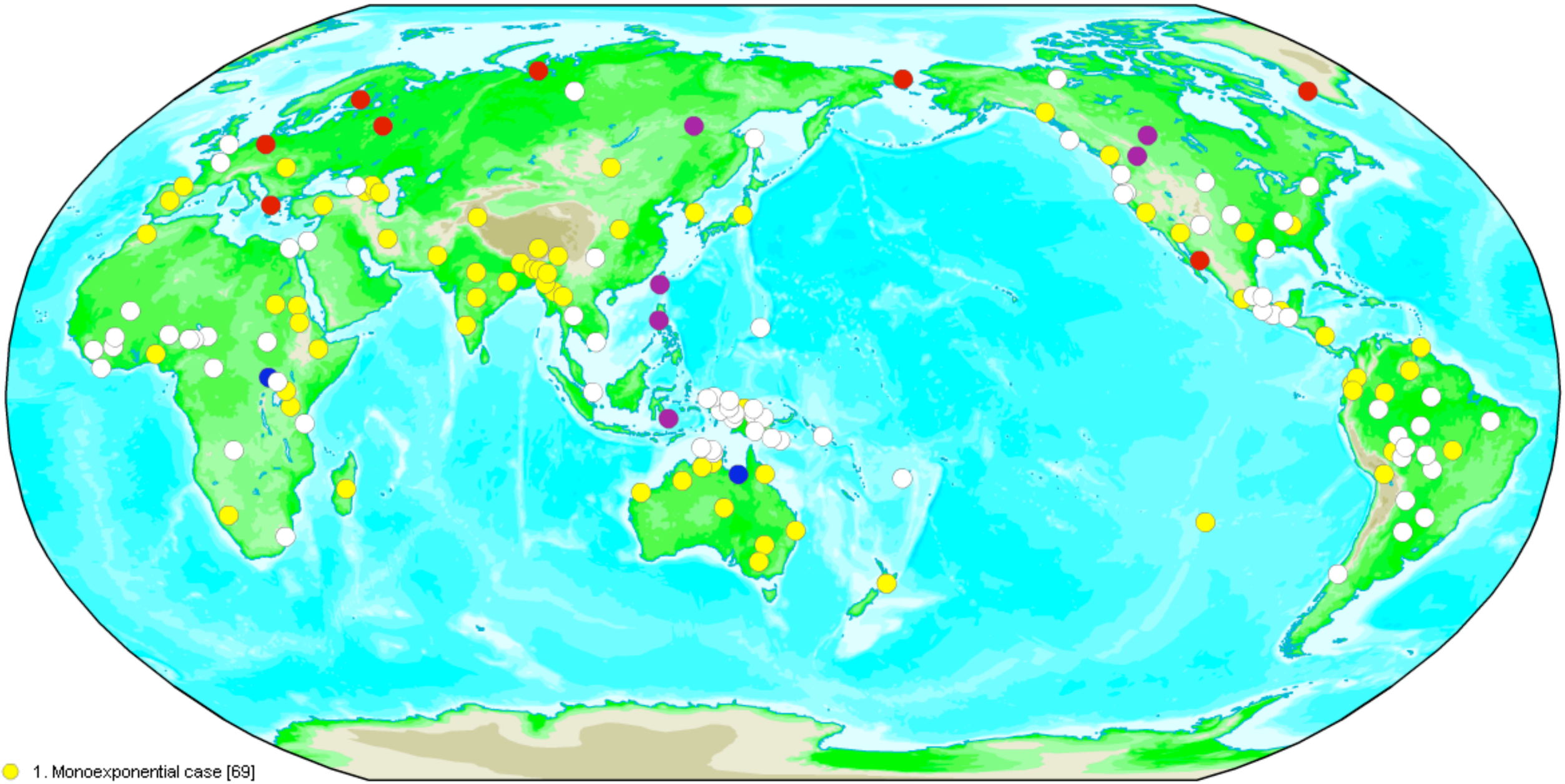
Coding Problems

- Types consisting of dissimilar languages
- Independent features combined in one map
- Definitional dependencies between maps
- No relative similarities available

Problem:

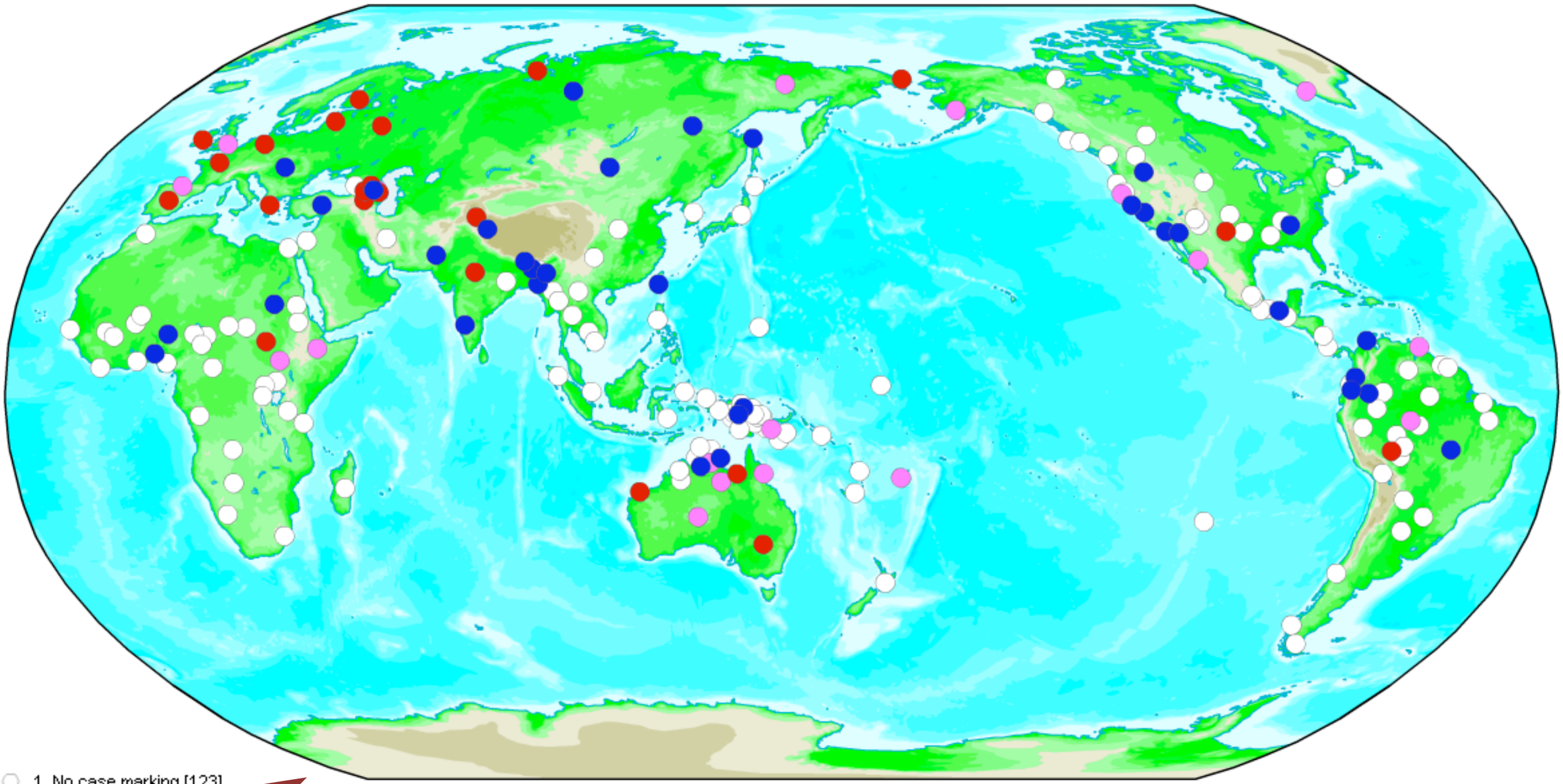
(Apparently) identical values

Map 21: Exponence of Selected Inflectional Formatives



1. Monoexponential case [69]
2. Case + number [8]
3. Case + referentiality [6]
4. Case + TAM [2]
5. No case [75]

Map 28: Case Syncretism

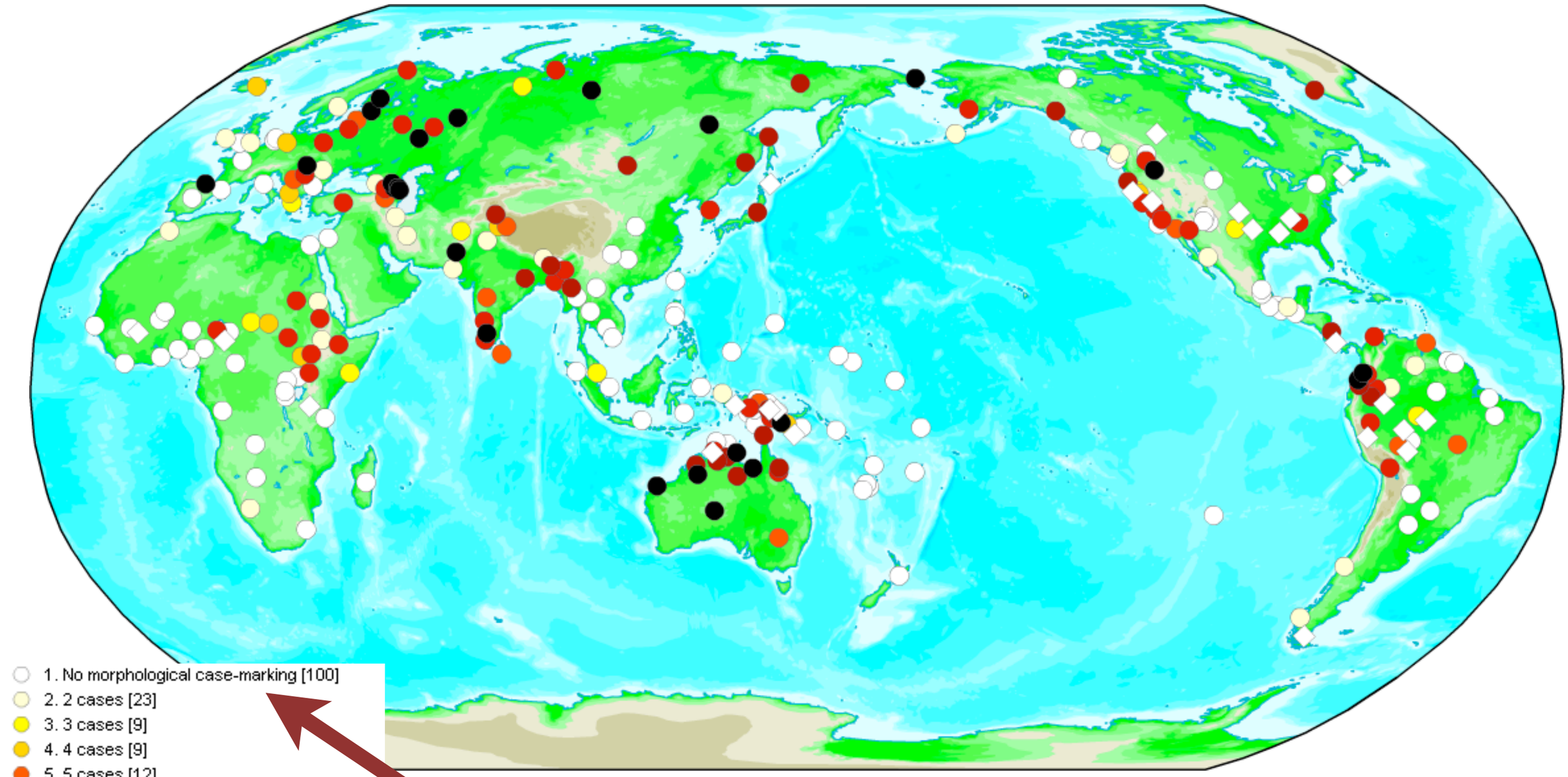


Different definitions and interpretation

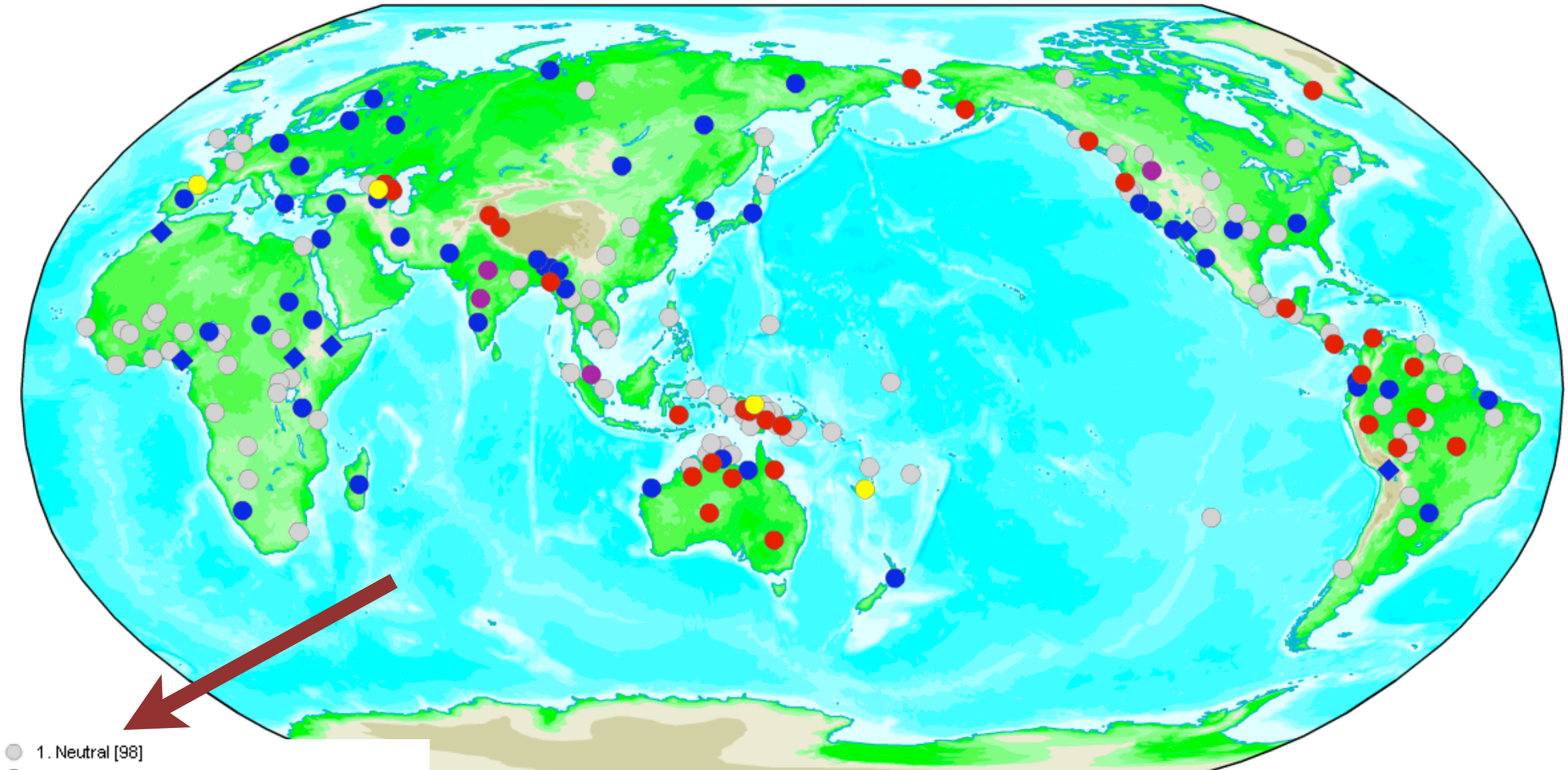
	no case (map 28)	case (map 28)
no case (map 21)	48	8
case (map 21)	28	44

