Quantitative Typology

Michael Cysouw

cysouw@eva.mpg.de
Max Planck Institute for Evolutionary Anthropology

What's possible?

What's where why?

Bickel, Balthasar. 2007. Typology in the 21st century: Major current developments. Linguistic Typology 11.239-251.

Course Outline

- Monday: Typological Tradition
- Tuesday: Comparative Measurements
- Wednesday: Universals, Hierarchies, Maps
- Thursday: Explanations and Models
- Friday: Typology and Corpus Linguistics

Course Outline

- **Monday: Typological Tradition**
- Tuesday: Comparative Measurements
- Wednesday: Universals, Hierarchies, Maps
- Thursday: Explanations and Models
- Friday: Typology and Corpus Linguistics

Traditional Typology

A. Choose languages

B. Classify these languages into types

C. Interpret the frequency of types

Traditional Typology

A. Choose languages

B. Classify these languages into types

C. Interpret the frequency of types

Choice of Languages (Sampling)

- Tradition: sample genealogically (proportionally from linguistic families)
- Indeed: don't take 20 Indo-European languages and 5 other (*pace* Greenberg...)
- Watch out for large areal consistencies !
- Watch out for internal variation in families !

Future of Sampling

- Instead of 100 languages from 100 families take e.g. 20 families with 5 languages each
  - compare family-internal variation to between-family variation
- even better: sample along genealogical trees!
  - investigate coevolution of characteristics

