

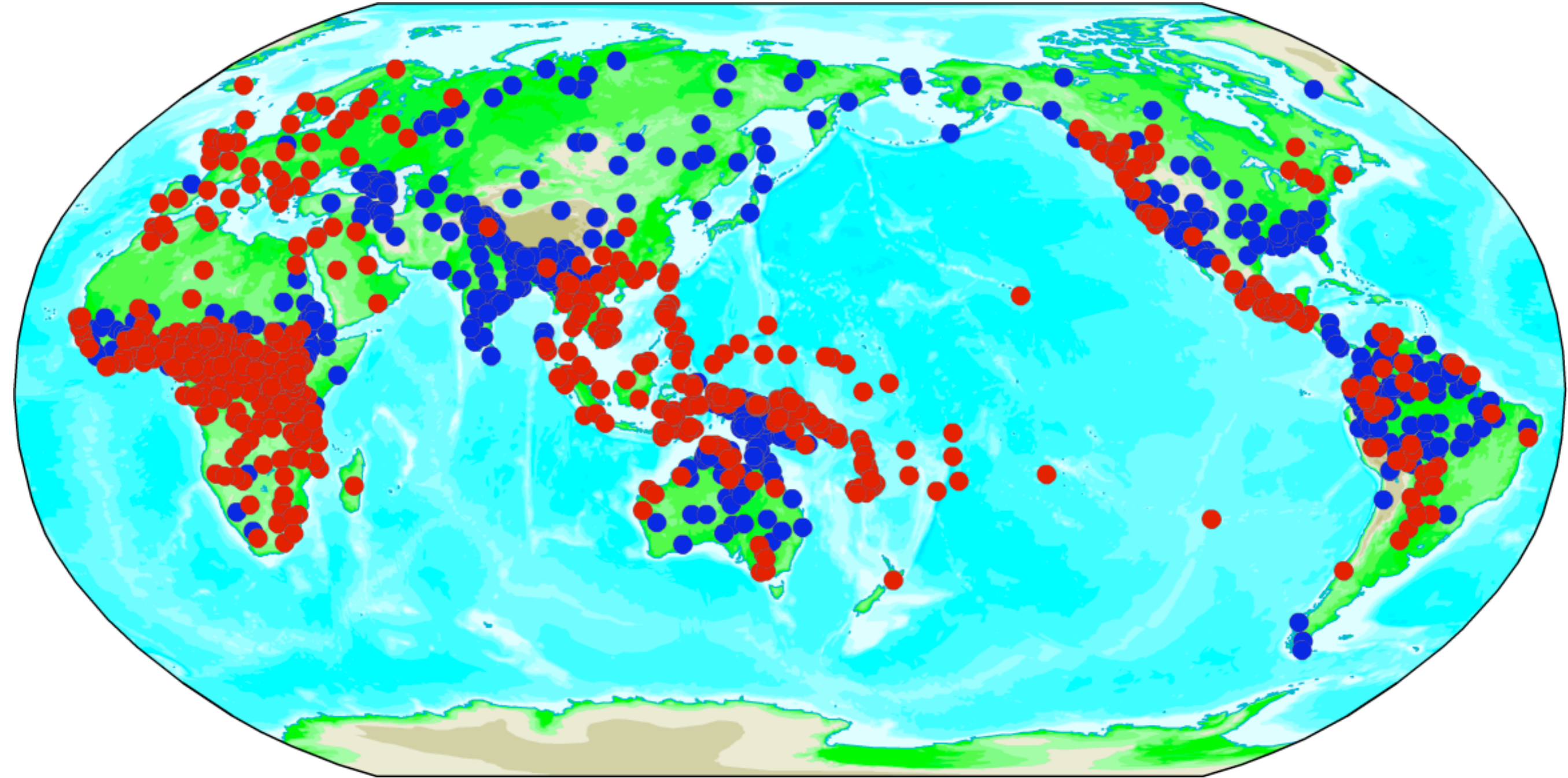
A world map with a light blue background. The map is overlaid with numerous semi-transparent colored dots in shades of purple, green, and red. These dots are scattered across the continents, with a higher density in Europe and Africa. The dots represent typological frequencies as mentioned in the title.

# **On the probability distribution of typological frequencies**

Michael Cysouw

Max Planck Institute for Evolutionary Anthropology, Leipzig

# VO vs. OV word order



Dryer, Matthew S. (2005) 'Order of object and verb' in Martin Haspelmath et al. (eds.)  
*World Atlas of Language Structures* (Oxford: Oxford University Press), 338-41.

	Verb - Object	Object - Verb	Total
Noun - Relative Clause	<b>370</b>	<b>96</b>	<b>466</b>
Relative Clause - Noun	<b>5</b>	<b>109</b>	<b>114</b>
Total	<b>375</b>	<b>205</b>	<b>580</b>

	Verb - Object	Object - Verb	Total
Noun - Relative Clause	<b>63.8%</b>	<b>16.6%</b>	<b>80.3%</b>
Relative Clause - Noun	<b>0.9%</b>	<b>18.8%</b>	<b>19.7%</b>
Total	<b>64.7%</b>	<b>35.3%</b>	<b>100%</b>

	Verb - Object	Object - Verb	Total
Noun - Relative Clause	<b>+ 3.96</b>	<b>- 5.35</b>	
Relative Clause - Noun	<b>- 8.00</b>	<b>+ 10.82</b>	
Total			

“In a representative sample of languages, if no universal were involved, i.e. if the distribution of types along some parameter were purely random, then we would expect each type to have roughly an equal number of representatives. To the extent that the actual distribution departs from this random distribution, the linguist is obliged to state and, if possible, account for this discrepancy.”

What kind of random distributions are we dealing with?

# Size of Phoneme Inventory

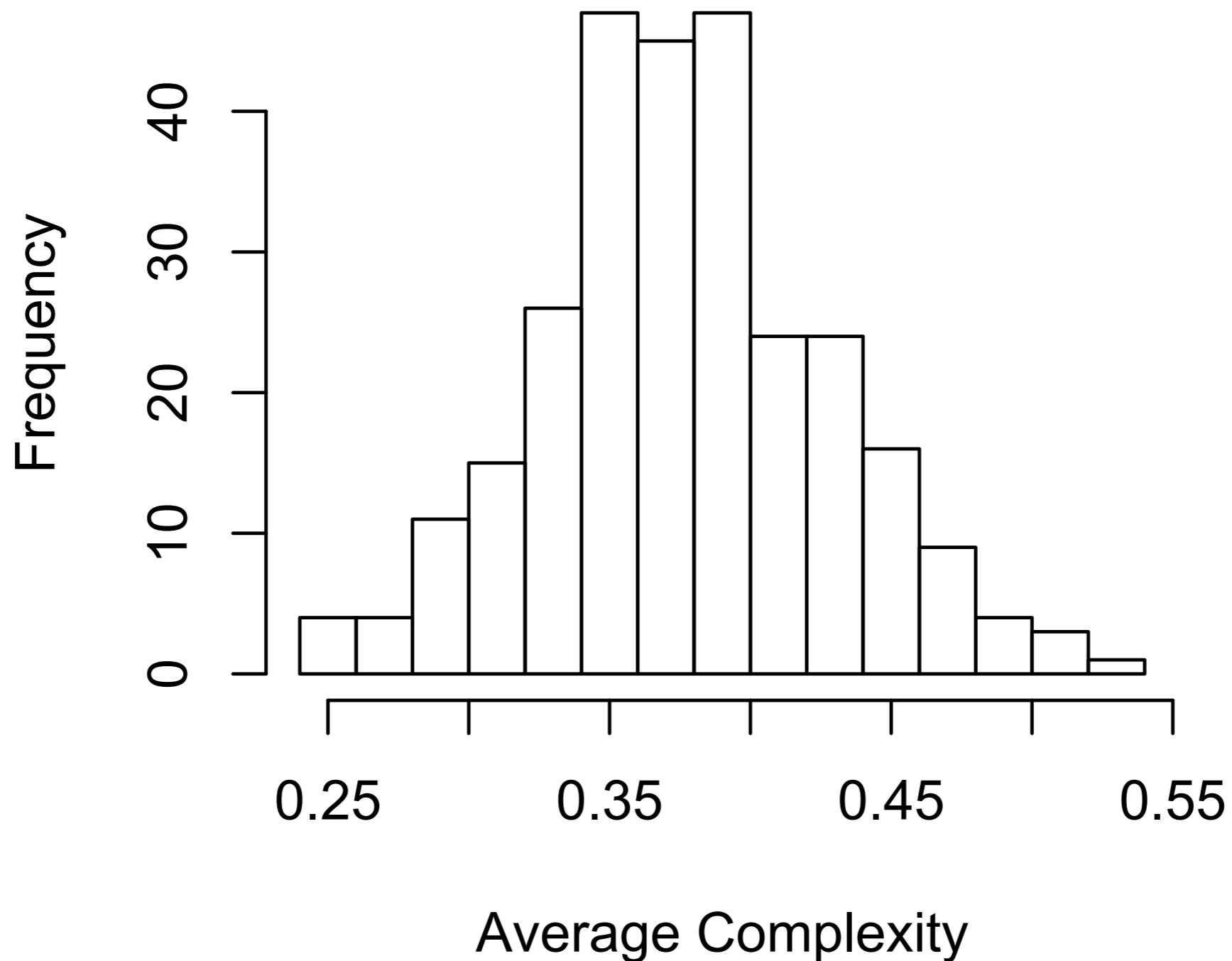
- Lehfeltdt (1975)  
**gamma distribution**
- Justeson & Stephens (1984)  
**log-normal distribution**
- Maddieson (2005)  
**normal distribution**