Problem:
Covert Dependencies

Map 49: Number of Cases



Map 98: Alignment of case marking of full noun phrases



1. Neutral [98]
2. Nominative - accusative (standard) [46]
3. Nominative - accusative (marked nominative) [6]
4. Ergative - absolutive [32]
5. Tripartite [4]
6. Active-inactive [4]

Marking of full noun phrases

	neutral alignment	non-neutral alignment
no case distinctions	62	12
case distinctions	17	77

Solution:
Combine features into one feature with (very) many values

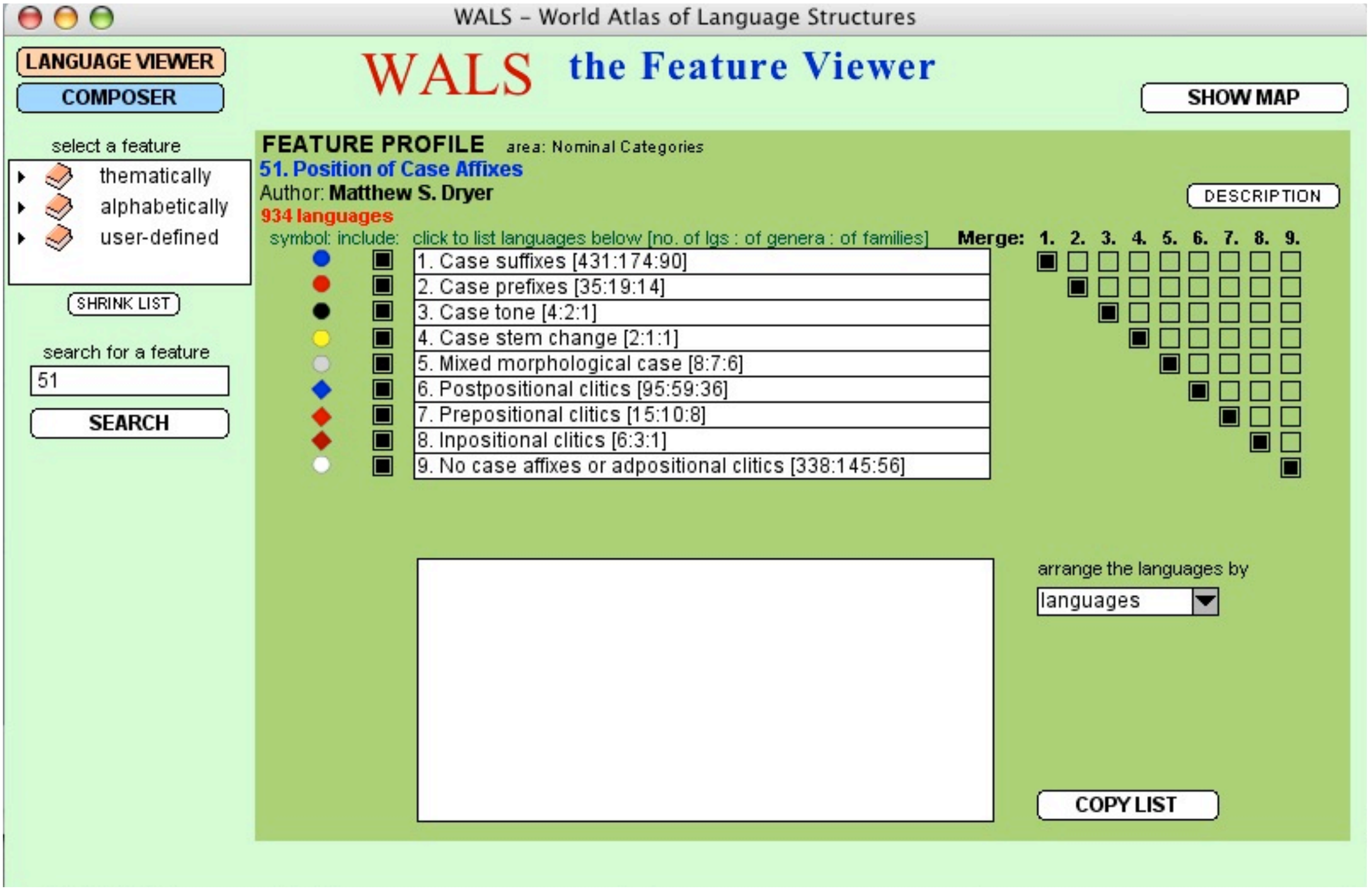
Coding Problems

- Types consisting of dissimilar languages
- Independent features combined in one map
- Definitional dependencies between maps
- No relative similarities available

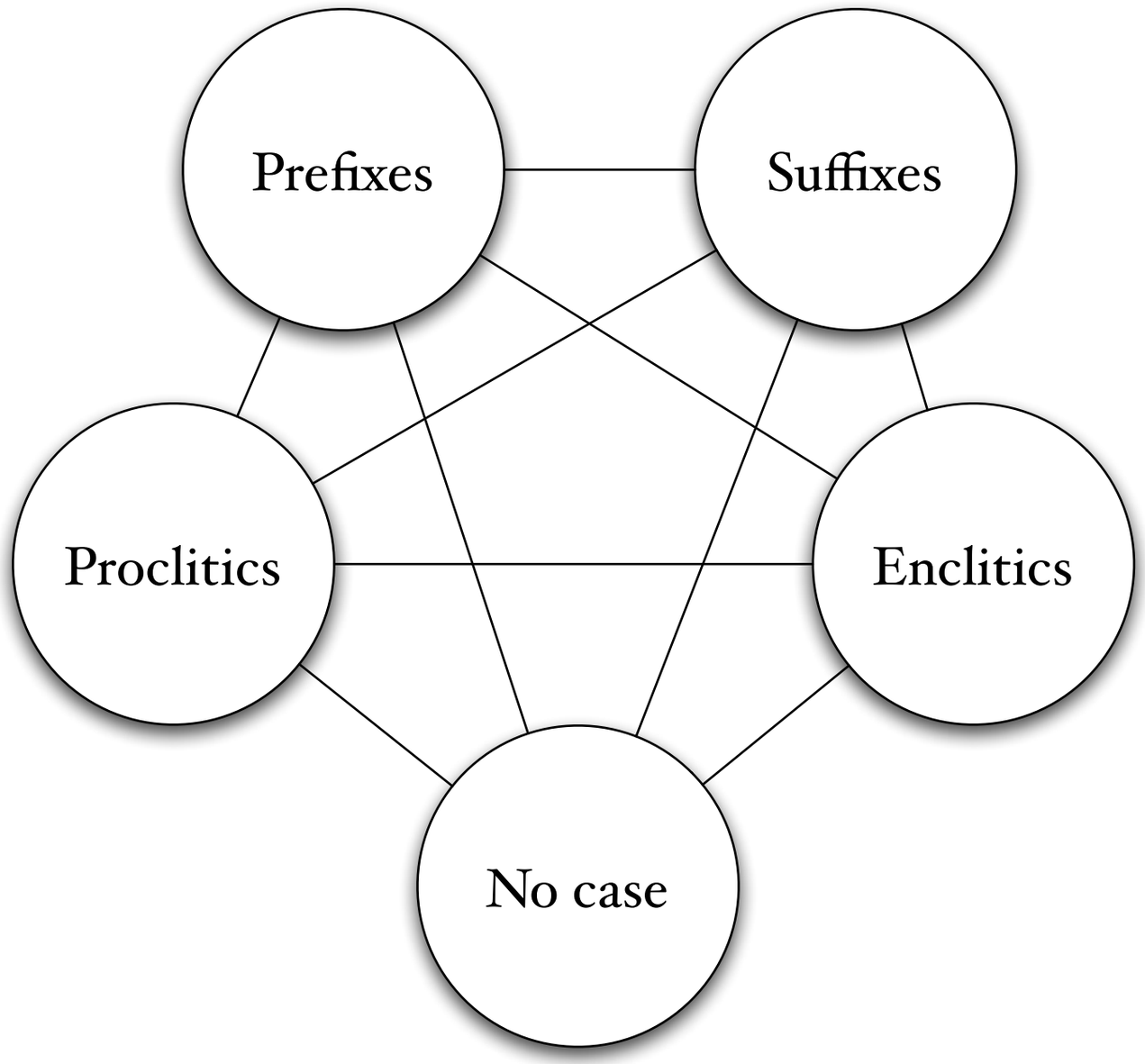
Coding Problems

- Types consisting of dissimilar languages
- Independent features combined in one map
- Definitional dependencies between maps
- No relative similarities available