Traditional Typology

A. Choose languages

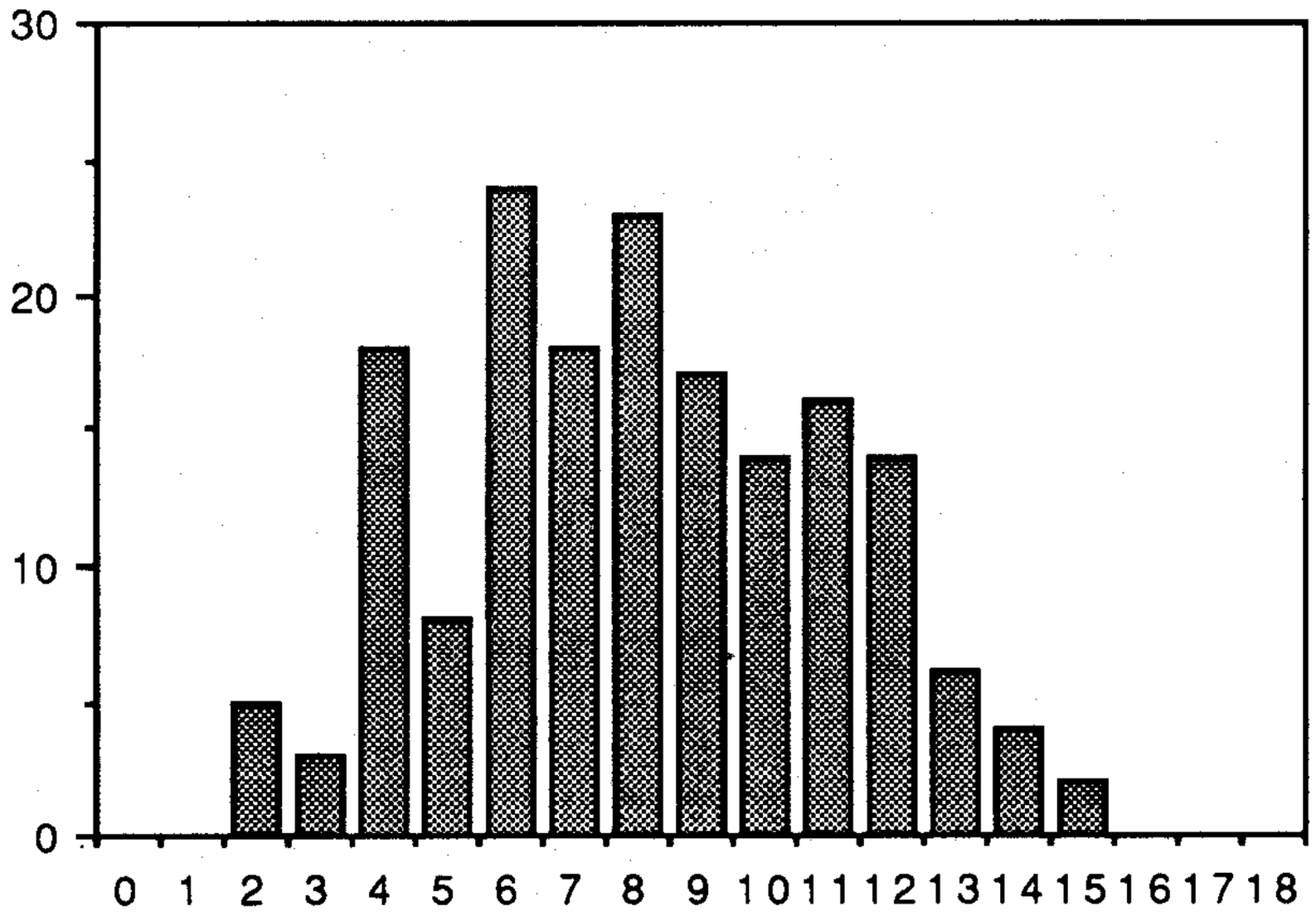
B. Classify these languages into types

C. Interpret the frequency of types

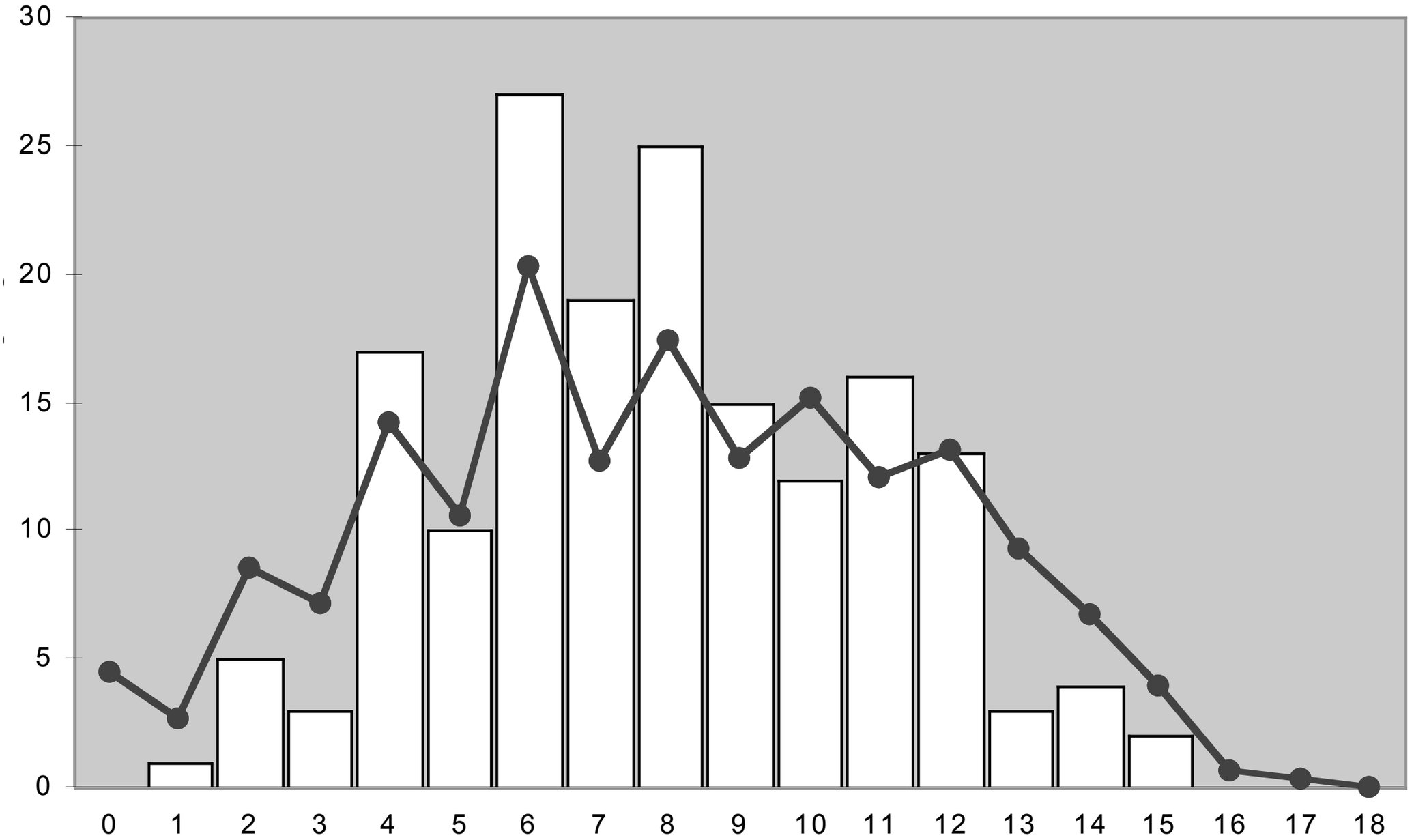
How to classify ?

- Not much methodology around:
  - ‘anything goes’ (Feyerabend)
  - as long as it brings results
- Watch out with summing up parameters !

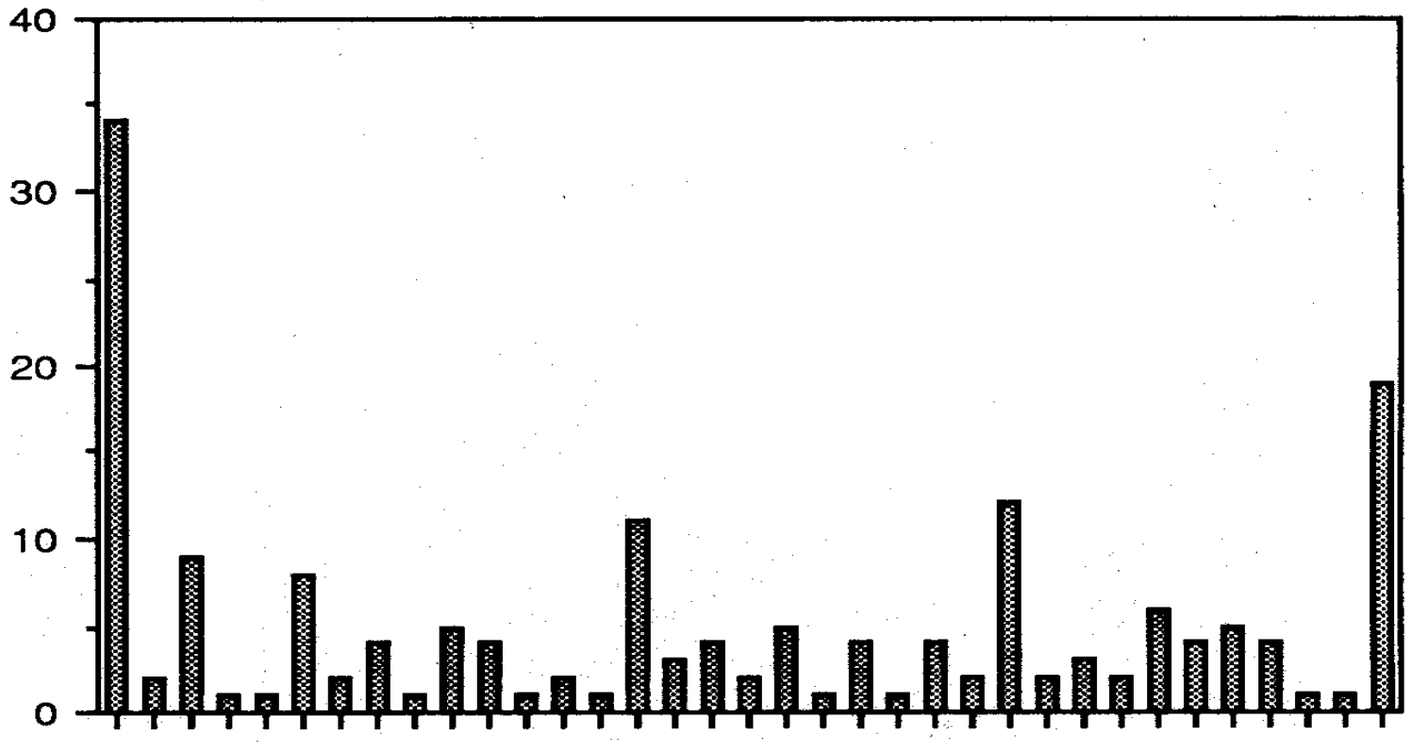
Sum of Head and Dependent marking: 'complexity':