# Linguistic “Complexity”

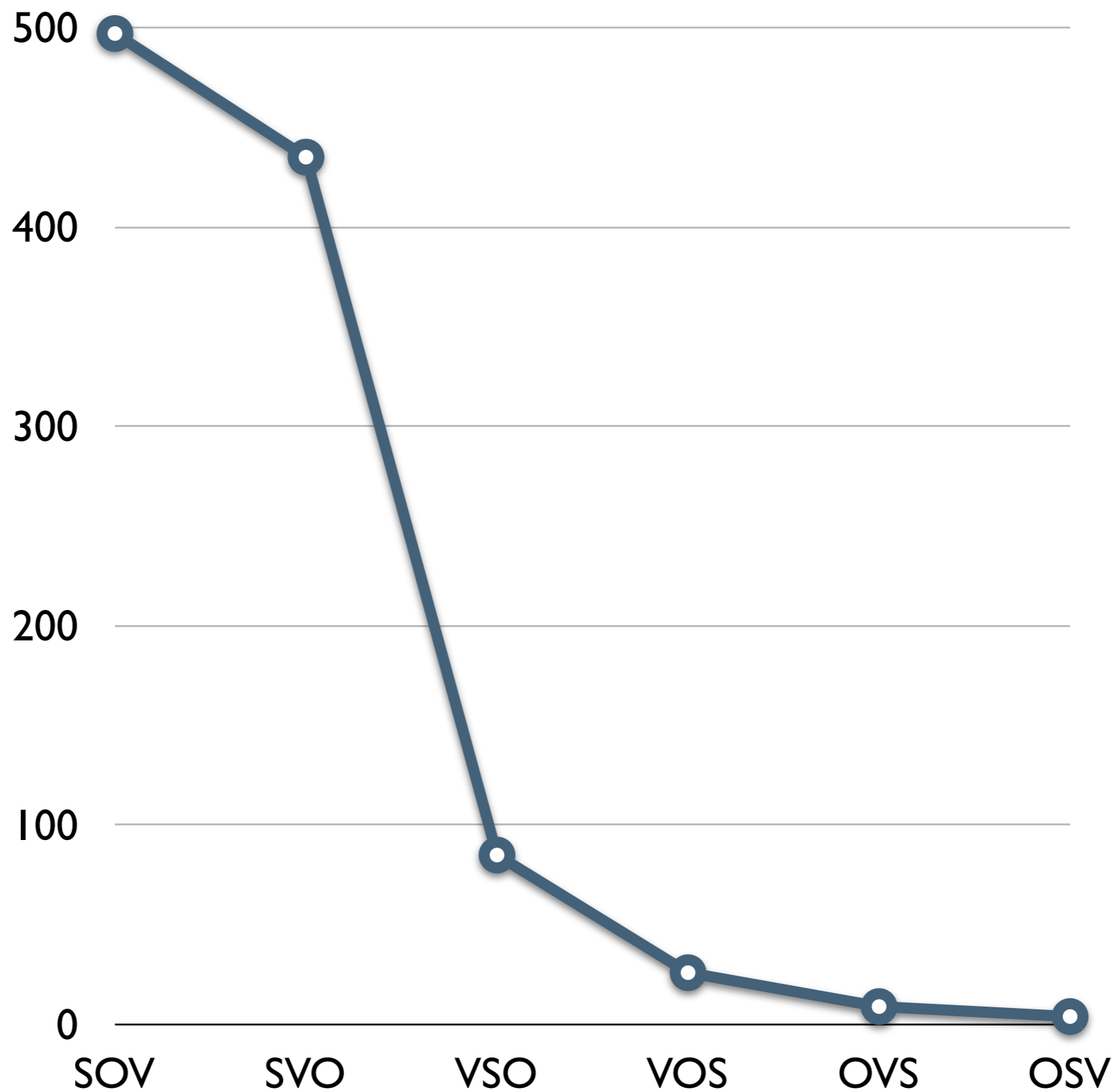
- Nichols (1992)  
**normal distribution**
- Nichols et al. (2006)  
**“bell-shaped” distribution**