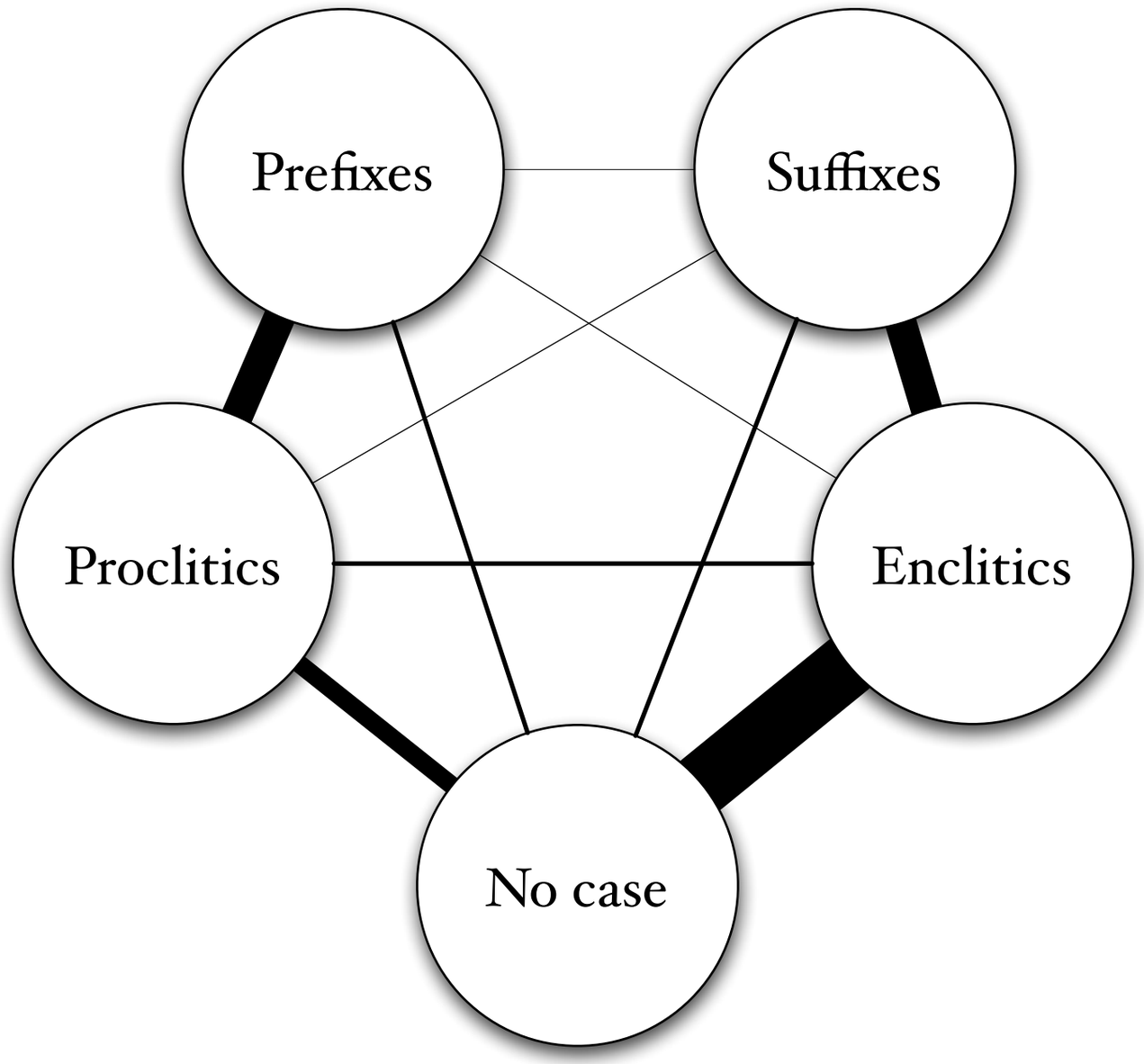
Map 51: Position of Case affixes



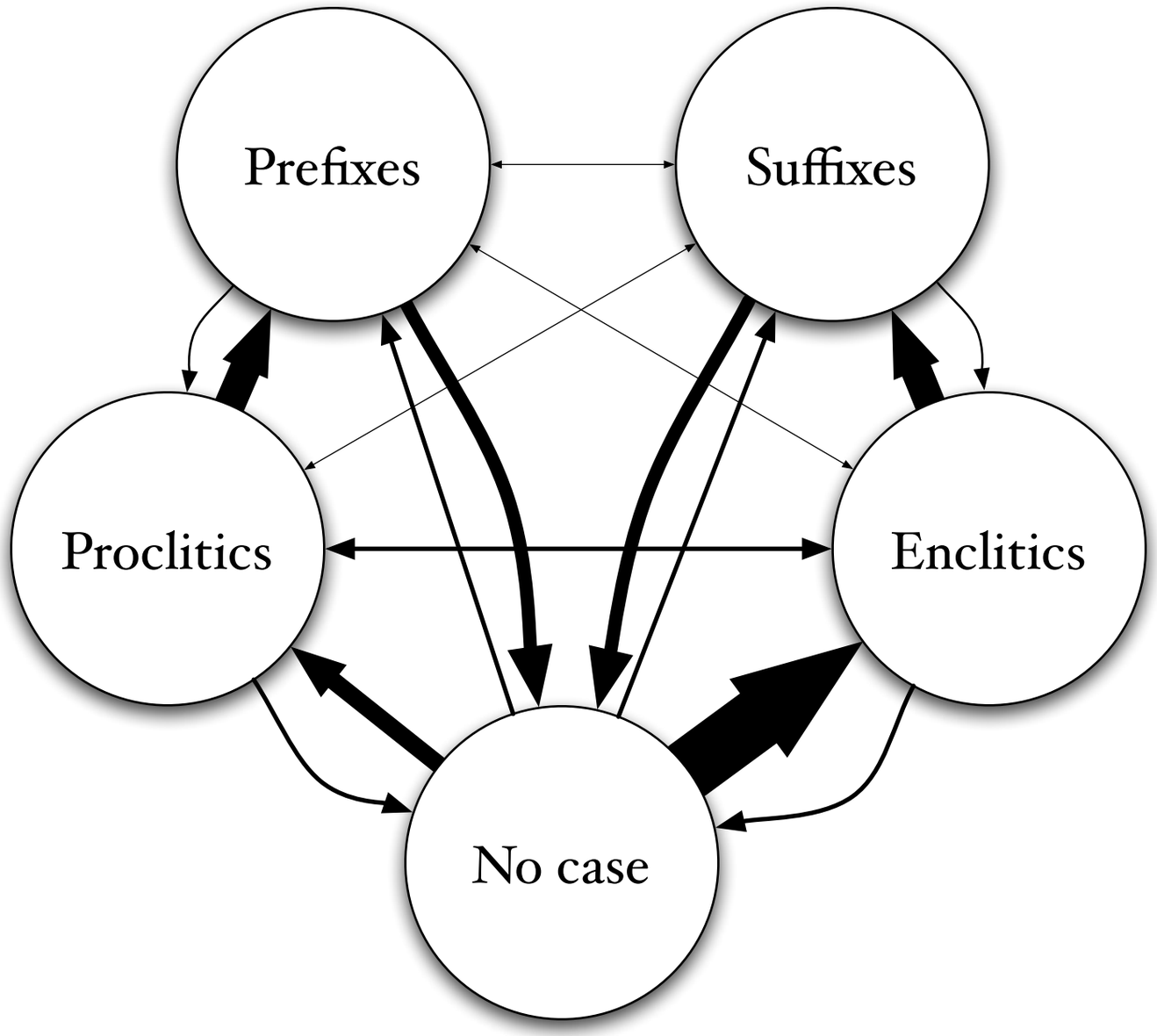
Undifferentiated typology



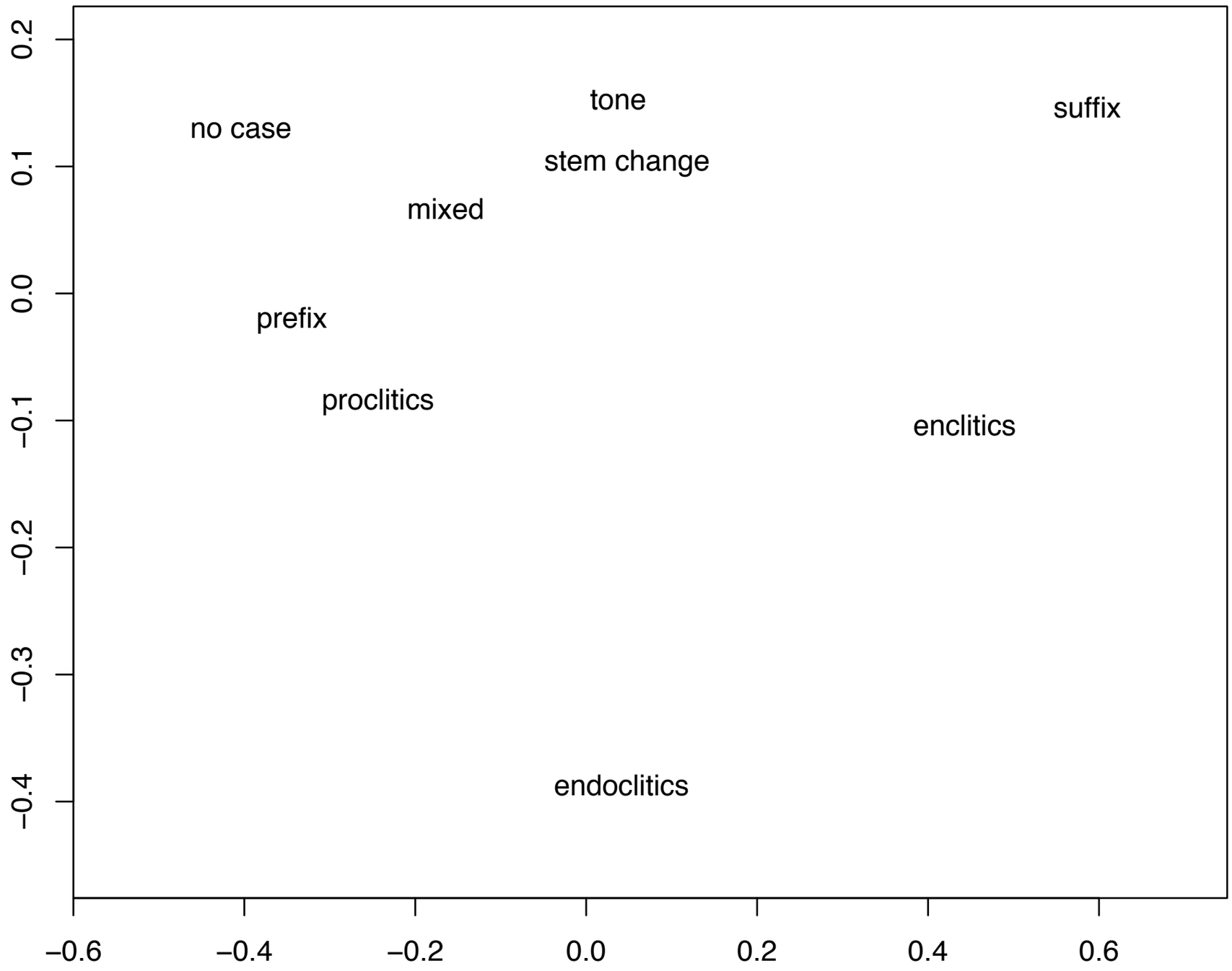
Similarities



Transitional probabilities



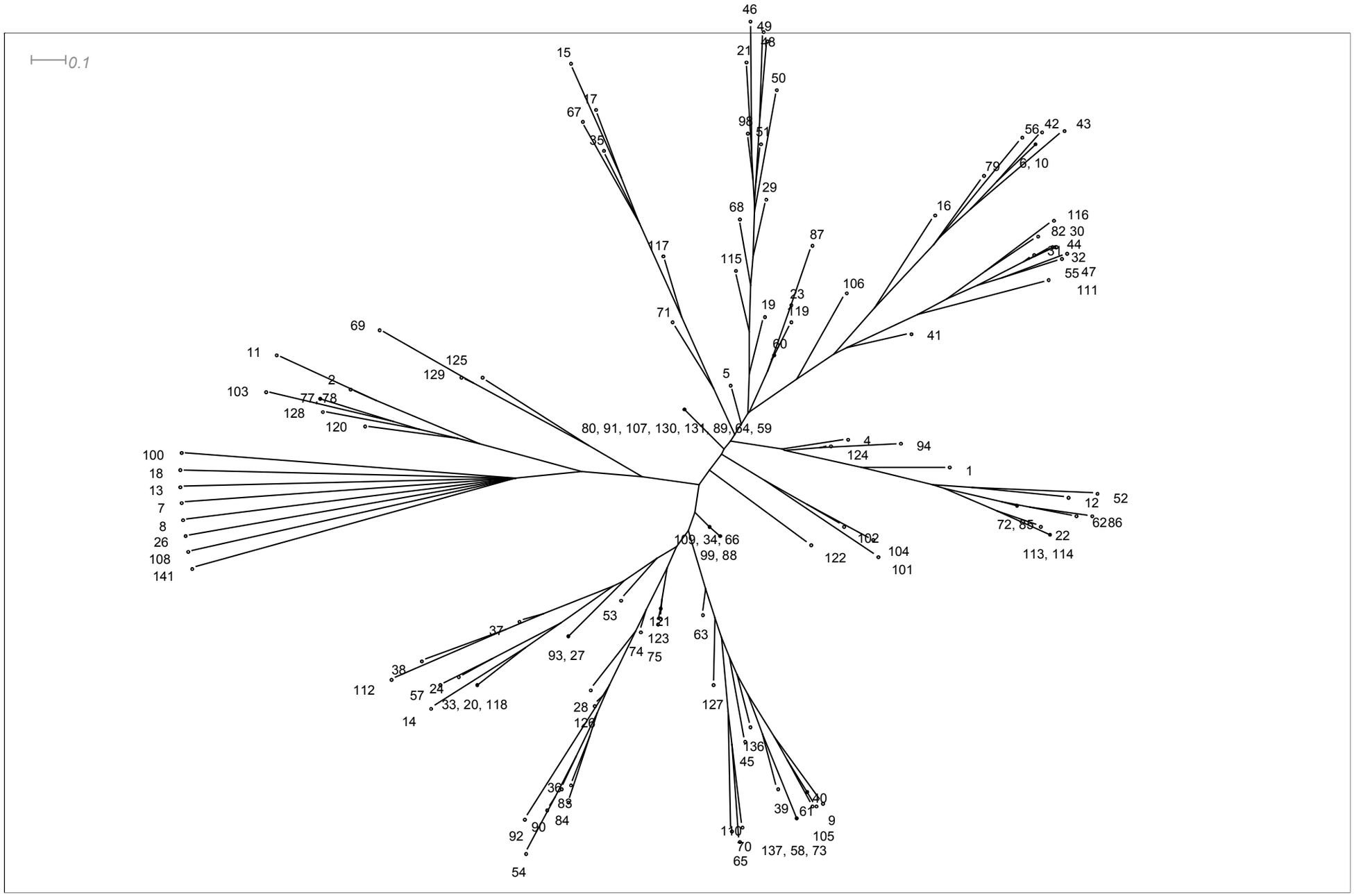
Estimate **similarities** from low-level genetic groups



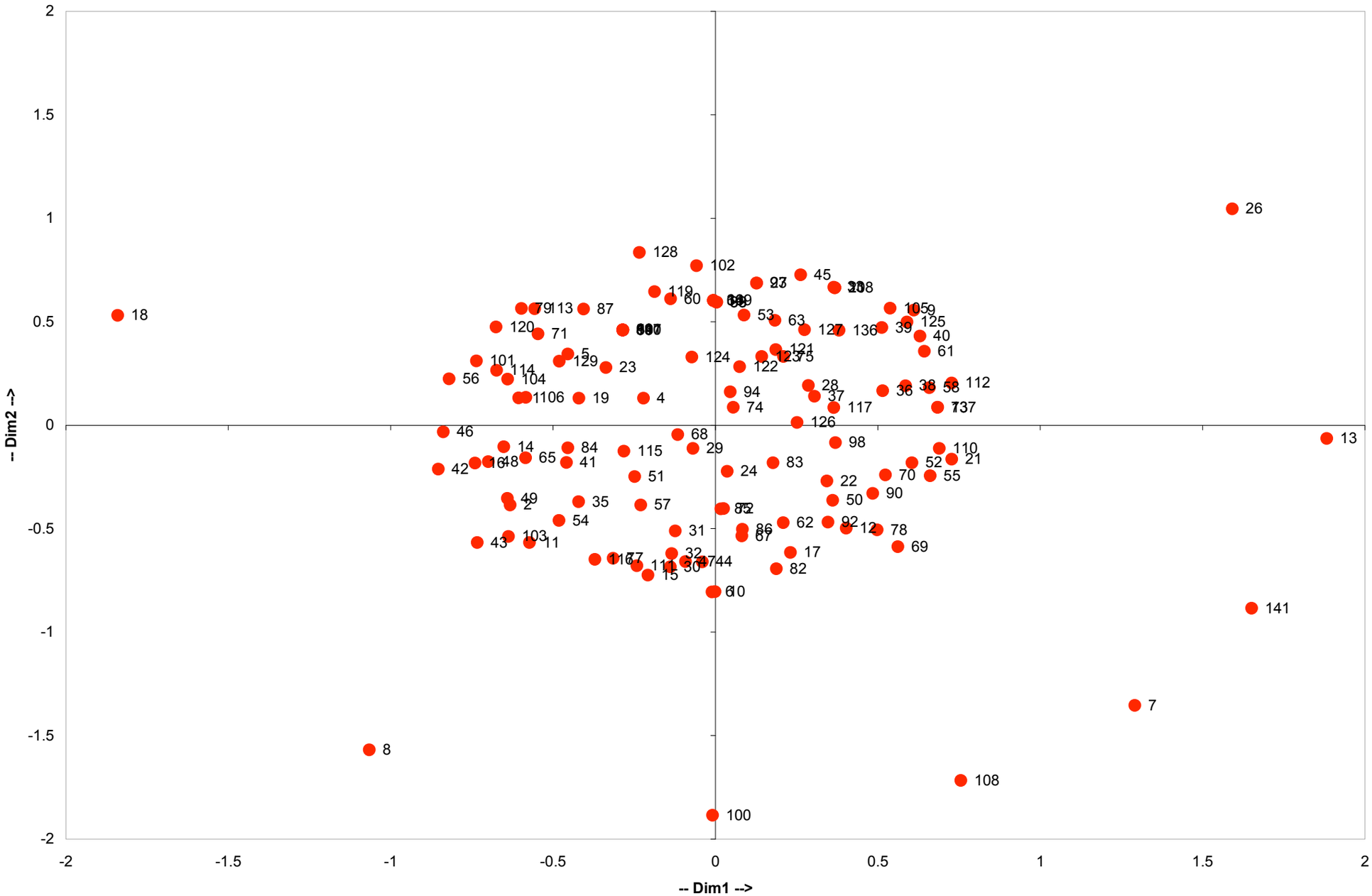
Exploration

Clustering of **features**

Looks promising...



… but it is an artifact

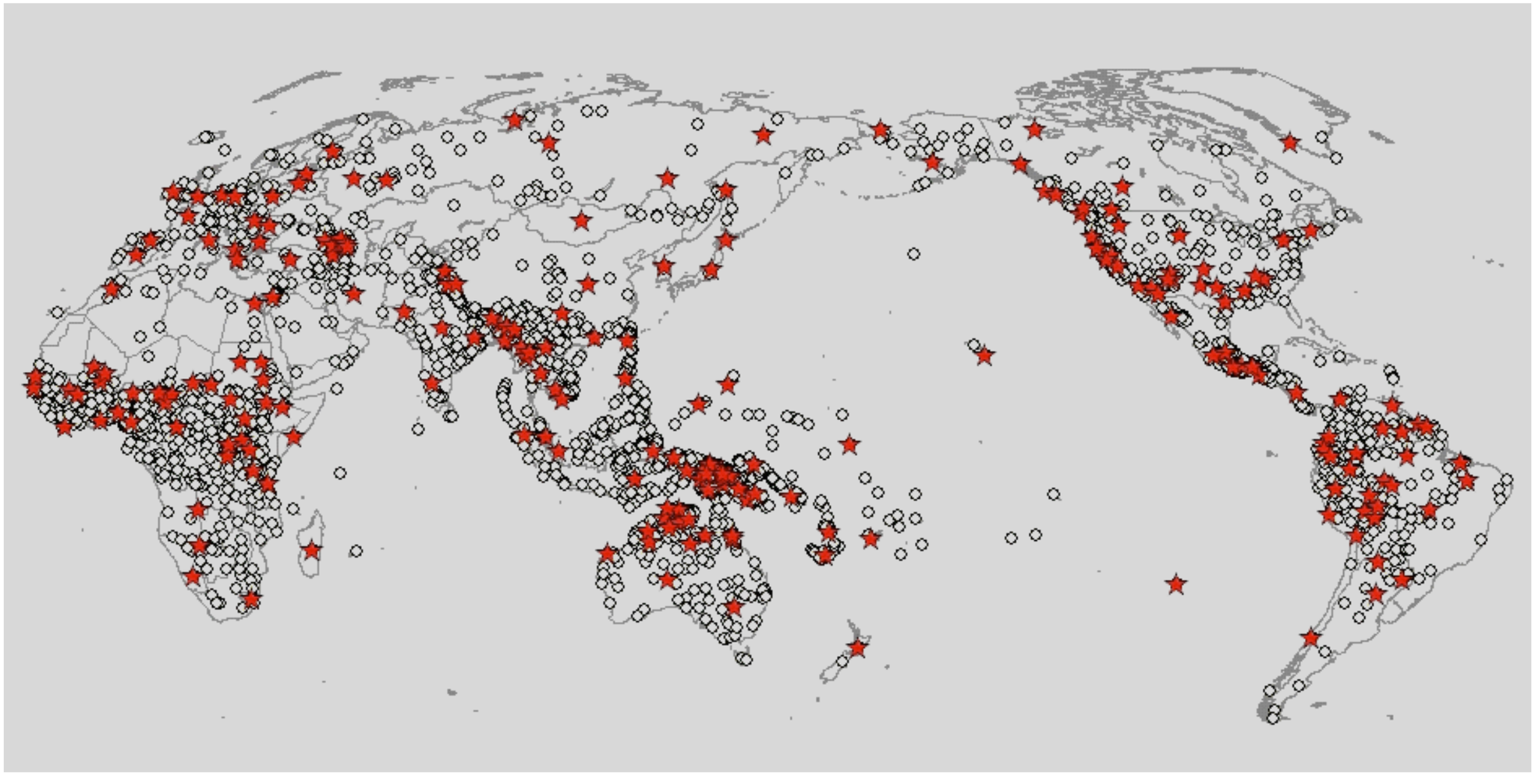


Evaluation of **Sampling** (by Östen Dahl)