'… the complexity (Dependent points plus Head points …) has a roughly normal distribution. Neither zero complexity nor the theoretical maximum complexity of [18] points (9 Head points plus 9 Dependent points …) occurs. the highest attested complexity is 15, found in only two languages. Figure 4 shows the complexity values attested in my sample. … The normal distribution and preference for moderate complexity shown in the overall sample are echoed in most … areas, with high complexity predominating in only two.' (Nichols 1992: 88-89)


30
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

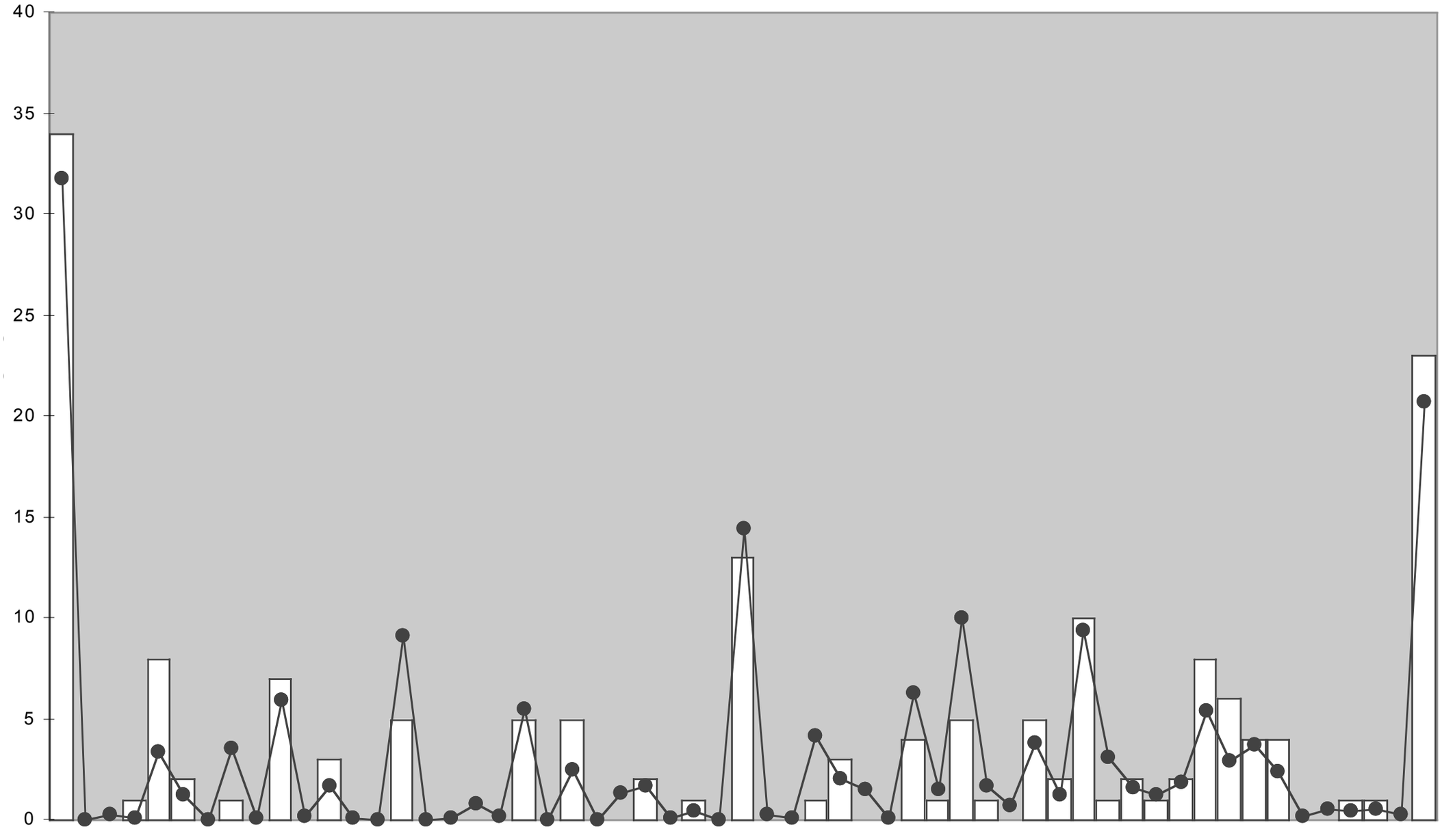
Ratio of Dependent and Head points: indicating the relative strength of head or dependent marking in a language.



'… computing the ration of dependent to head marking … gives us 35 different ratios among the 174 sample languages. Their distribution is shown in figure 1. It is bimodal, with the greatest peaks at the extremes of exclusive head marking (ration of zero since D = 0) and exclusive dependent marking (since H = 0, an actual ratio cannot be computed as it has a zero denominator). The other ratios, whose without zeroes, run from 0.14 (two languages) to 8.00 (one language). The highest frequencies are:

0.00	34 languages (radically head marking)	
0.17	9 languages	
0.50	8 languages	[should be '0.33', MC]
1.00	11 languages	
2.00	12 languages	
H = 0	19 languages (radically dependent marking)	

… The other three frequency peaks suggest that preferred patterns cluster at perceptually simple ratios: two to one, one to one, and one to two. Overall, then , we have a preferecne for neatness of some sort: polar types, two-to-one ratios and even splits.' (Nichols 1992: 72-73)