# All data from WALS



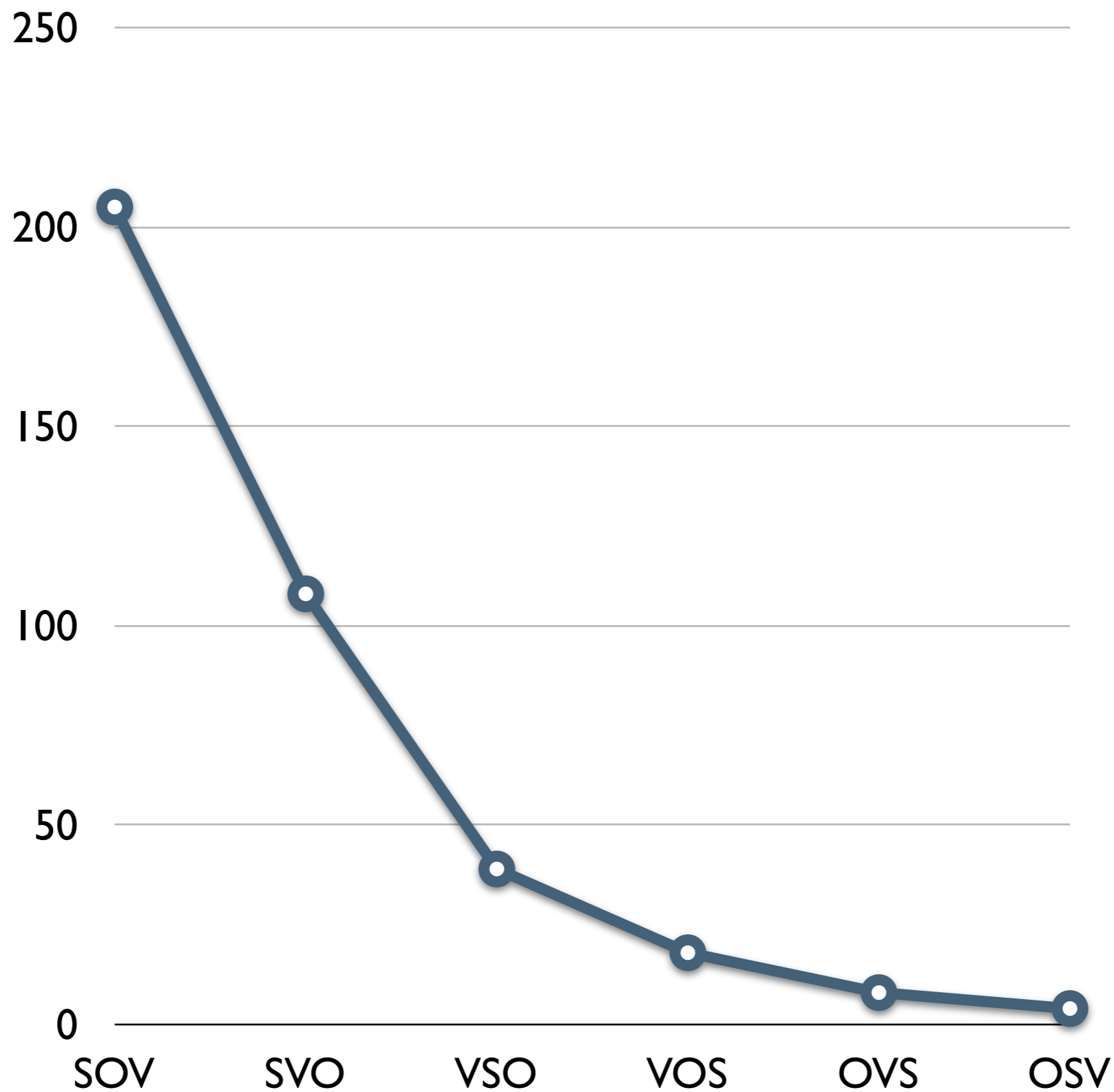
Shapiro-Wilk  $p = 0.02$

Languages



Dryer, Matthew S. (2005) Order of subject, object and verb' in Martin Haspelmath et al. (eds.) *World Atlas of Language Structures* (Oxford: Oxford University Press), 330-33.

Genera

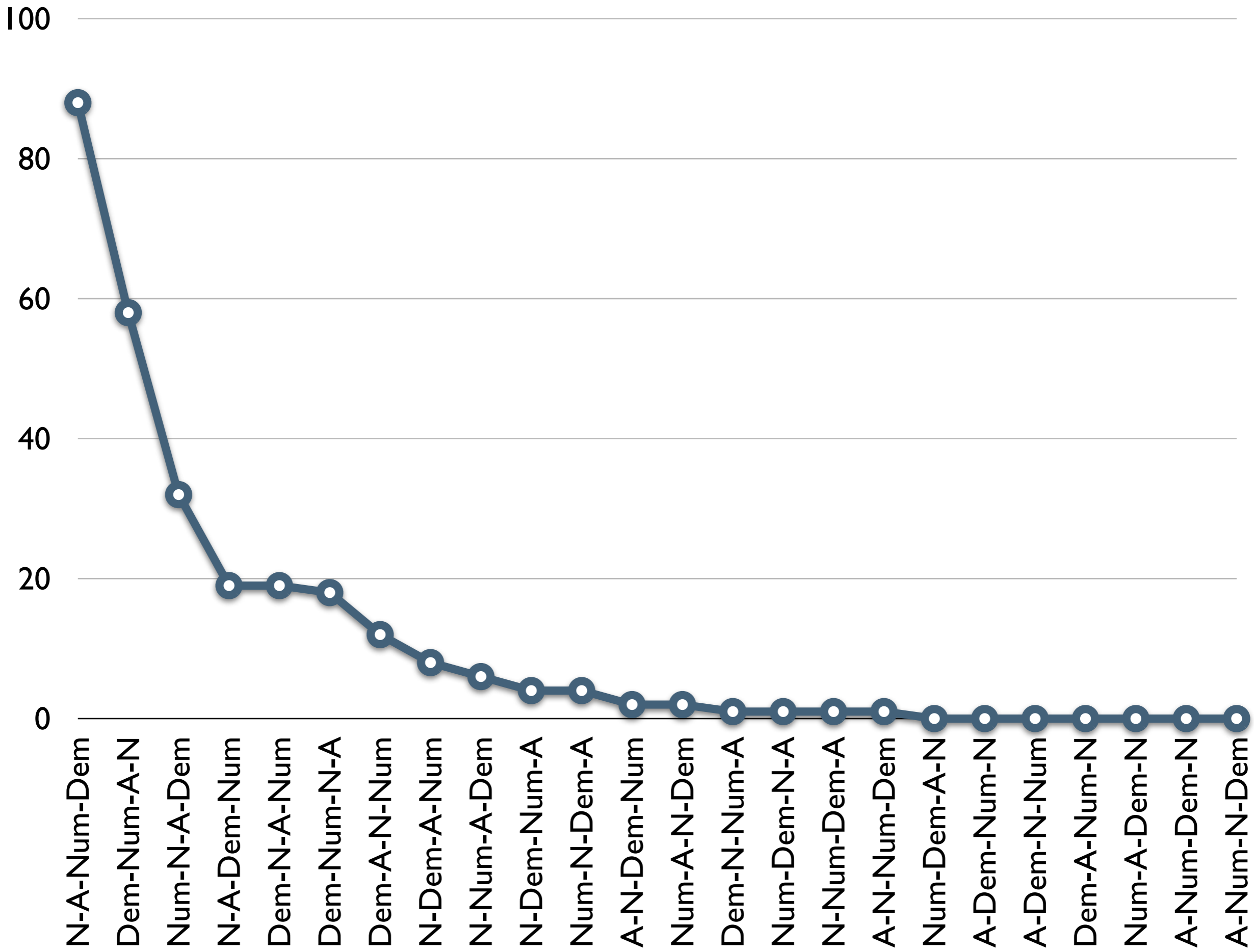


Dryer, Matthew S. (2005) Order of subject, object and verb' in Martin Haspelmath et al. (eds.) *World Atlas of Language Structures* (Oxford: Oxford University Press), 330-33.

# Noun Phrase word order

- “Those three red books”
- Demonstrative - Numeral - Adjective - Noun
- Dem-Num-A-N

Cinque, Guglielmo (2005) ‘Deriving Greenberg’s Universal 20 and Its Exceptions’ *Linguistic Inquiry* 36 (3), 315 - 332.  
Dryer, Matthew S. (2006) ‘On Cinque on Greenberg’s universal 20’  
Presentation at MPI-EVA Leipzig.



**Why should we believe any claimed distribution ?**

# Searching for distributions

- Empirical data won't help much
- Fitting data to a distribution always results in a more or less good fit.
- It is unclear what it would mean for an empirical distribution to have a “reasonably good fit”



We need some kind of theoretical notion about how typological distributions arise !