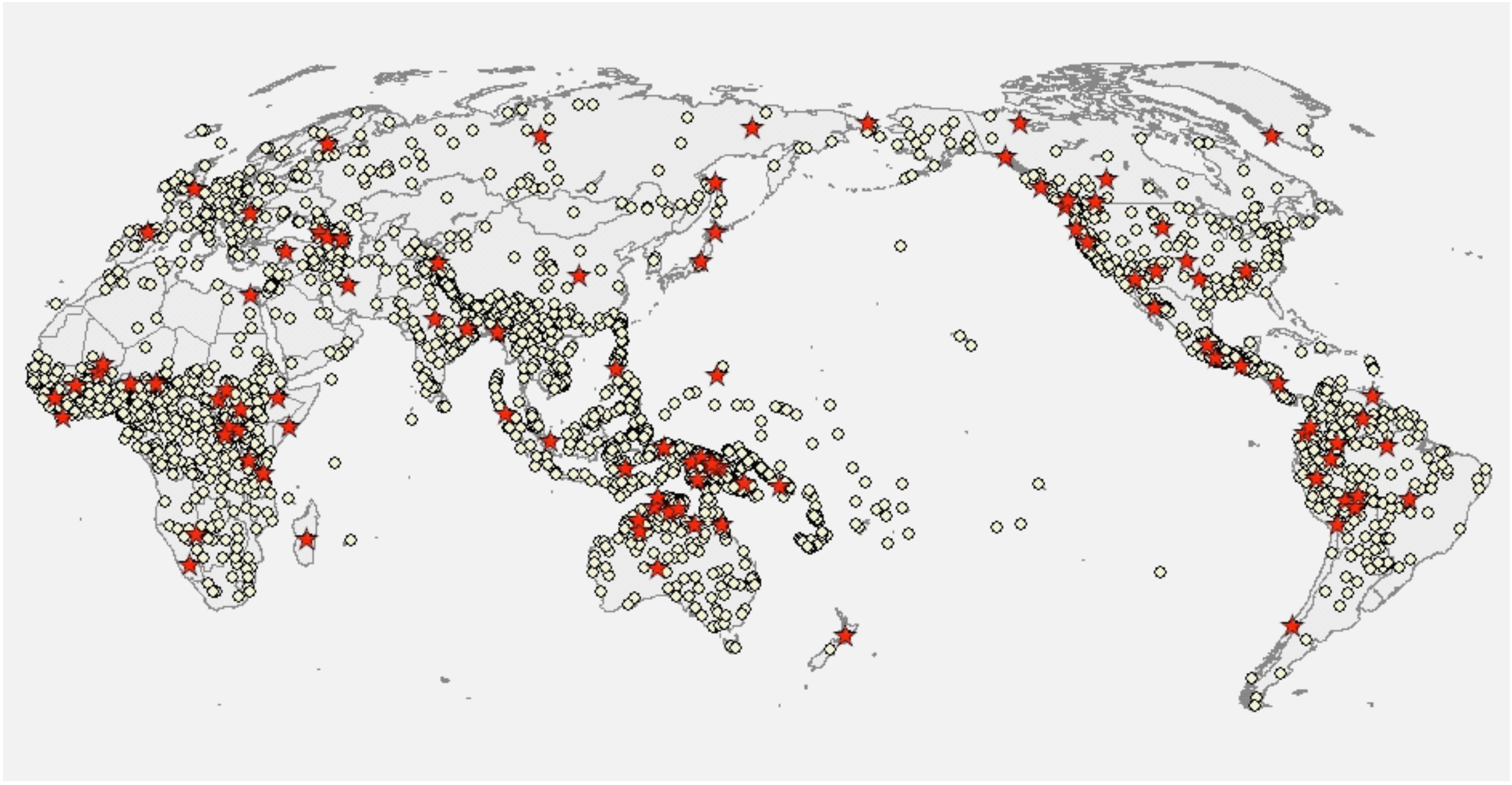
Which languages to investigate ?

- Typological samples normally take one language from each genealogical group
- But: genealogical groups are mostly based on lexical & phonological characteristics
- Is this a good basis to investigate syntax or morphology ?

Genealogical
200 language sample

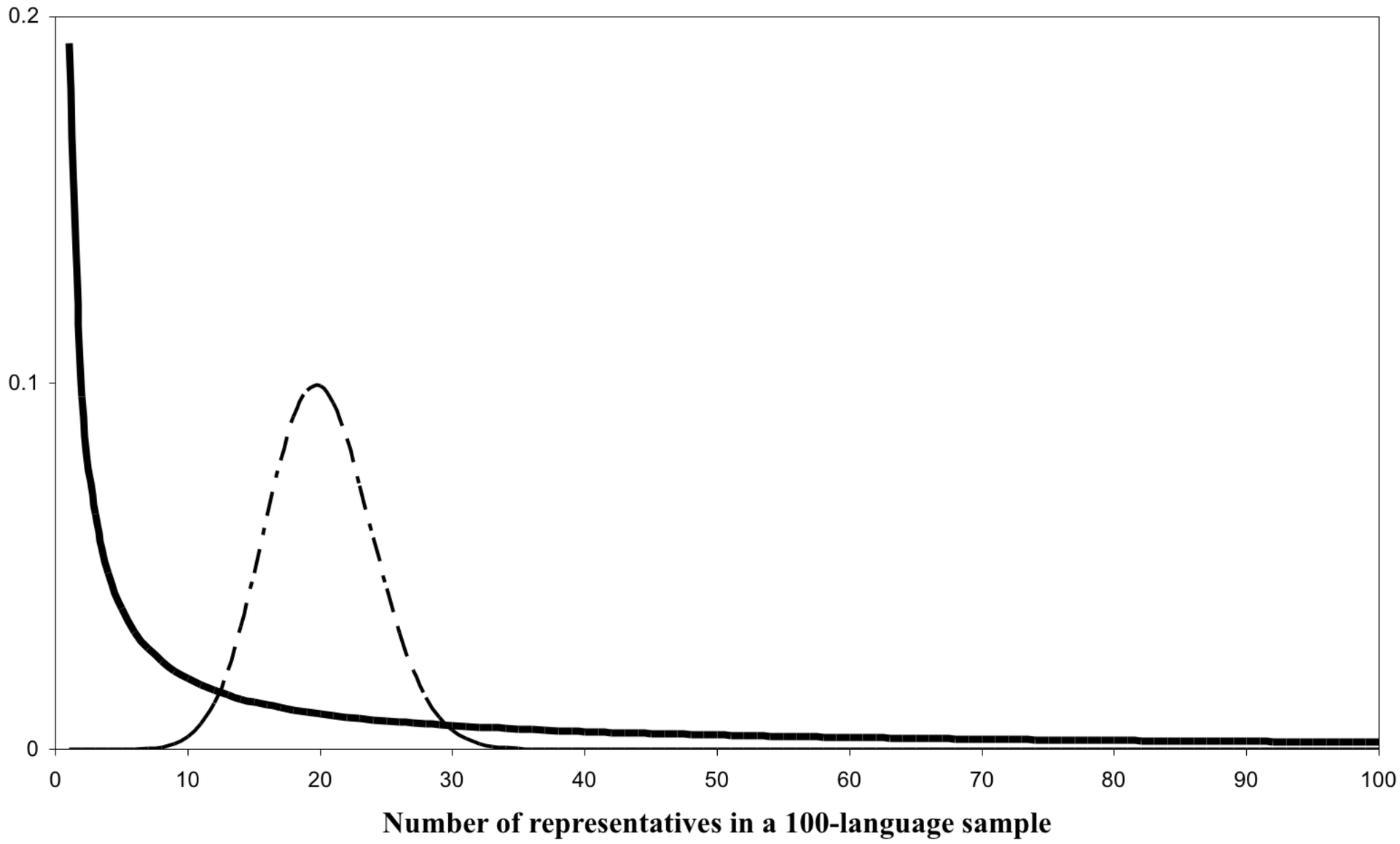


Typological 100 language sample

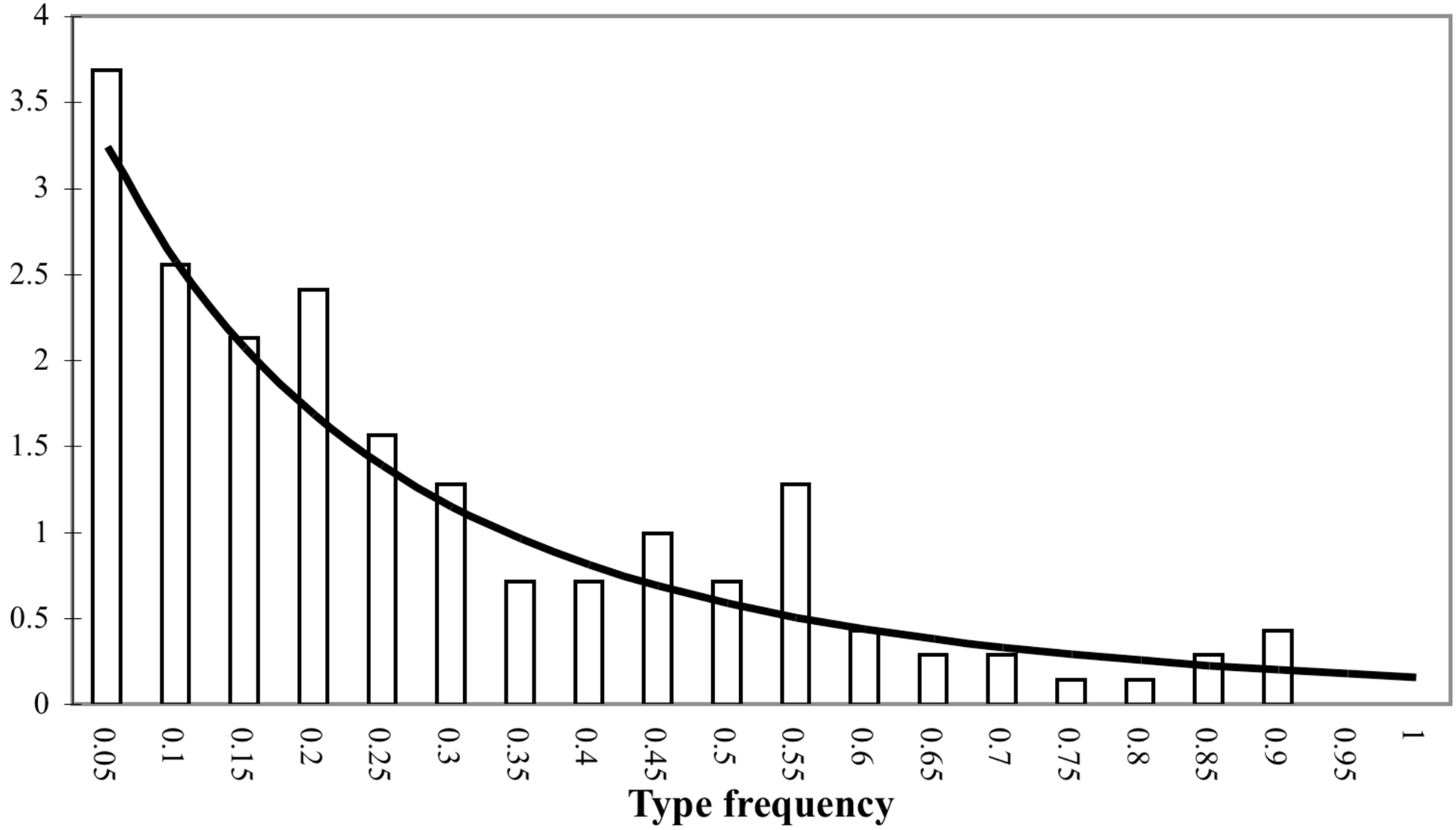


Evaluation of **Expectations** (by Elena Maslova)

With five values, what do we expect to find ?



WALS lets us find out what we should expect



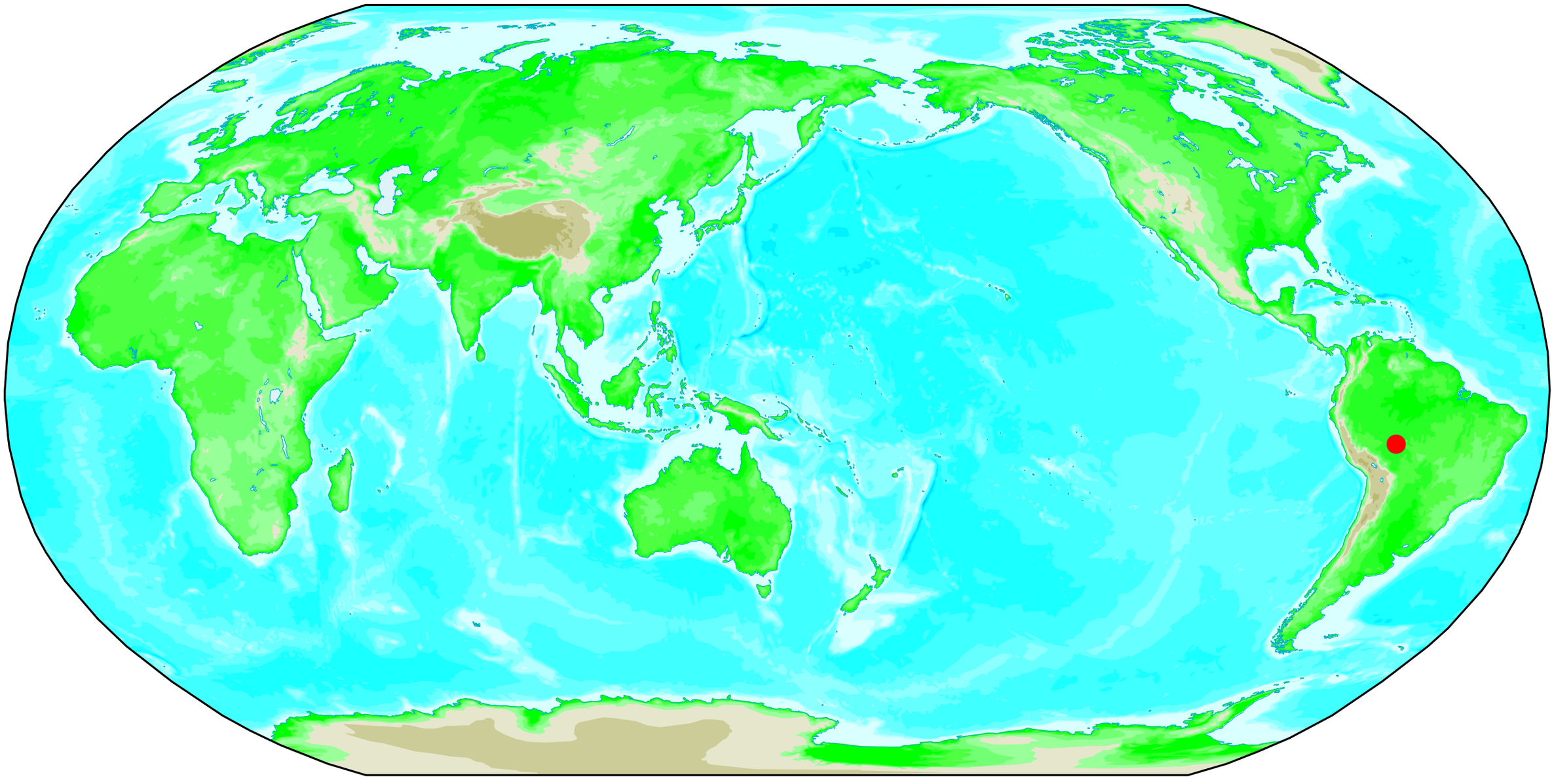
Investigating **rare characteristics**

And the winners are:

In the category:

‘Most Unusual Individual Language’