40
35
30
25
20
15
10
5
0

Traditional Typology

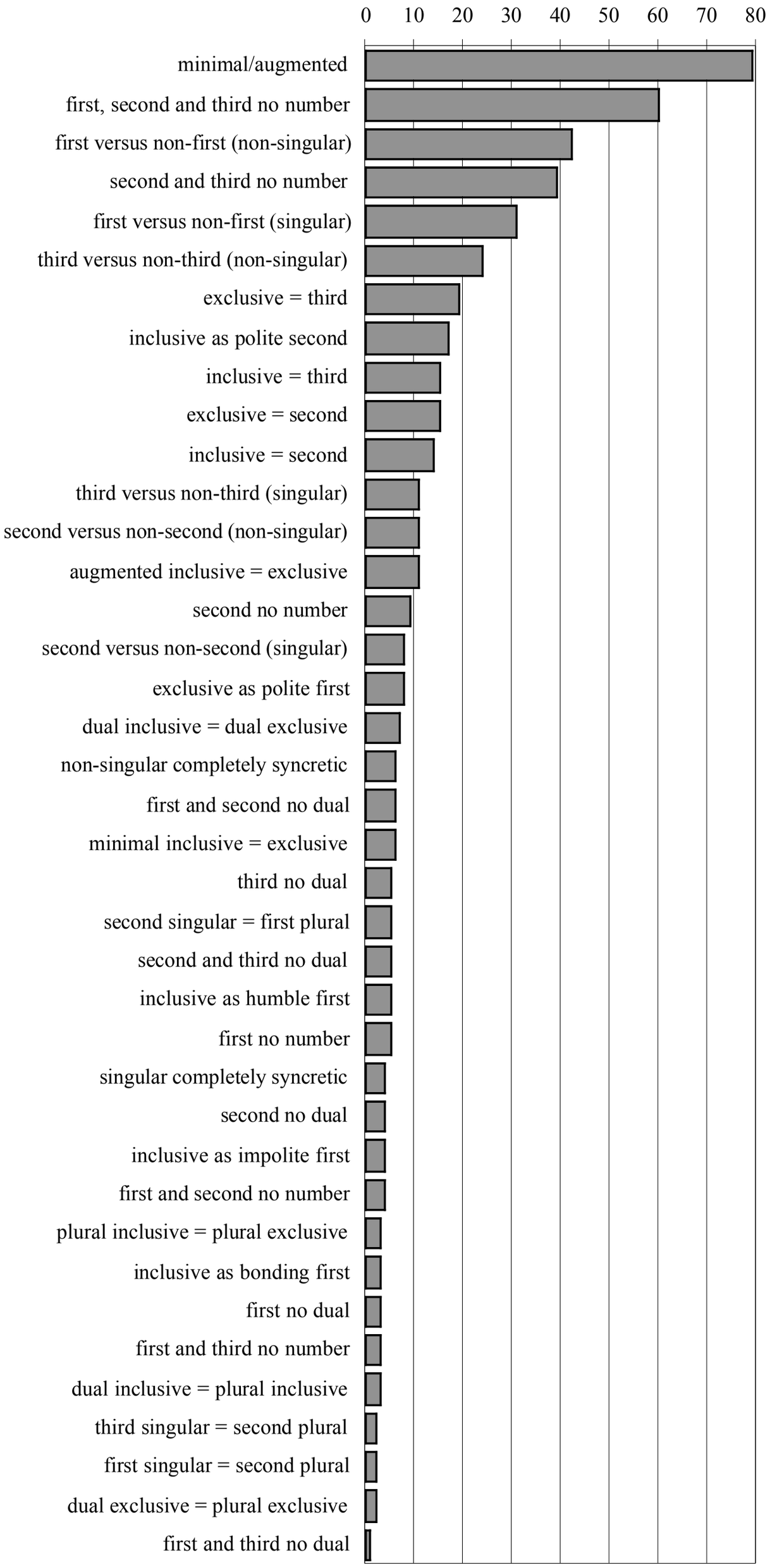
A. Choose languages

B. Classify these languages into types

C. Interpret the frequency of types

Distributions

- What would we expect?
  - "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." (Comrie, 1989, 20.)
- Probably not true: we should expect many small types and just few large ones


0
10
20
30
40
50
60
70
80
minimal/augmented
first, second and third no number
first versus non-first (non-singular)
second and third no number
first versus non-first (singular)
third versus non-third (non-singular)
exclusive = third
inclusive as polite second
inclusive = third
exclusive = second
inclusive = second
third versus non-third (singular)
second versus non-second (non-singular)
augmented inclusive = exclusive
second no number
second versus non-second (singular)
exclusive as polite first
dual inclusive = dual exclusive
non-singular completely syncretic
first and second no dual
minimal inclusive = exclusive
third no dual
second singular = first plural
second and third no dual
inclusive as humble first
first no number
singular completely syncretic
second no dual
inclusive as impolite first
first and second no number
plural inclusive = plural exclusive
inclusive as bonding first
first no dual
first and third no number
dual inclusive = plural inclusive
third singular = second plural
first singular = second plural
dual exclusive = plural exclusive
first and third no dual

Implicational Universals

- The typological tradition
- Statistical view of things

The typological tradition

- Implicational Universal
- Bidirectional Universal (Equivalence)
- Implicational Hierarchy
- Nested Implicational Universal

Greenberg (1963)

- *Universal 3*: Languages with dominant VSO order are always prepositional
- *Universal 2:* In languages with prepositions, the genitive almost always follows the governing noun, while in languages with postpositions it almost always precedes

Statistical view of things

	+	–	total
+	**10**	**31**	41
–	**2**	**12**	14
total	12	43	55

	+	–	total
+	$\frac{41}{55} \cdot \frac{12}{55} \cdot 55 = 8.9$	$\frac{41}{55} \cdot \frac{43}{55} \cdot 55 = 32.1$	41
–	$\frac{41}{55} \cdot \frac{43}{55} \cdot 55 = 32.1$	$\frac{14}{55} \cdot \frac{43}{55} \cdot 55 = 10.9$	14
total	12	43	55

	+	–	total
+	**+ 1.1**	**-1.1**	41
–	**-1.1**	**+1.1**	14
total	12	43	55

What do typologists say?

Smallest number	Kind of universal	Hypothetical distributions of a 100-language sample							
Zero	Exceptionless universal	33	34	26	48	20	60	14	72
		0	33	**0**	26	**0**	20	**0**	14
Five	Strong tendency	36	23	31	33	27	41	22	51
		5	36	**5**	31	**5**	27	**5**	22
Ten	Statistical tendency	38	14	33	24	30	30	25	40
		10	38	**10**	33	**10**	30	**10**	25
Fifteen	Maybe something			35	15	31	23	28	29
				15	35	**15**	31	**15**	28
Nineteen	Nothing					31	19	27	27
						19	31	**19**	27

What do statisticians say?