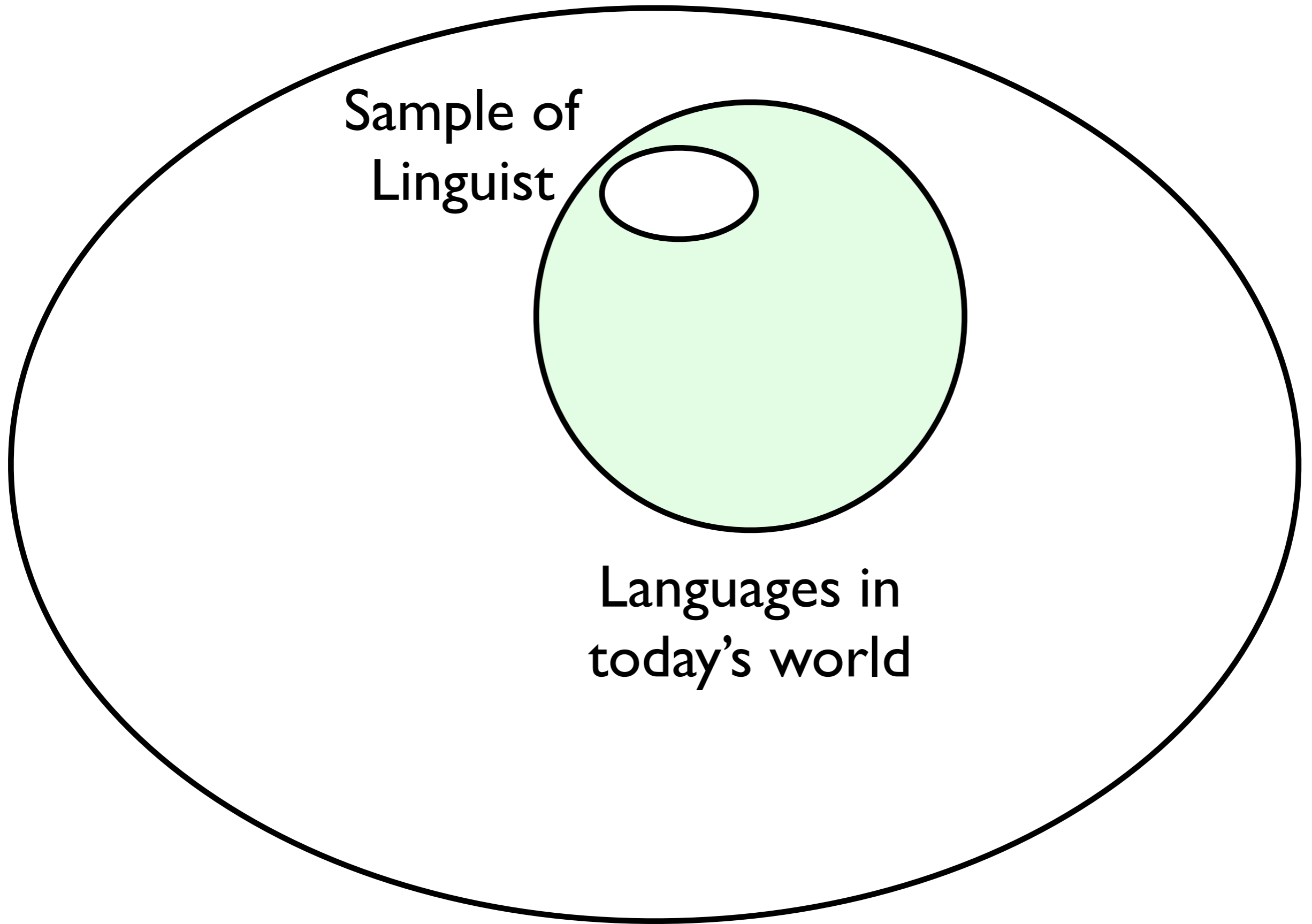
# Size of Phoneme Inventory

according to Justeson & Stephens (1984)

- phonemes are based on features
- the number of phonemic features are normally distributed across languages (!?)
- with  $n$  features one can make  $2^n$  combinations
- thus, the sizes of phoneme inventories are log-normally distributed

# Variation in typological studies

- Which languages are investigated ?
- How many languages are investigated ?
- On what basis are the types distinguished ?
- How many different types are distinguished ?



Sample of  
Linguist

Languages in  
today's world

All human languages, past & present

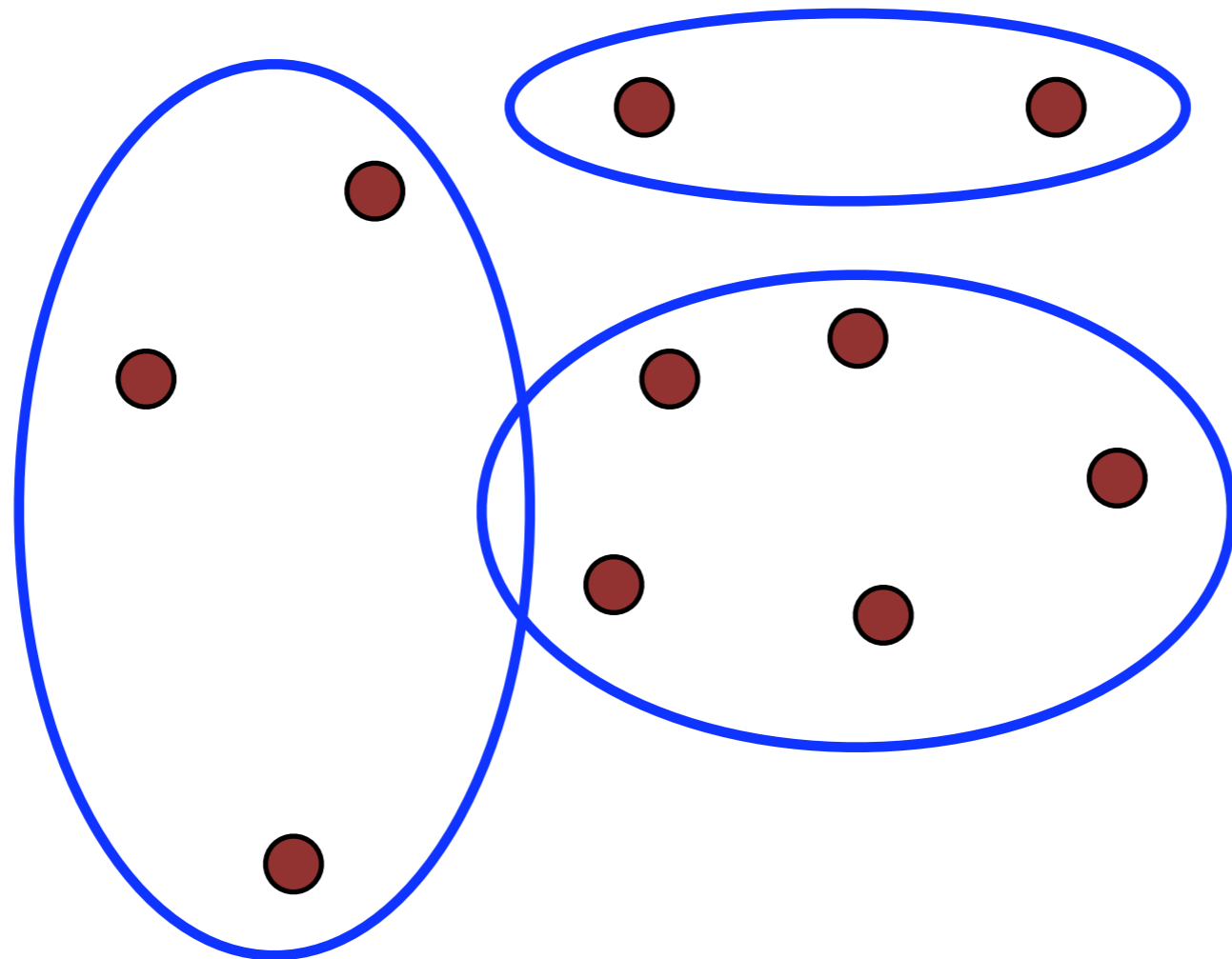
# Markov-chain queueing model

- Languages change
- Being of a certain type is like being in a queue of a shop being set up by a linguist
- Language enter the queue following a poisson distribution, and they exit the queue according to a poisson distribution
- The length of the queue is then negatively exponentially distributed