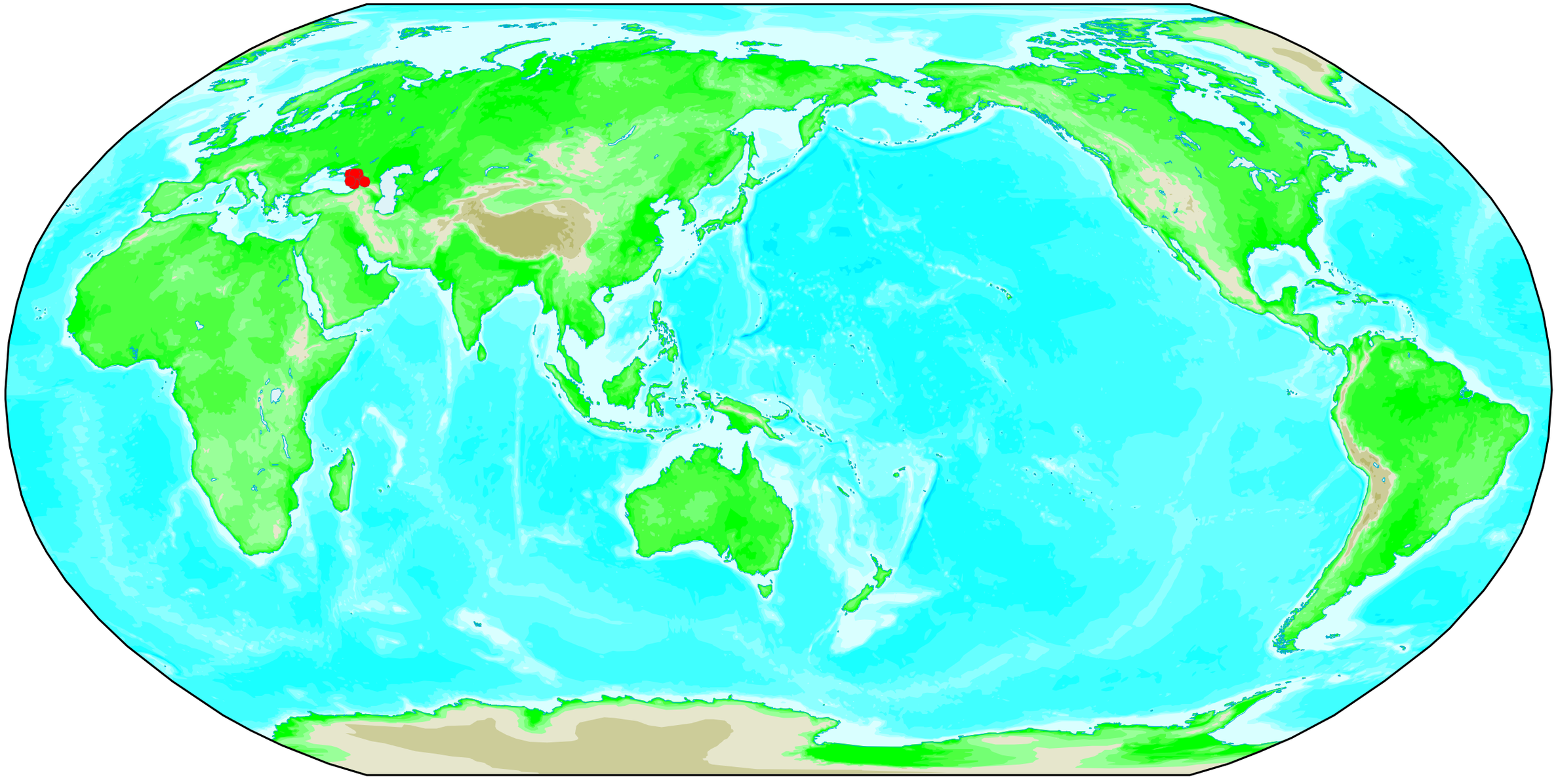
Wari’



In the category:

‘Most Unusual Genealogical Group’

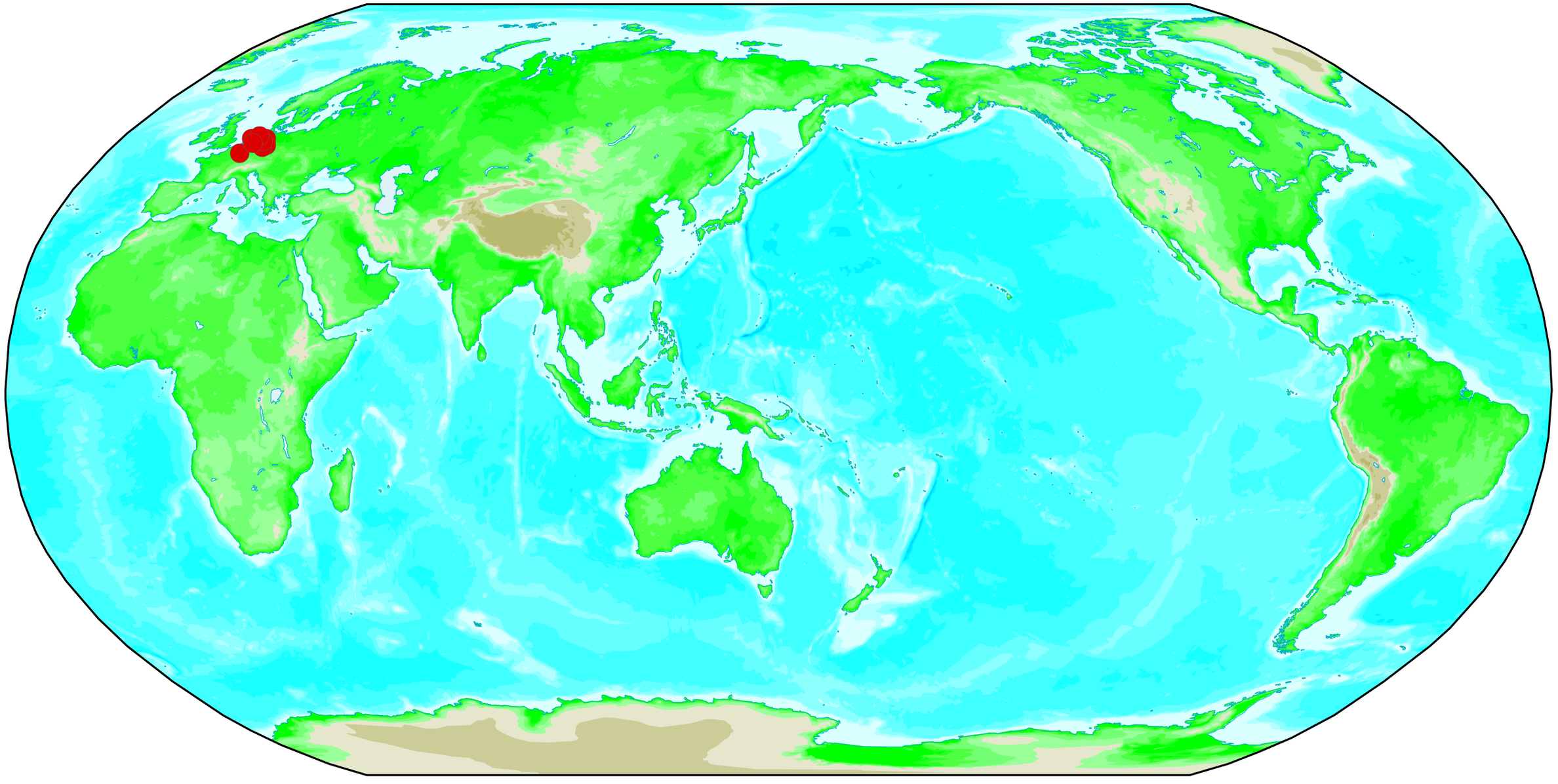
Northwest Caucasian



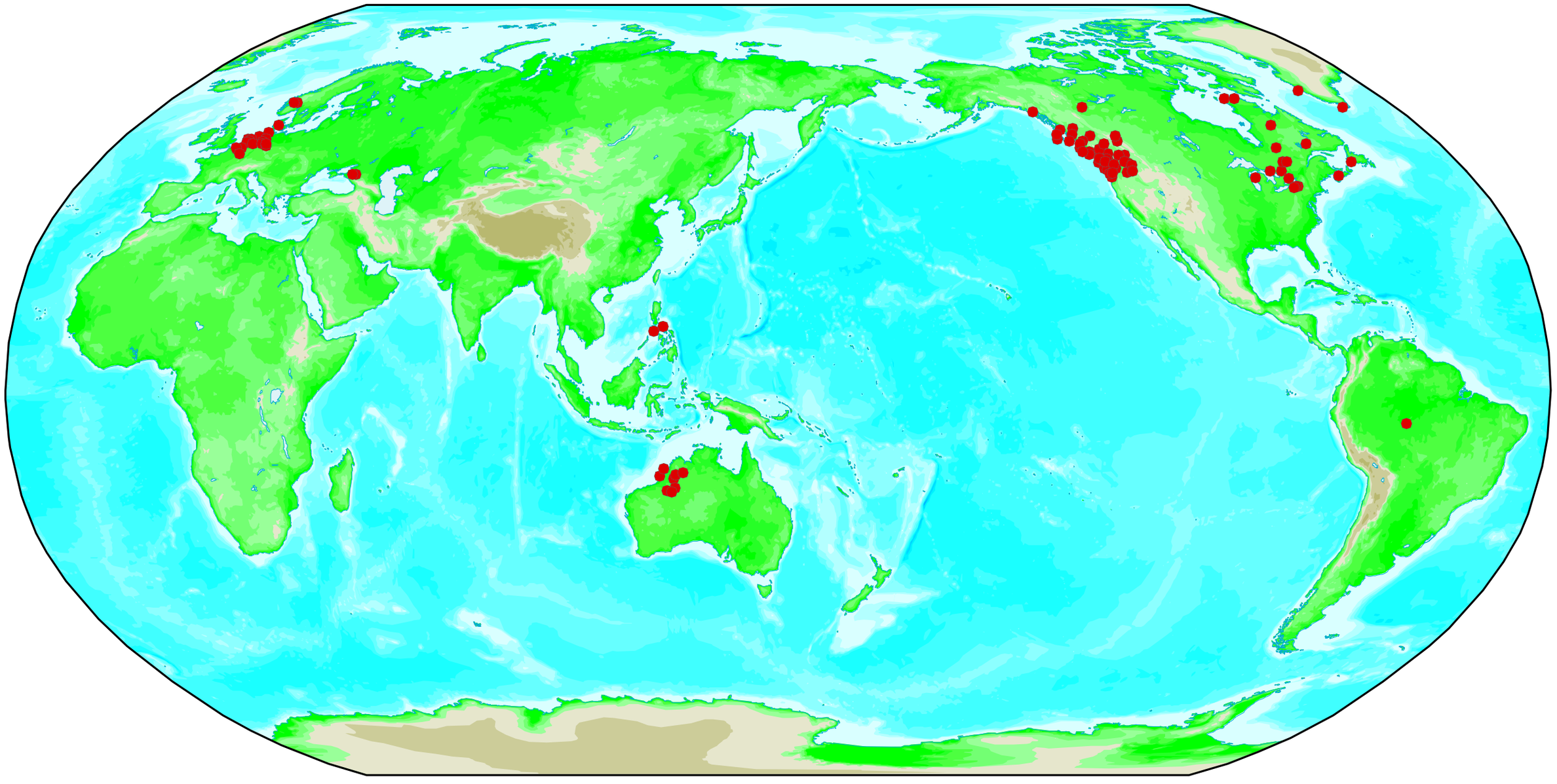
In the category:

‘Most Unusual Geographical Area’