Hypothetical distributions of a 100-language sample

33	34	26	48	20	60	14	72
0	33	**0**	26	**0**	20	**0**	14
36	23	31	33	27	41	22	51
5	36	**5**	31	**5**	27	**5**	22
38	14	33	24	30	30	25	40
10	38	**10**	33	**10**	30	**10**	25
		35	15	31	23	28	29
		15	35	**15**	31	**15**	28
				31	19	27	27
				19	31	**19**	27

Kind of interaction	Very strongly significant	Strongly significant	Significant	No interaction
Fisher's Exact two-tailed	$p < 0.000001$	$p < 0.001$	$p < 0.05$	$p > 0.2$

Course Outline

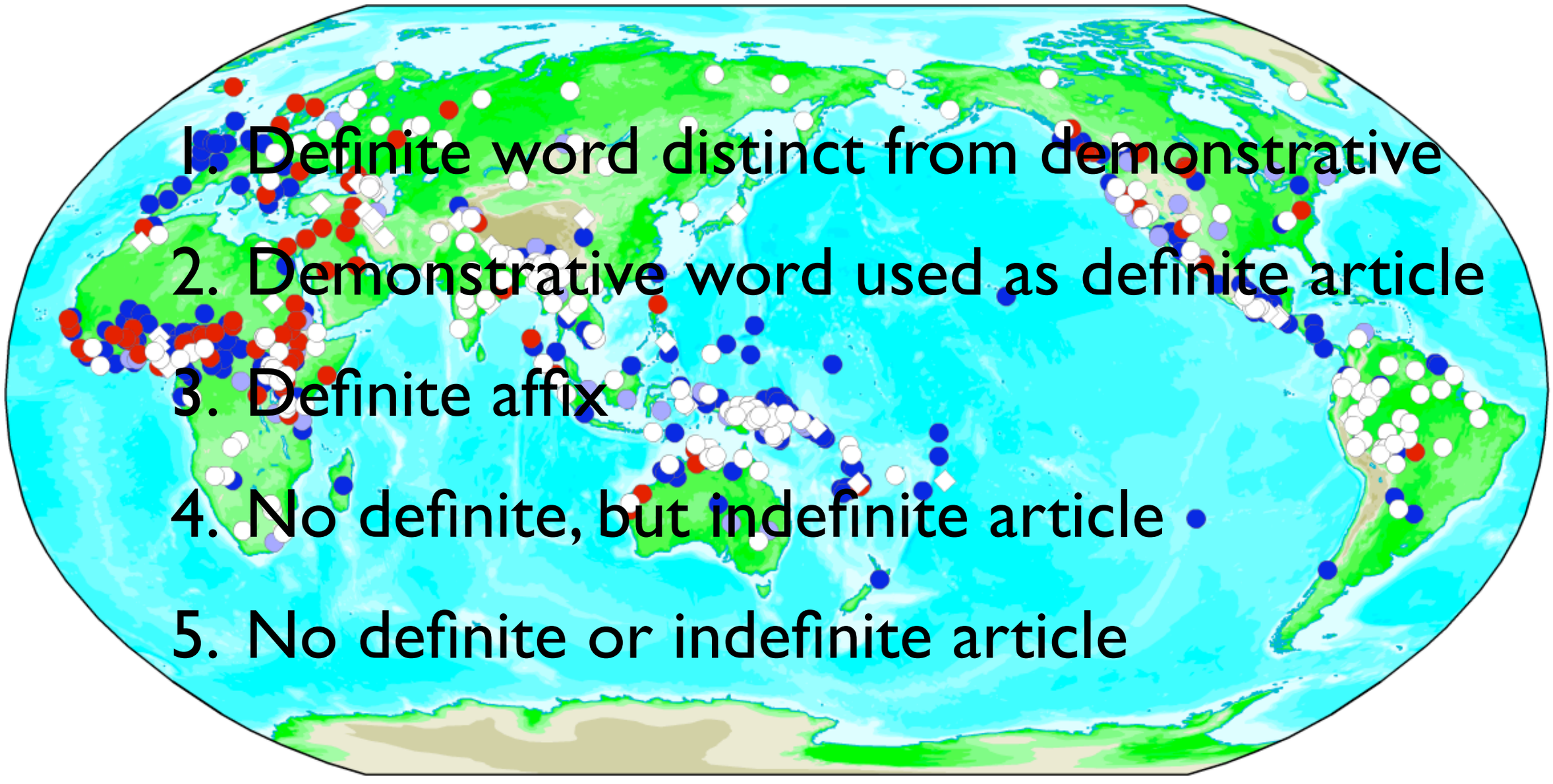
- Monday: Typological Tradition
- **Tuesday: Comparative Measurements**
- Wednesday: Universals, Hierarchies, Maps
- Thursday: Explanations and Models
- Friday: Typology and Corpus Linguistics

Measurement theory

- Stevens (1946)
  - from a psychological background
- proposed hierarchy of variables
  - nominal
  - ordinal
  - interval
  - ratio
- "yardstick" metaphor of measurement