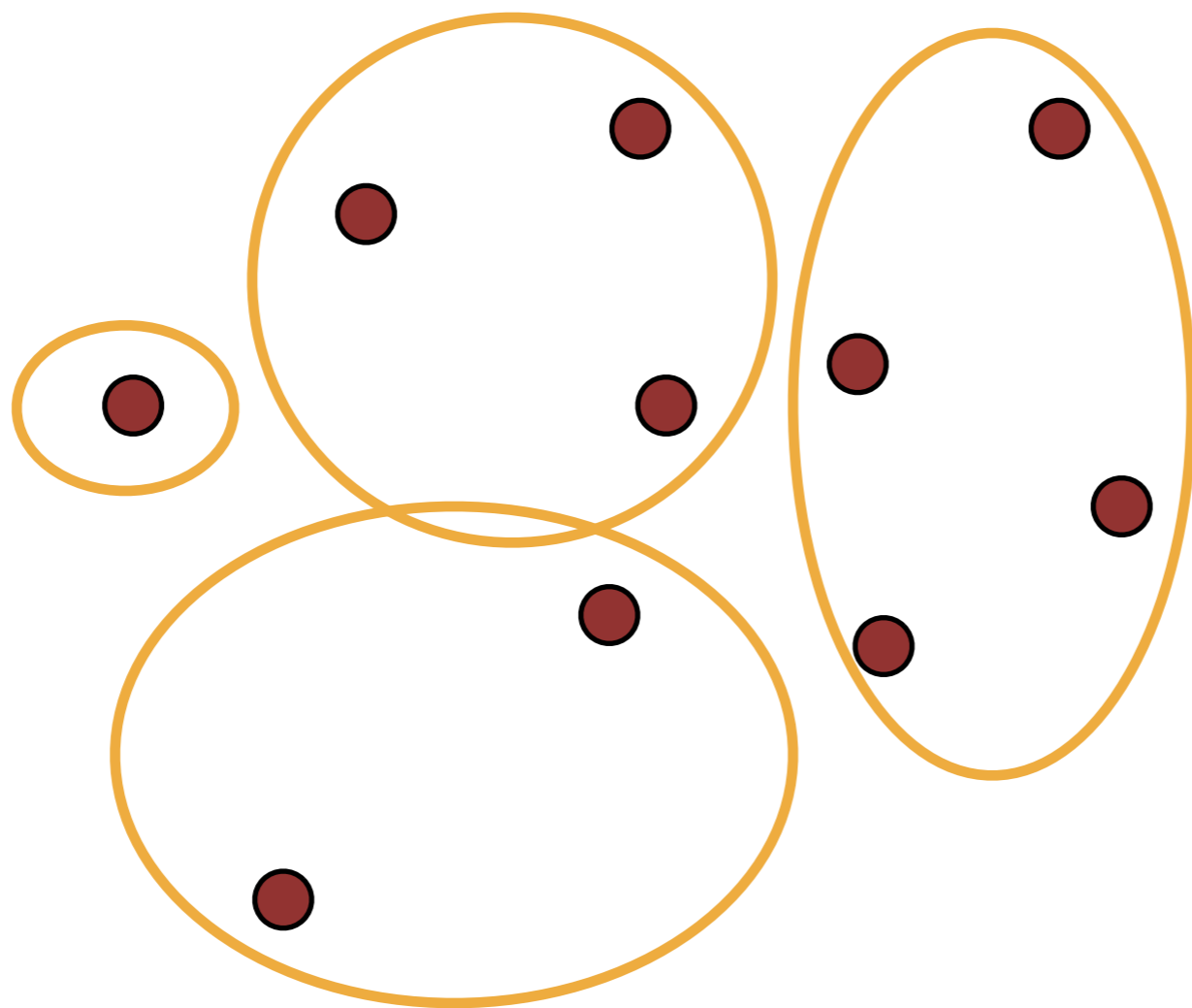
# B. Wälchli's data on motion events

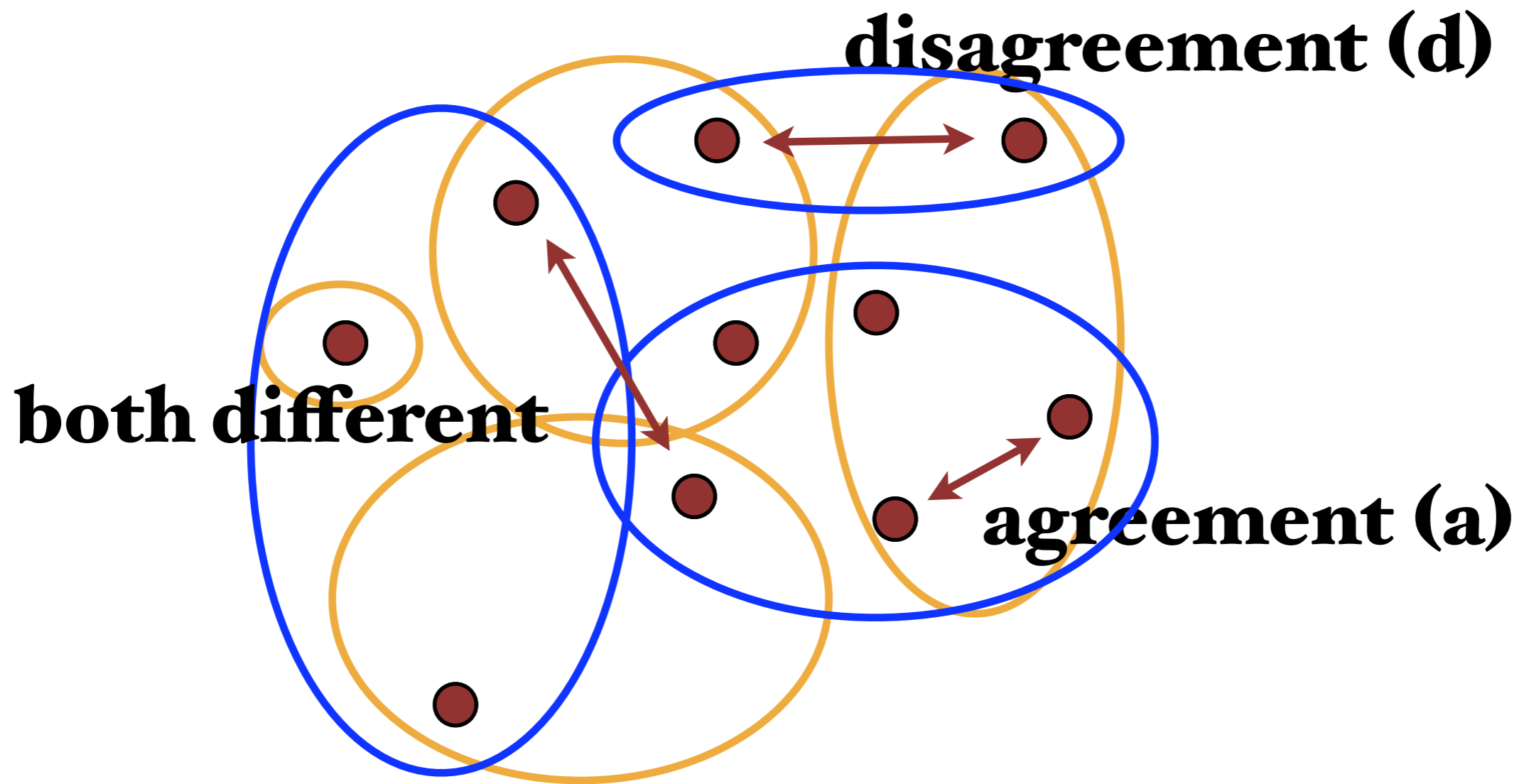
- 72 languages
- 335 clauses for each language from Bible
- lexical verbs describing motion events