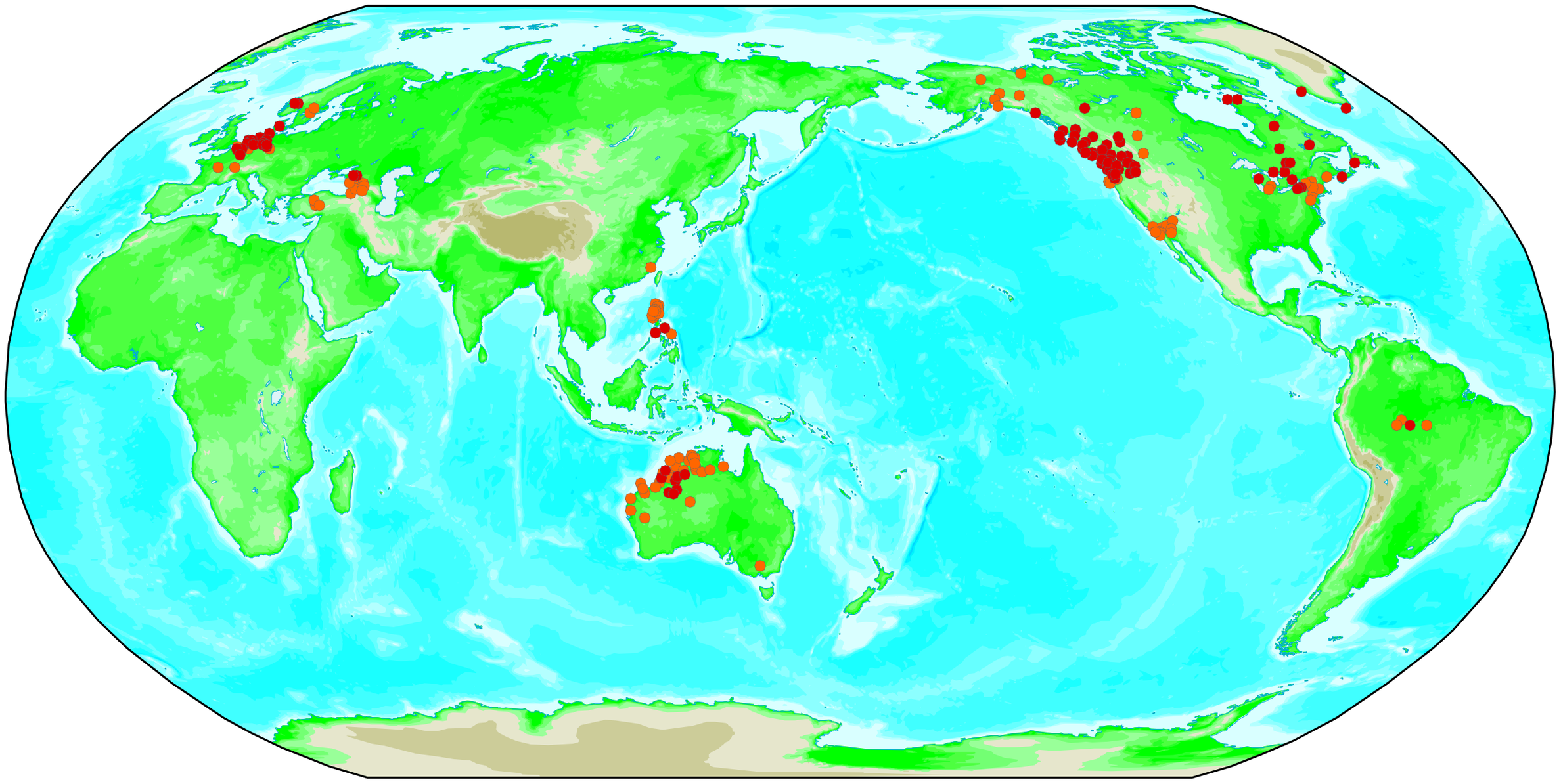
Northwest Continental Europe



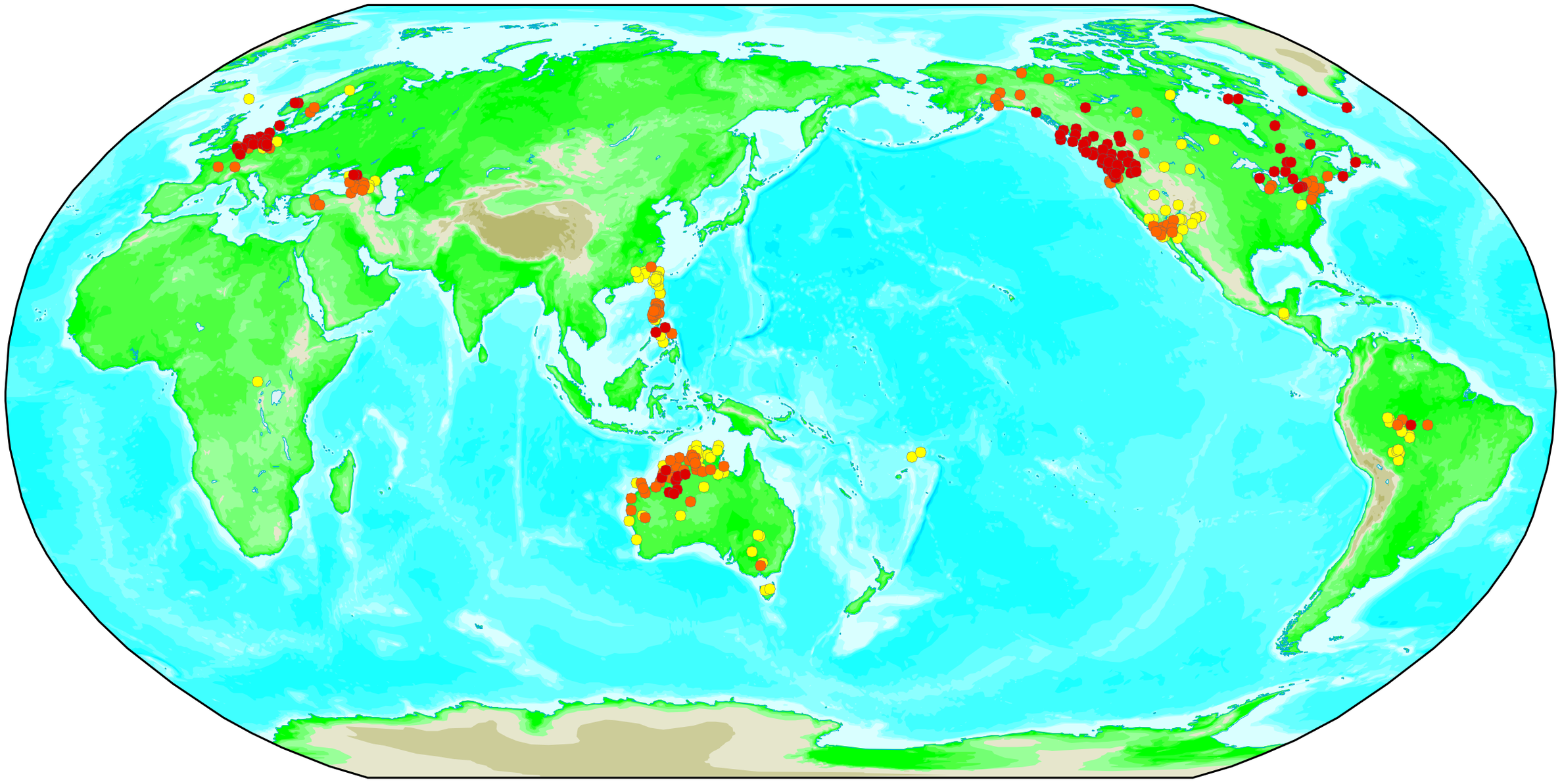
Top 100



Top 200

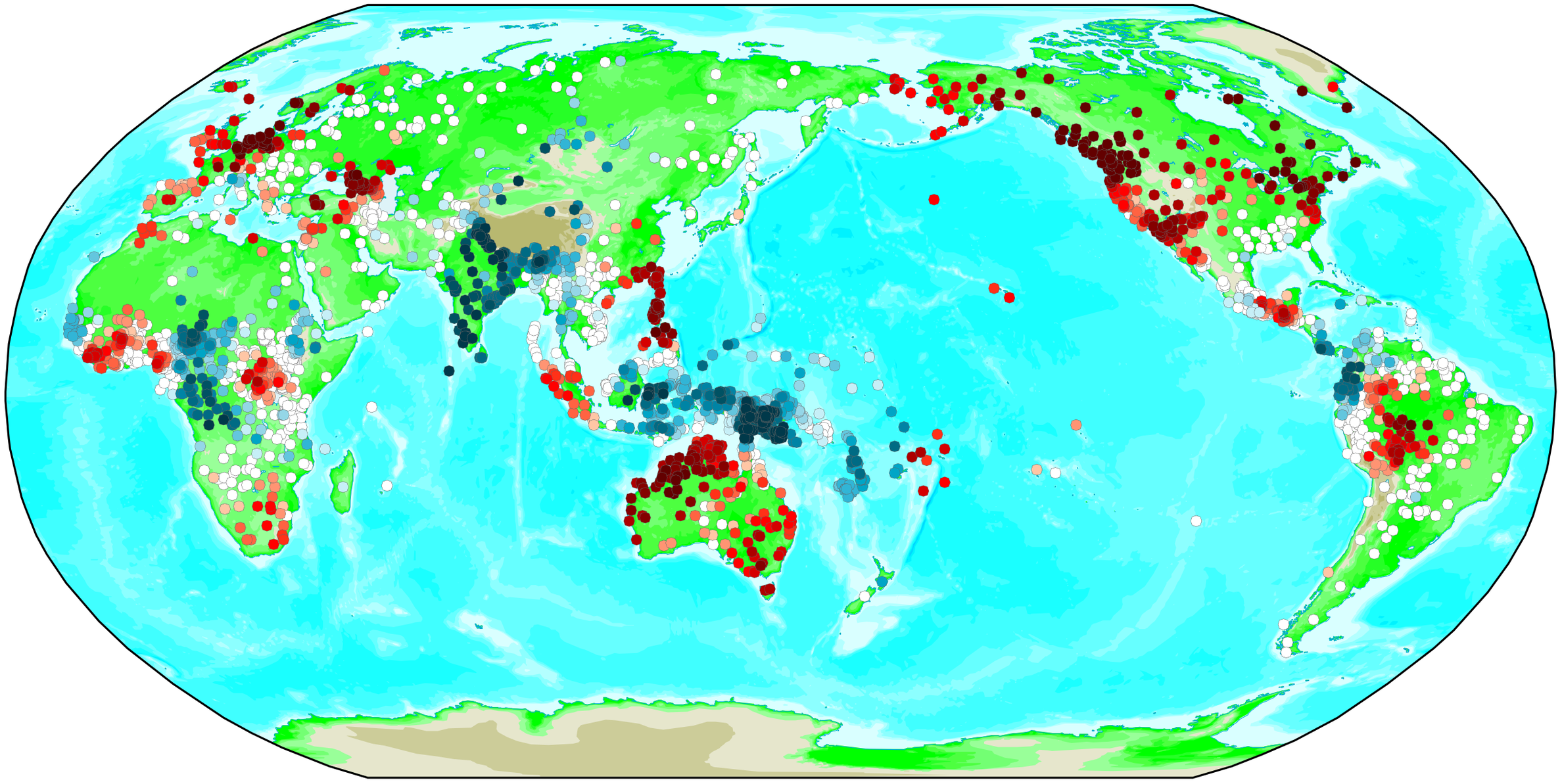


Top 300

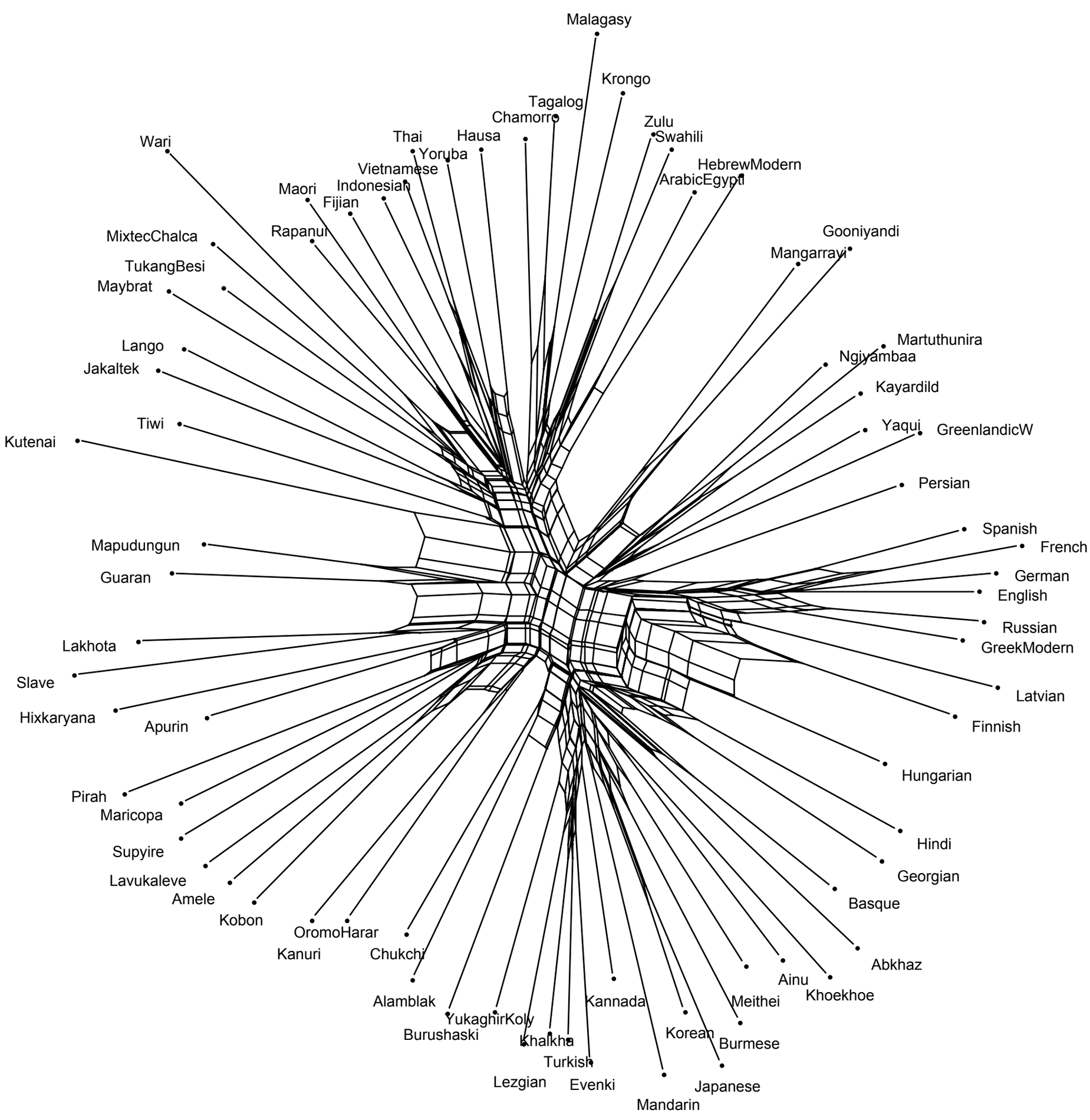


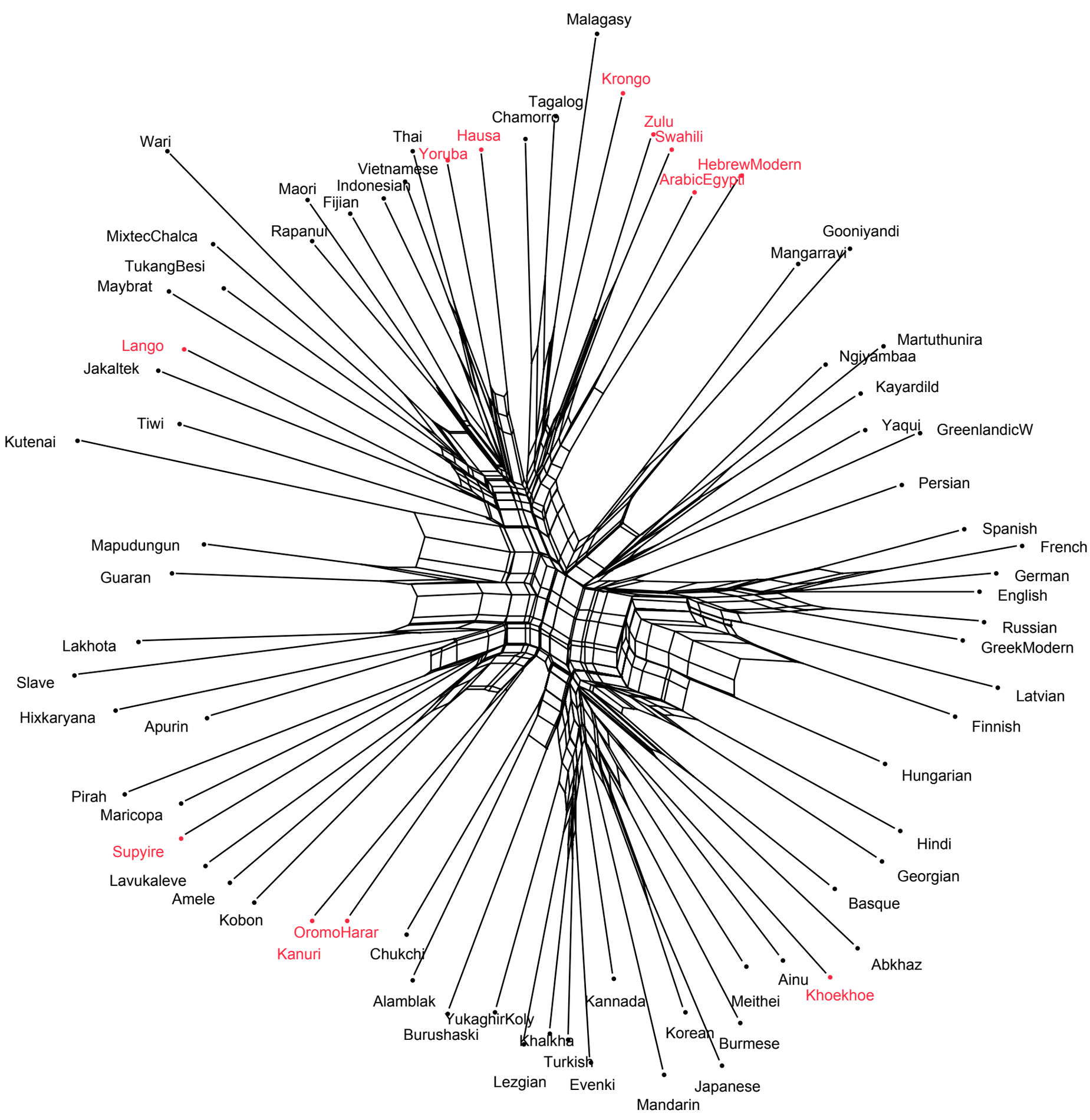
All languages

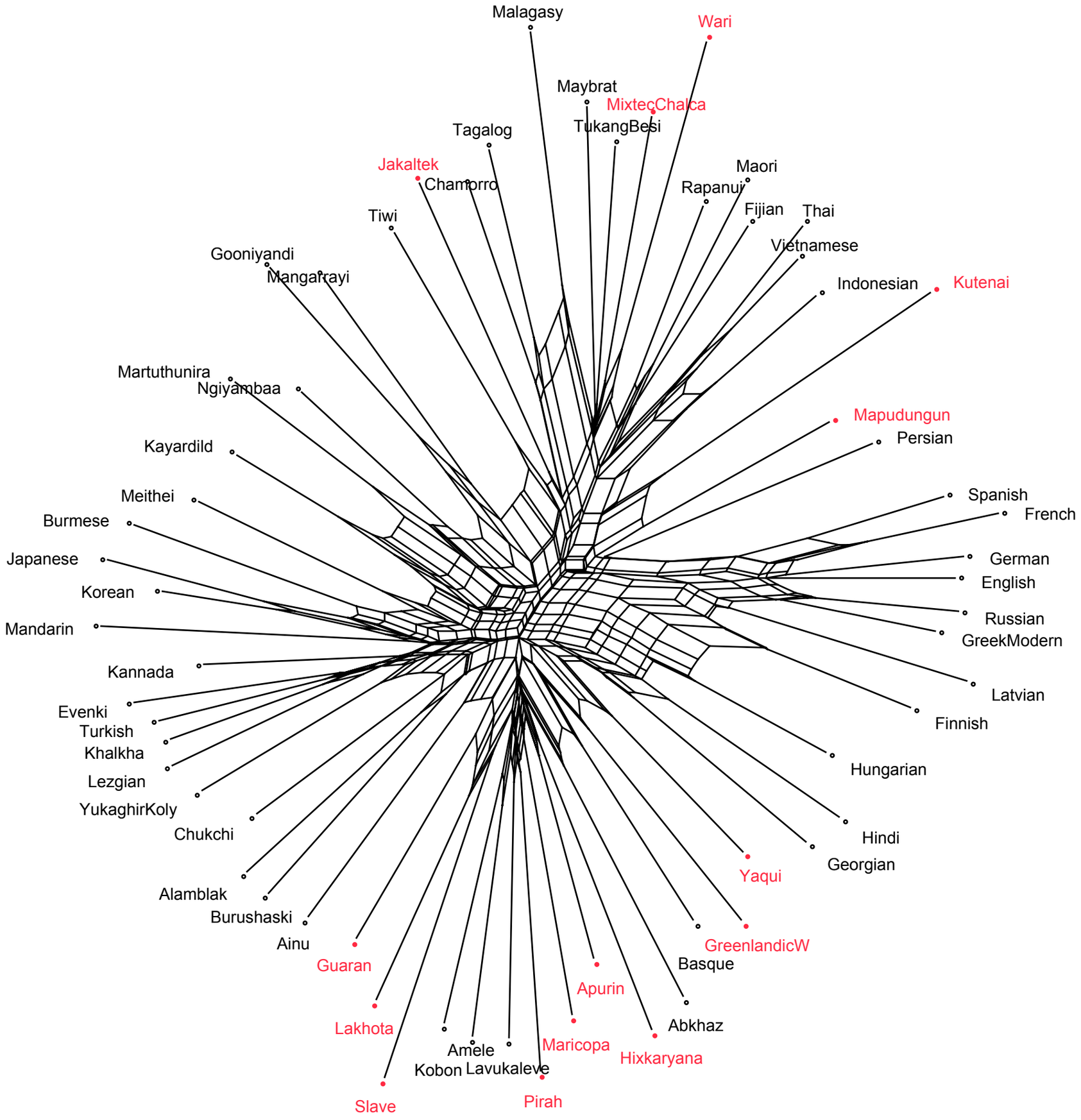
(red = rare, blue = common)

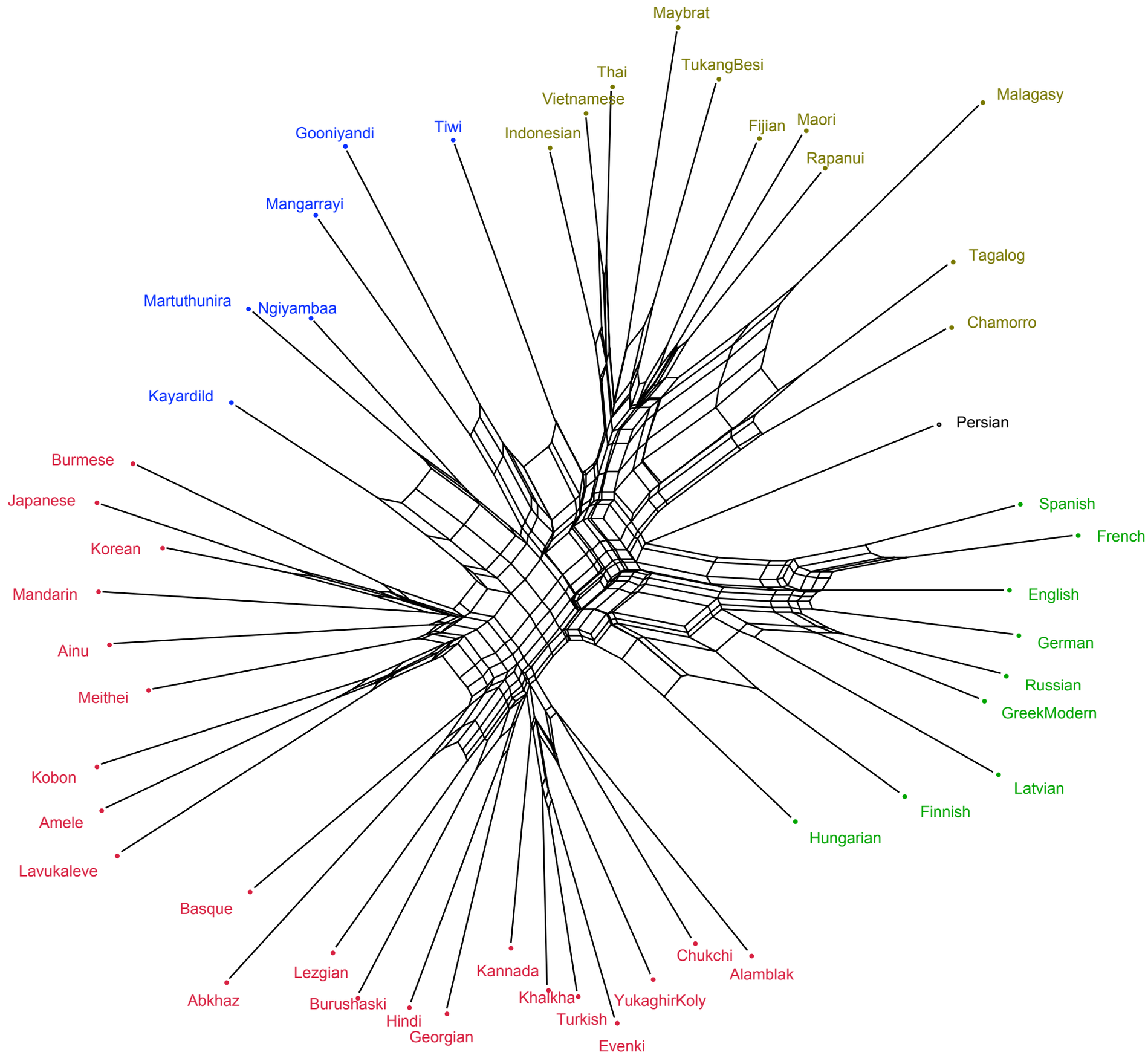


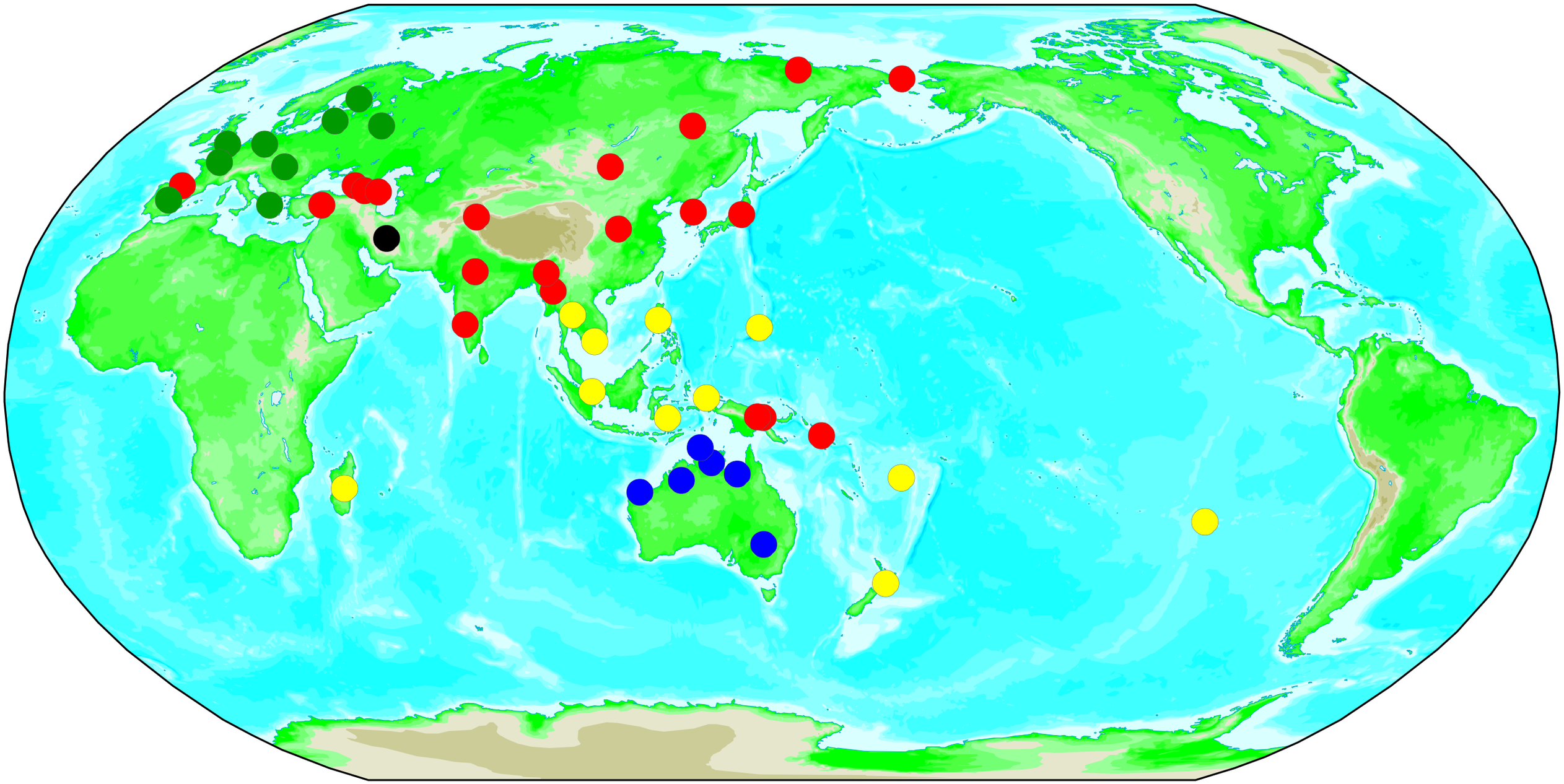
Investigating **language similarity**


Malagasy
Krongo
Tagalog
Chamorro
Zulu
Swahili
Thai
Hausa
Wari
Yoruba
HebrewModern
Vietnamese
ArabicEgypt
Maori
Indonesian
Fijian
Gooniyandi
Rapanui
MixtecChalca
Mangarrayi
TukangBesi
Maybrat
Martuthunira