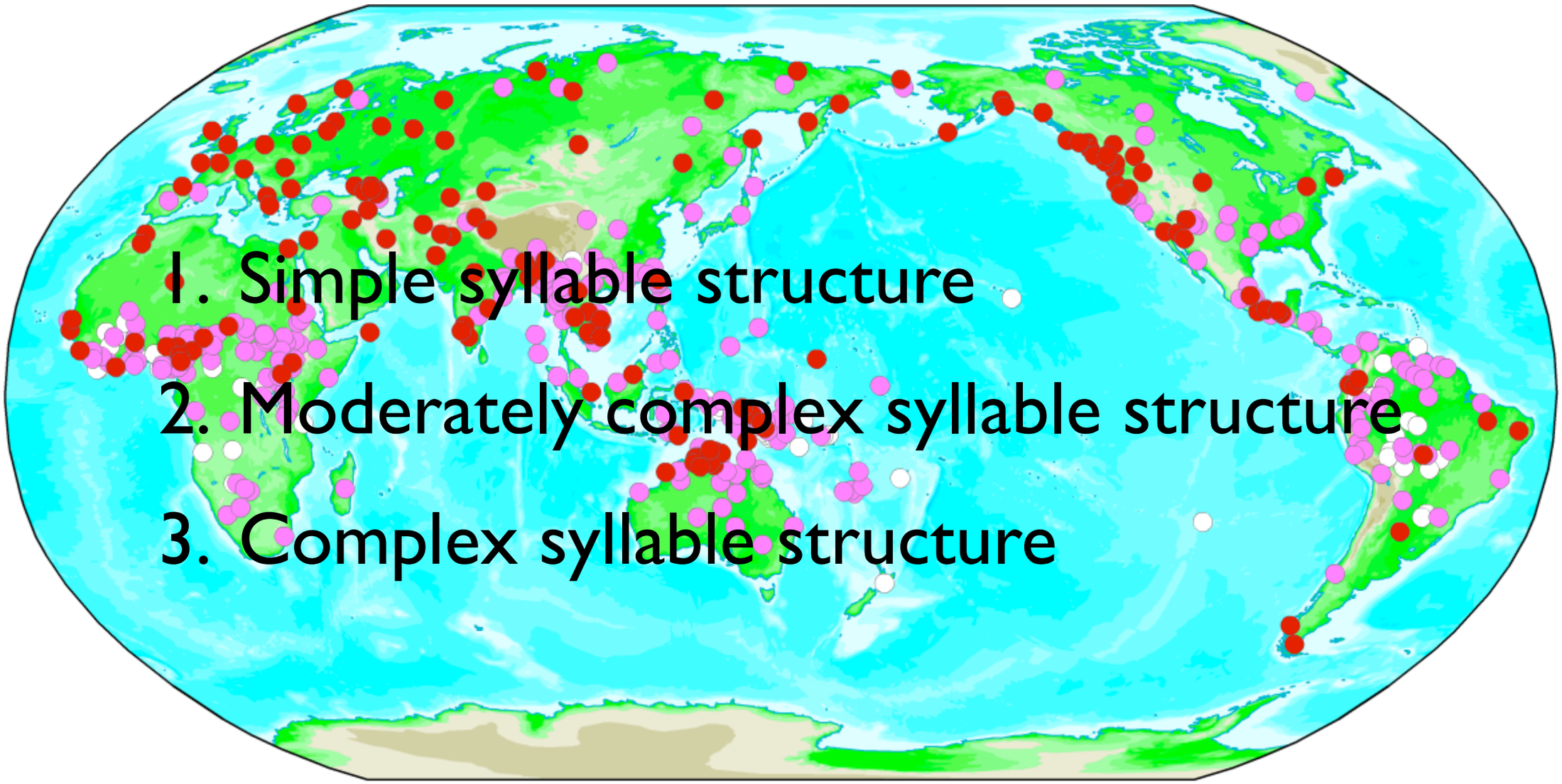
Stevens, S. S. (1946) 'On the theory of scales of measurement', *Science* 103 (2684): 677-680.

Categorization
(nominal variable)



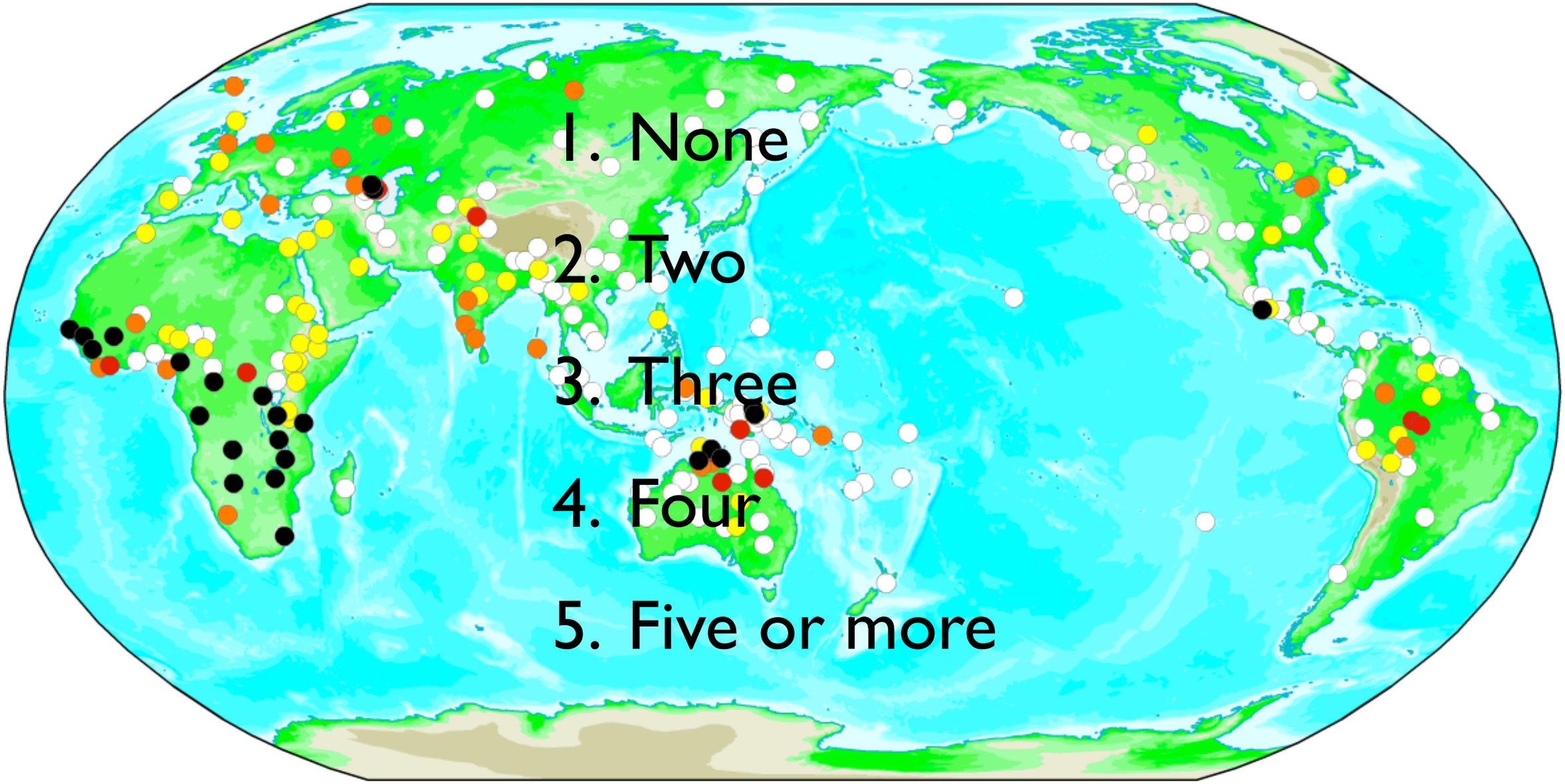
Dryer, Matthew S. (2005) 'Definite article' in: Martin Haspelmath, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *World Atlas of Language Structures*. Oxford: Oxford University Press, 154-157.