	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







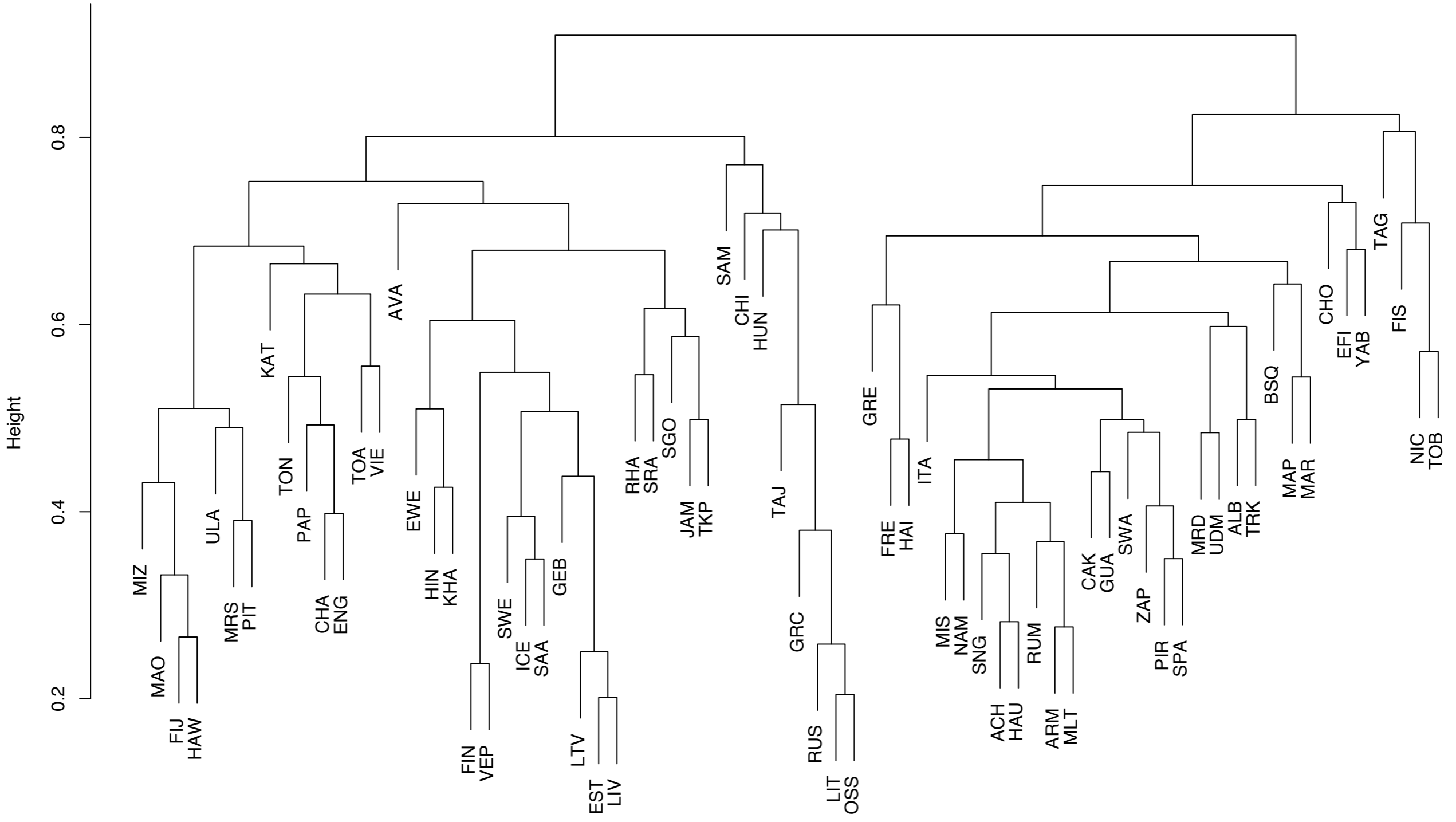


Jaccard similarity:

$$\frac{a}{a+d}$$

	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

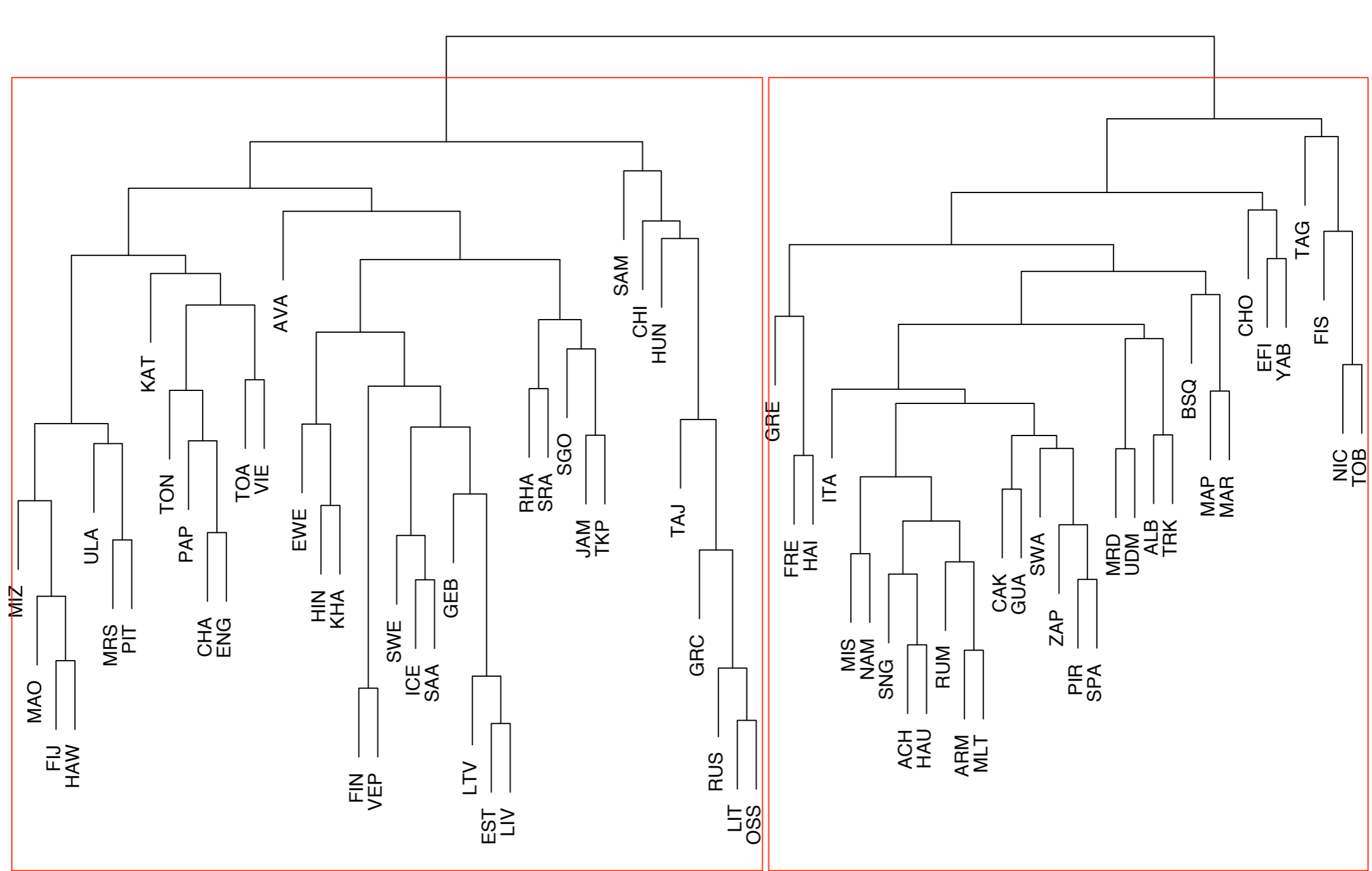




hclust (\*, "complete")