Lango
Ngiyambaa
Jakaltek
Kayardild
Yaqui
Tiwi
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
Amele
Abkhaz
Kobon
Ainu
OromoHarar
Khoekhoe
Kanuri
Chukchi
Alamblak
Kannada
Meithei
YukaghirKoly
Korean
Burushaski
Burmese
Khalkha
Turkish
Lezgian
Evenki
Japanese
Mandarin

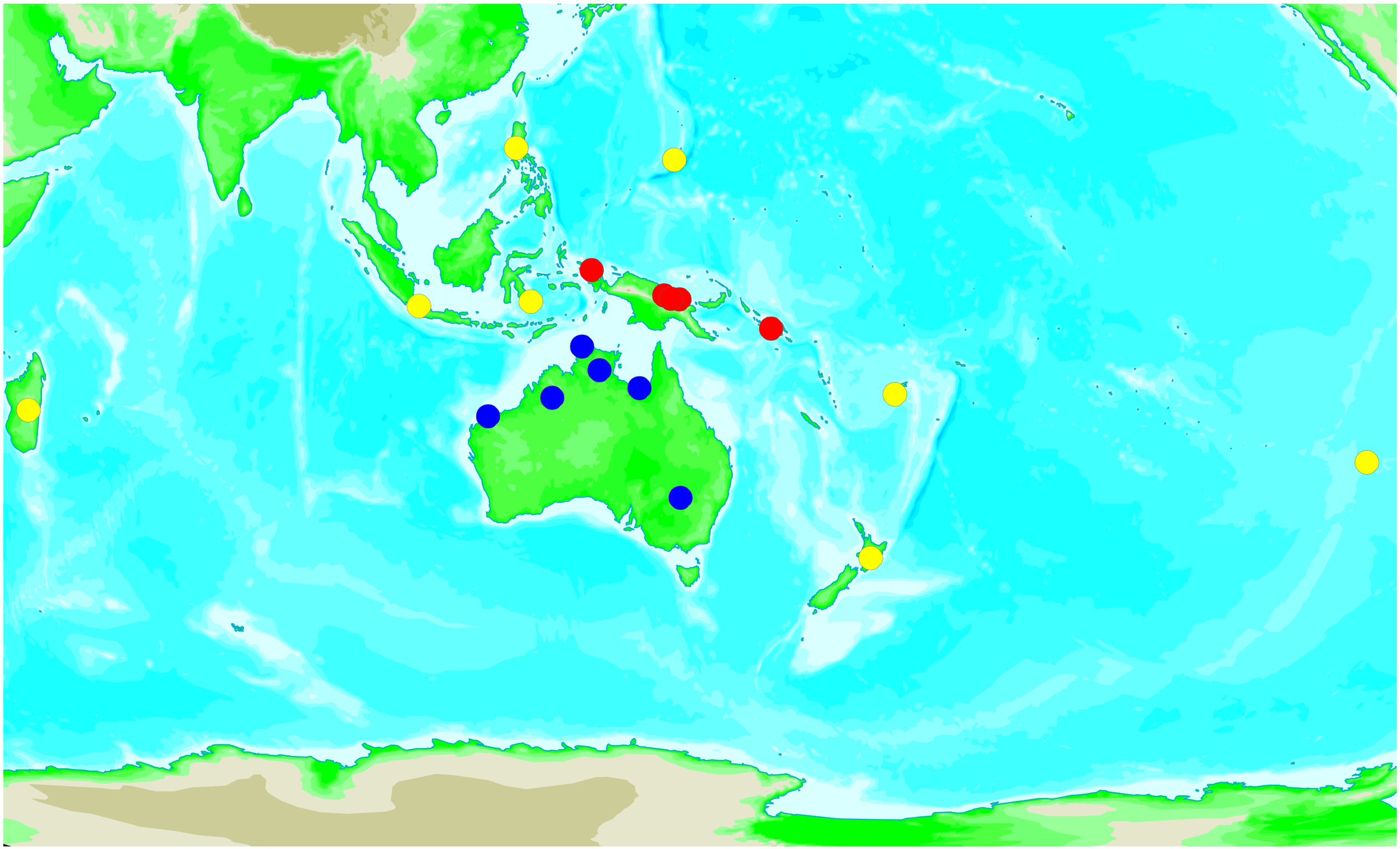

Malagasy
Krongo
Tagalog
Chamorro
Zulu
Swahili
Thai
Hausa
Yoruba
Wari
Vietnamese
HebrewModern
ArabicEgypt
Maori
Indonesian
Fijian
Rapanui
Gooniyandi
MixtecChalca
Mangarrayi
TukangBesi
Maybrat
Martuthunira
Lango
Ngiyambaa
Jakaltek
Kayardild
Tiwi
Yaqui
GreenlandicW
Kutenai
Persian
Spanish
Mapudungun
French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