Linearly ordered categorization
(interval variable)



Maddieson, Ian (2005) 'Syllable structure' in: Martin Haspelmath, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *World Atlas of Language Structures*. Oxford: Oxford University Press, 54-57.

Count
(ratio variable)



Corbett, Greville G. (2005) 'Number of genders' in: Martin Haspelmath, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *World Atlas of Language Structures*. Oxford: Oxford University Press, 126-129.

Continuum
(ratio variable)

Language	Average wordlength
Hmong Nua	3.72
English	5.05
German	6.23
Cashinahua	6.42
Bugis	6.45
Inuktitut	14.99

n)

Measurement theory

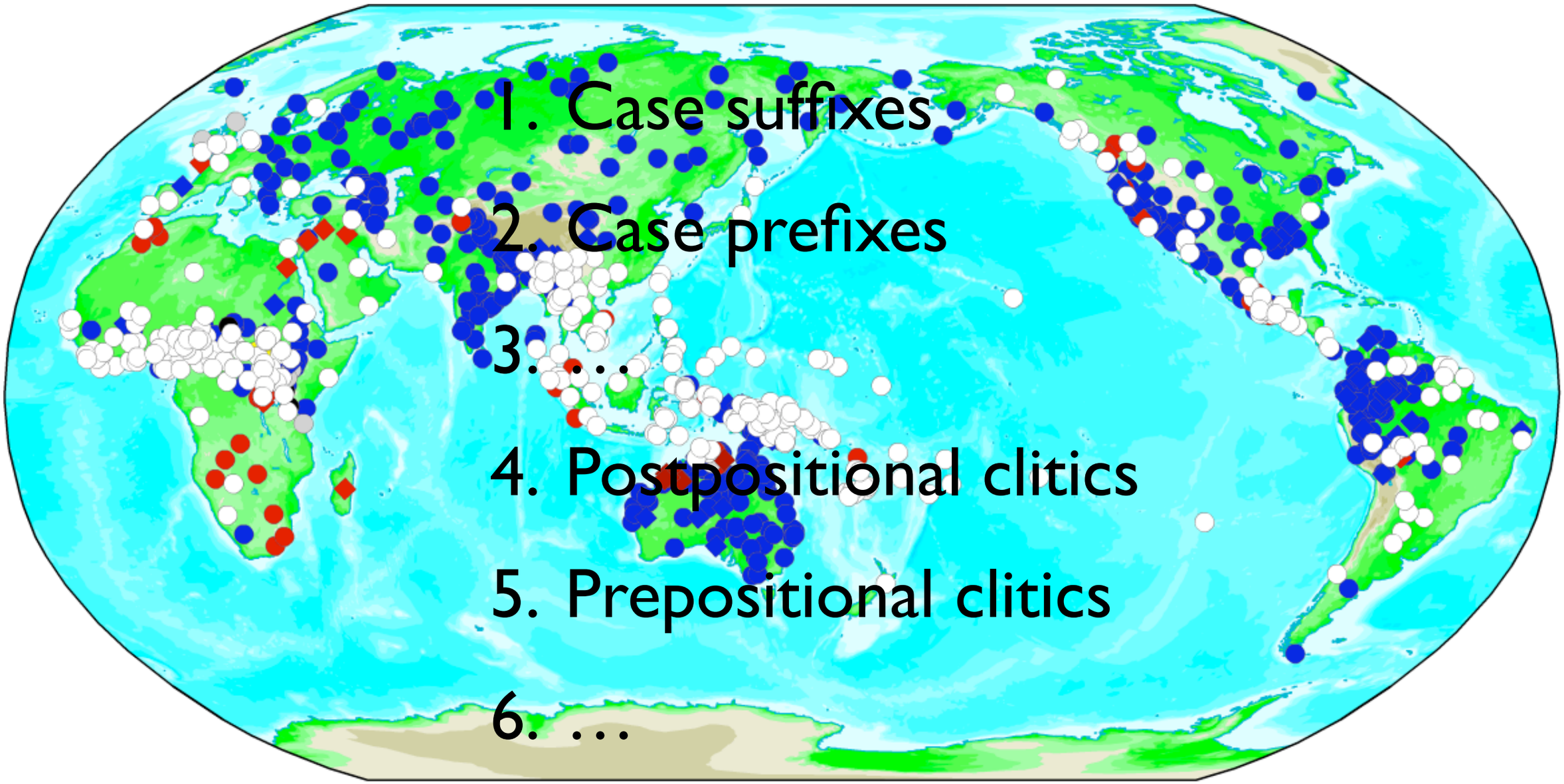
- Stevens (1946)
  - from a psychological background
- proposed hierarchy of variables
  - nominal
  - ordinal
  - interval
  - ratio
- "yardstick" metaphor of measurement

Stevens, S. S. (1946) 'On the theory of scales of measurement', *Science* 103 (2684): 677-680.

Problems

- More measurements wanted
  - more specification in categorization
  - full pairwise comparisons
- Difficult to combine measurements of different kinds

More specification for categorizations



Dryer, Matthew S. (2005) 'Position of case affixes' in: Martin Haspelmath, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *World Atlas of Language Structures*. Oxford: Oxford University Press, 210-213.