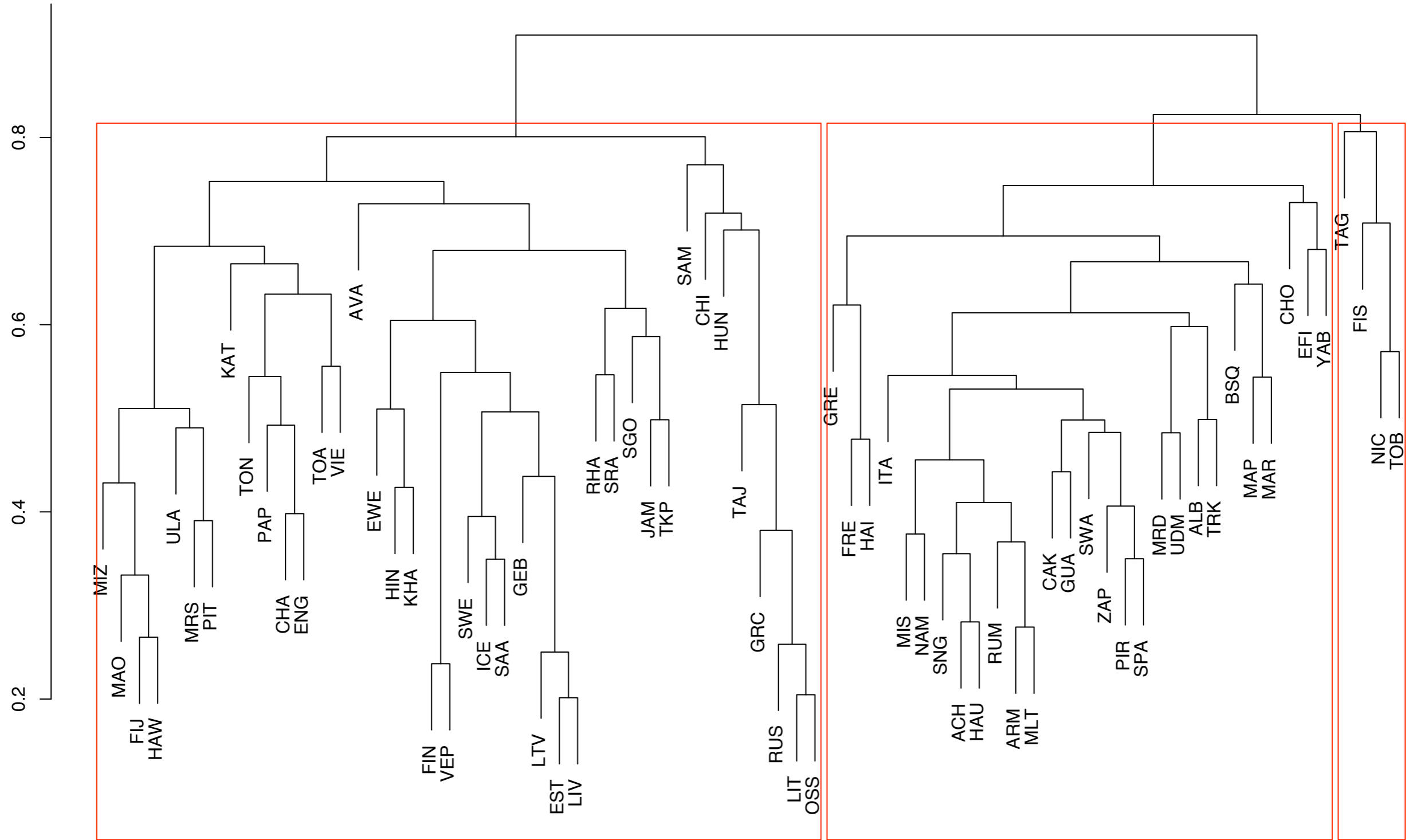
Height

0.2 0.4 0.6 0.8



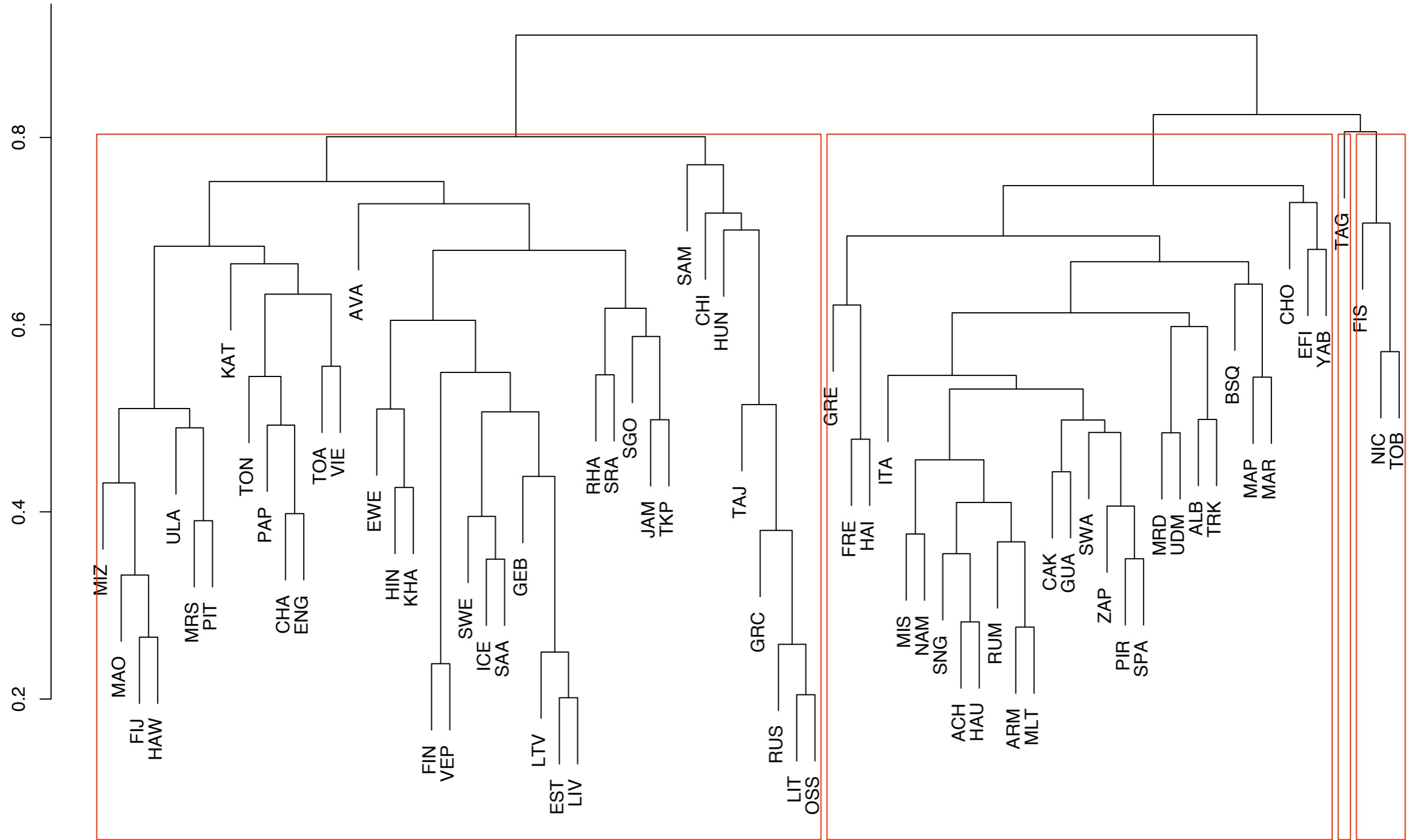
hclust (\*, "complete")