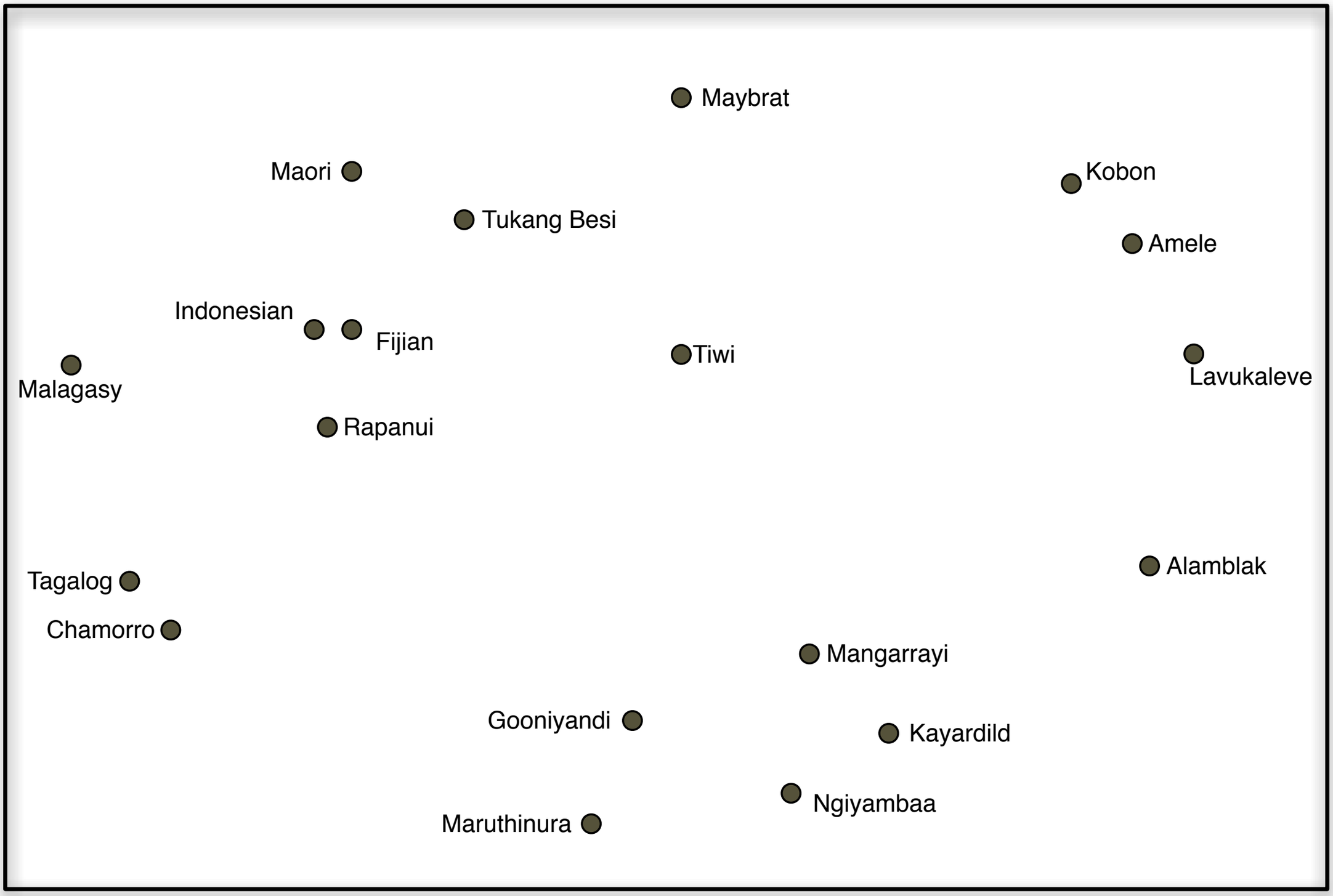

Maybrat
TukangBesi
Thai
Vietnamese
Malagasy
Gooniyandi
Tiwi
Indonesian
Fijian
Maori
Rapanui
Mangarrayi
Tagalog
Martuthunira
Ngiyambaa
Chamorro
Kayardild
Persian
Burmese
Spanish
Japanese
French
Korean
English
Mandarin
German
Ainu
Russian
Meithei
GreekModern
Kobon
Latvian
Amele
Finnish
Lavukaleve
Hungarian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



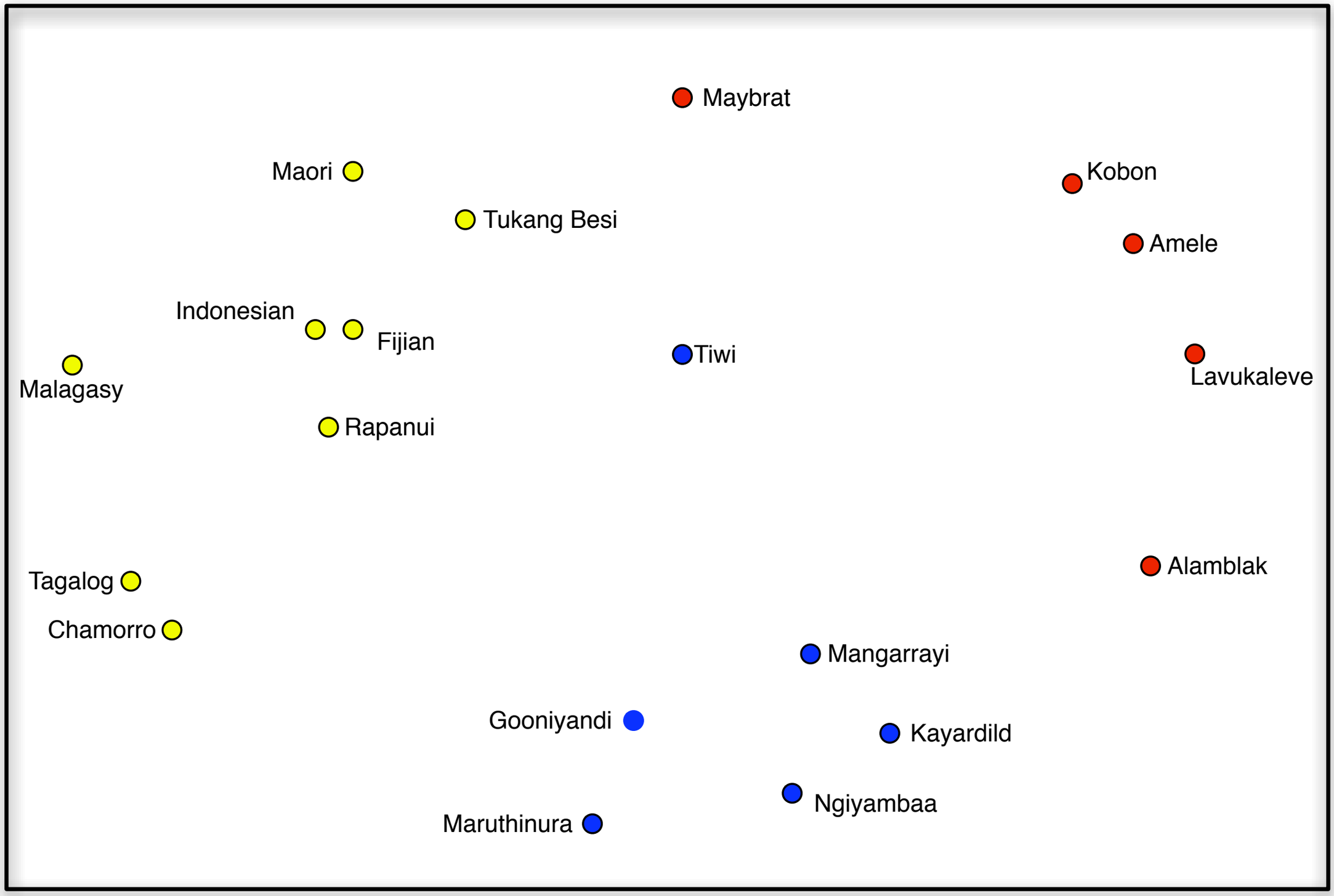
A closer look at geography: the case of **Oceania**



MDS of typological distances

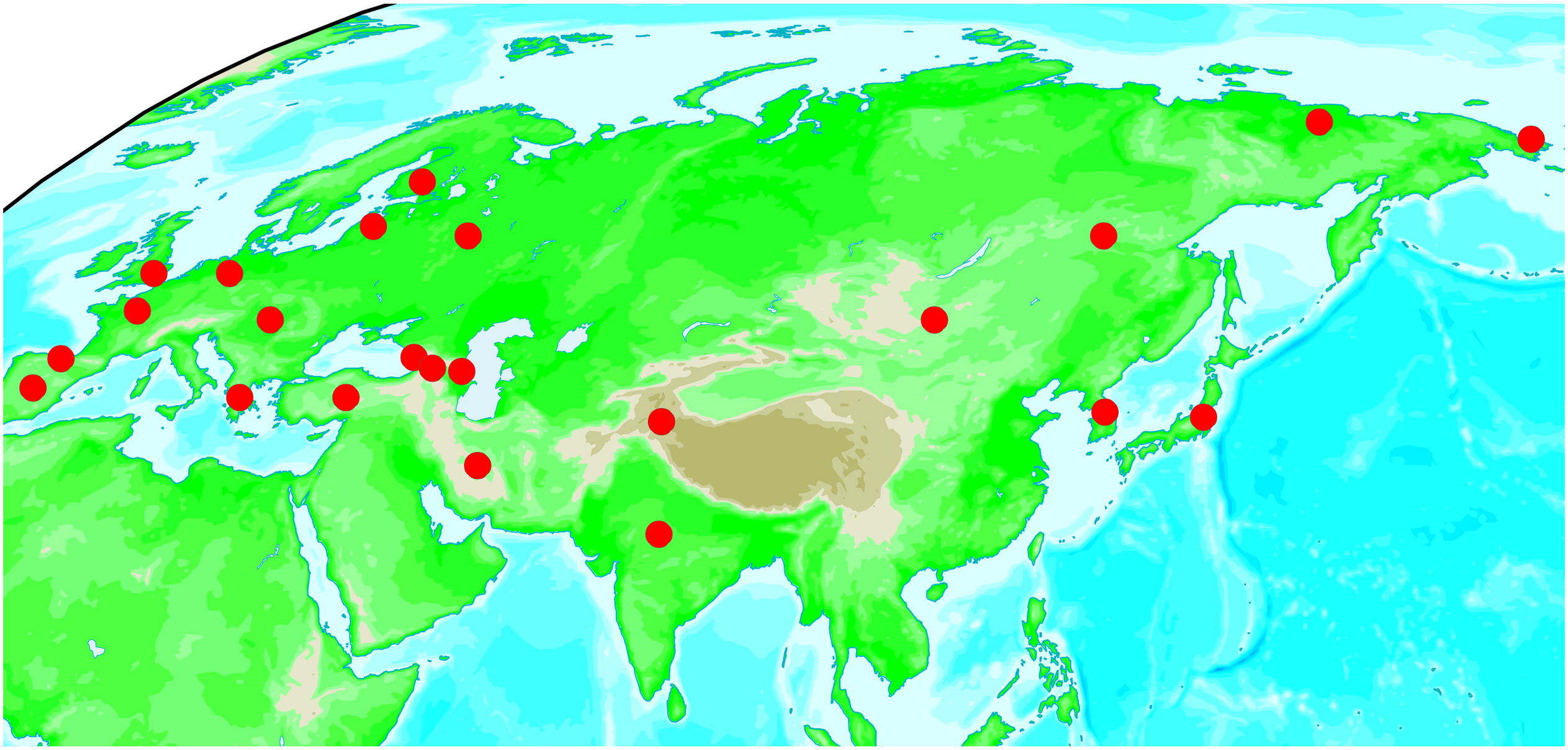


MDS of typological distances

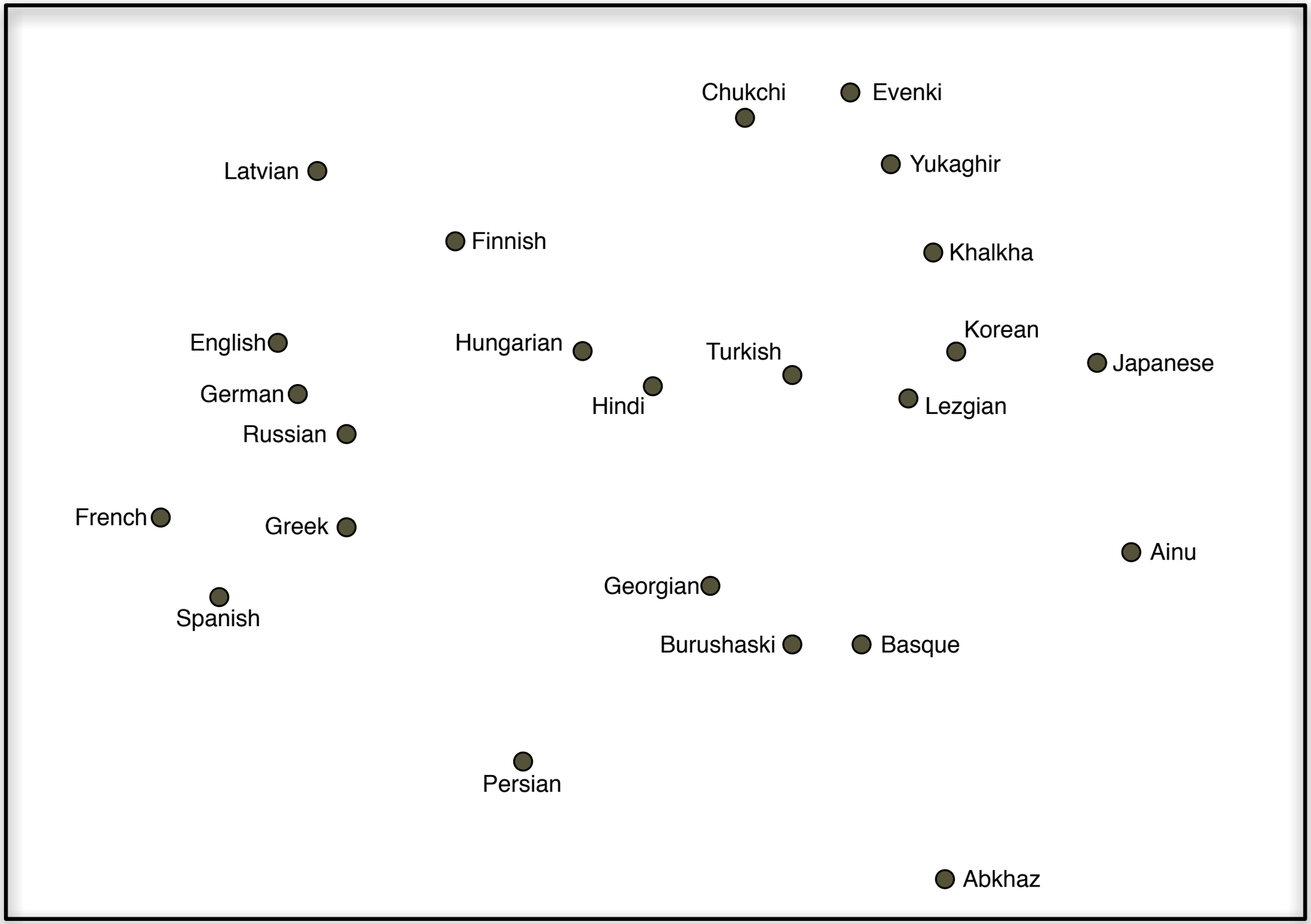


Maybrat
Kobon
Tukang Besi
Amele
Tiwi
Lavukaleve
Alamblak
Mangarrayi
Gooniyandi
Kayardild
Ngiyambaa
Maruthinura

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



MDS of typological distances



WALS forever !