Height



hclust (\*, "complete")

Height

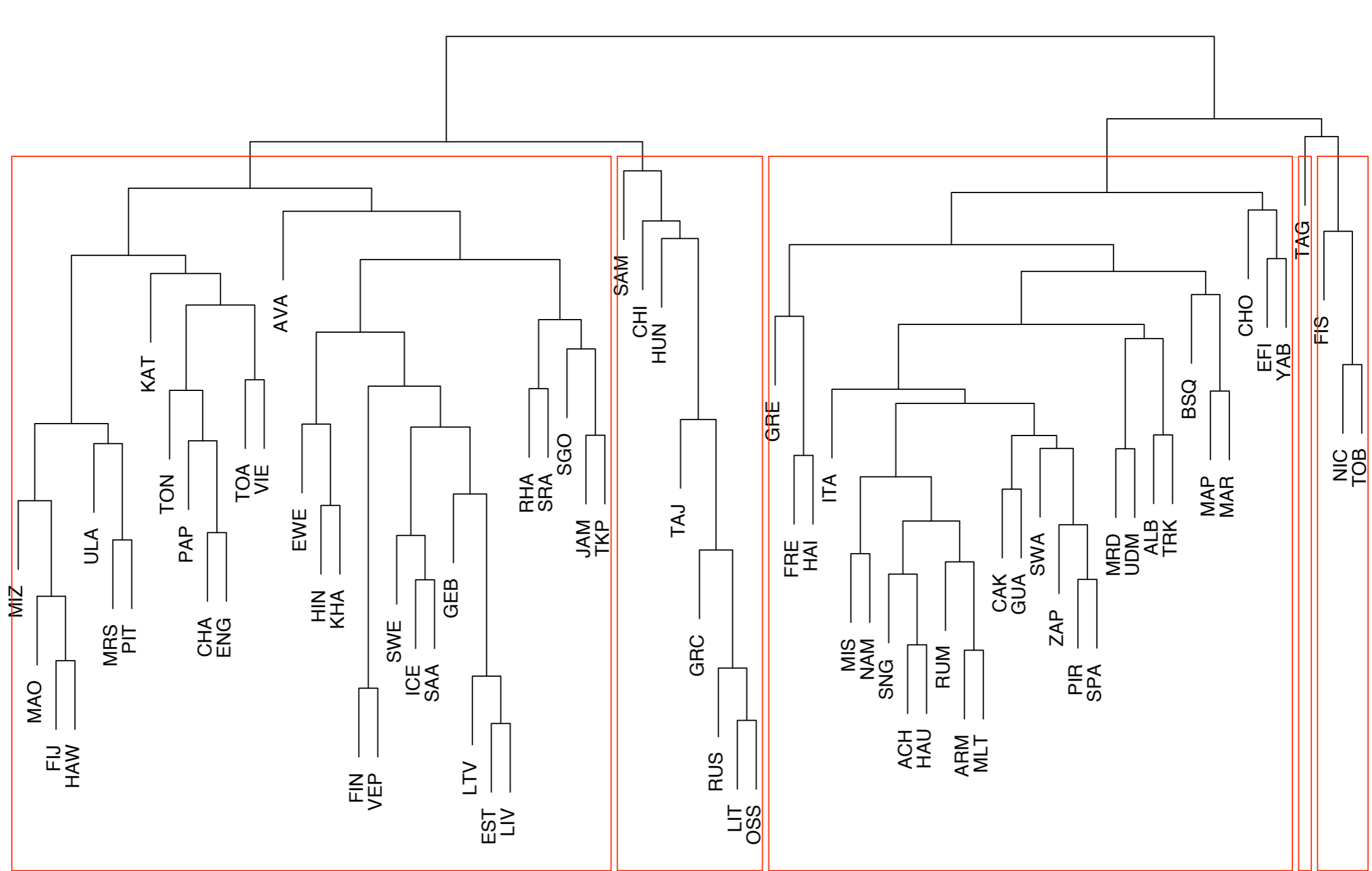


hclust (\*, "complete")



Height

0.2 0.4 0.6 0.8



hclust (\*, "complete")



























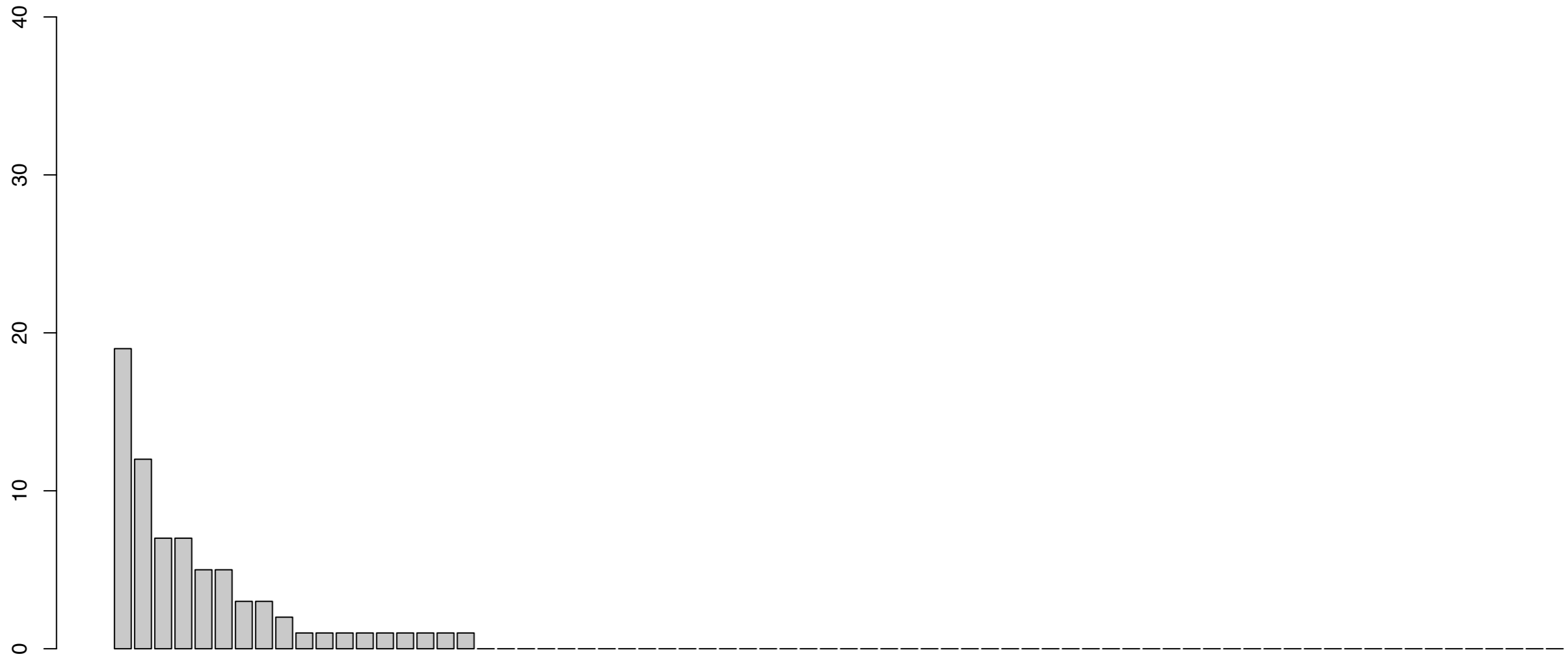




















































# Where is this going?

- Typological distributions are shaped by different forces:
  - ▶ language change
  - ▶ choices made by linguists
- Negative exponential distributions for unordered typologies are a promising idea
- The distributions are clearly not even !