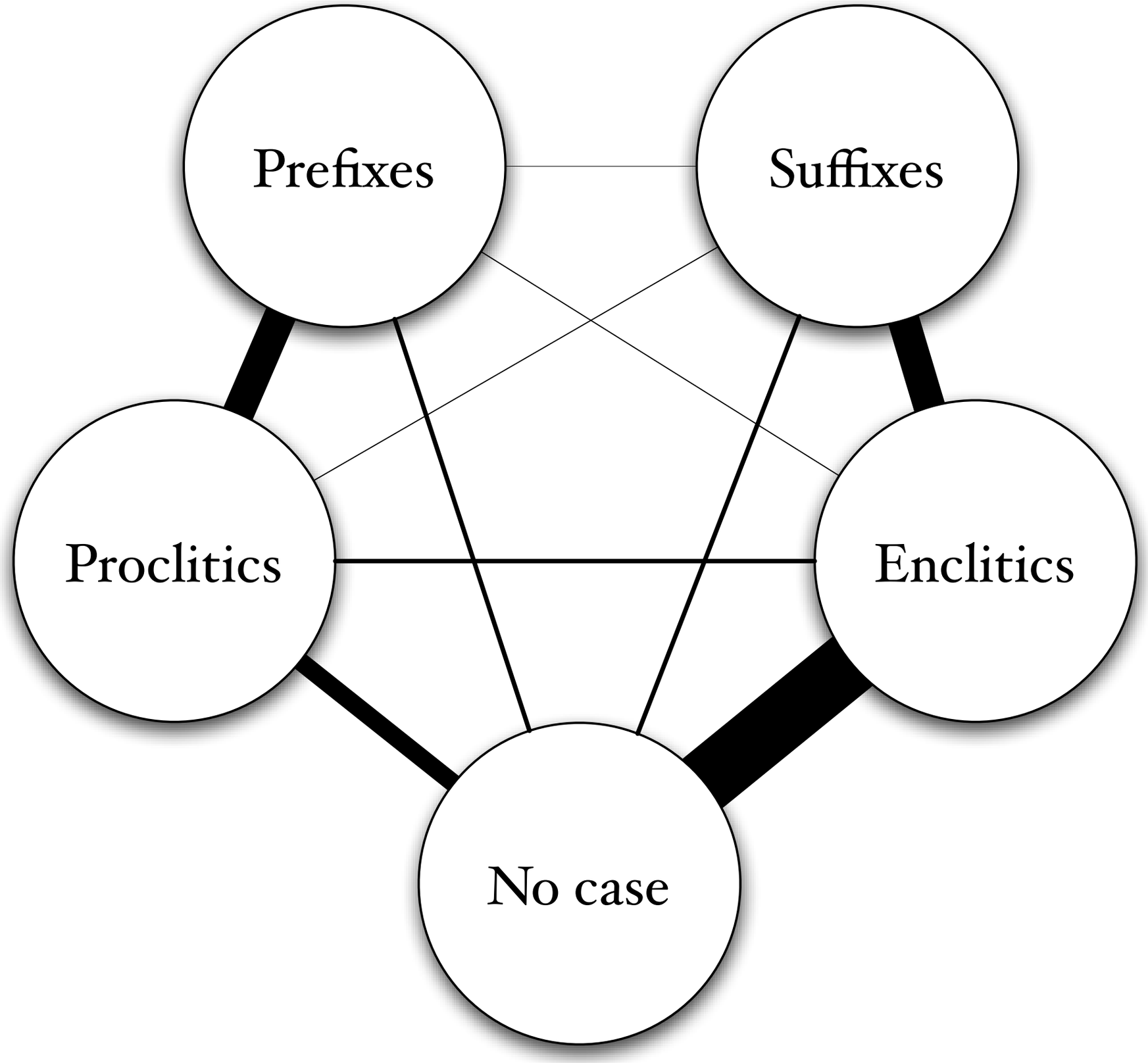
Prefixes

Suffixes

Proclitics

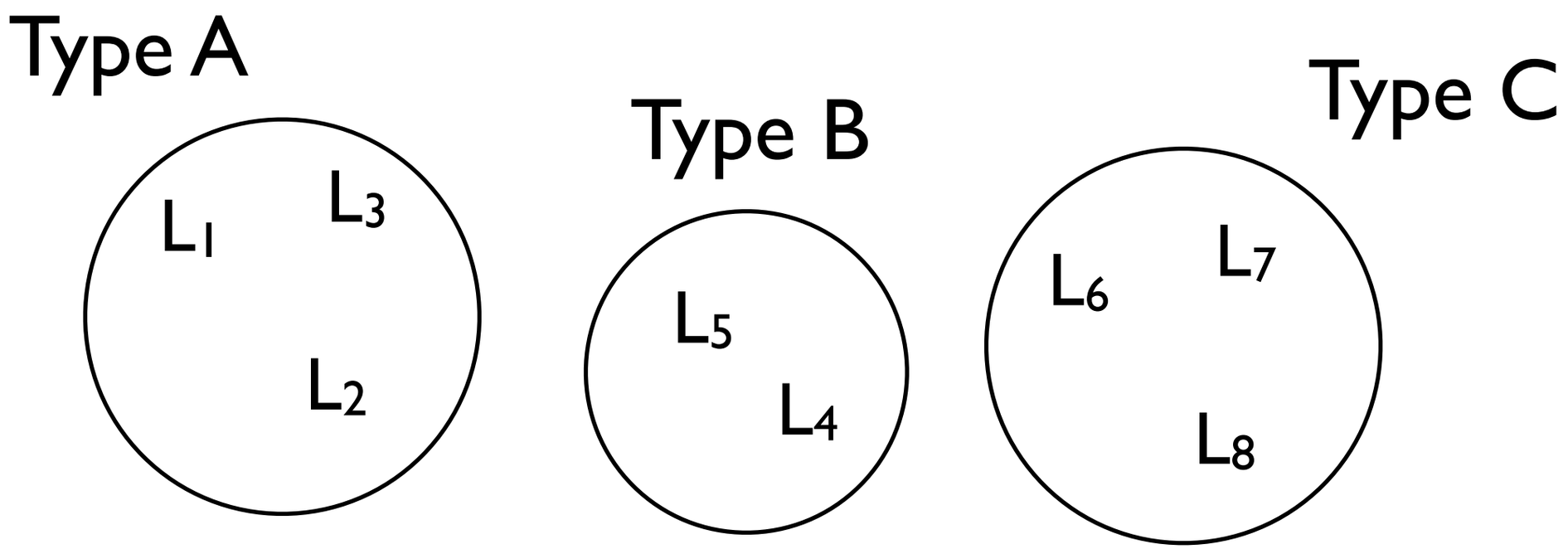
Enclitics

No case


Prefixes
Suffixes
Proclitics
Enclitics
No case

Relational metaphor of measurement

- Express typology as pairwise language-to-language similarities
- Such a typology consists of data with separate interpretation of the meaning of the data


Type A
L1
L3
L2
Type B
L5
L4
Type C
L6
L7
L8

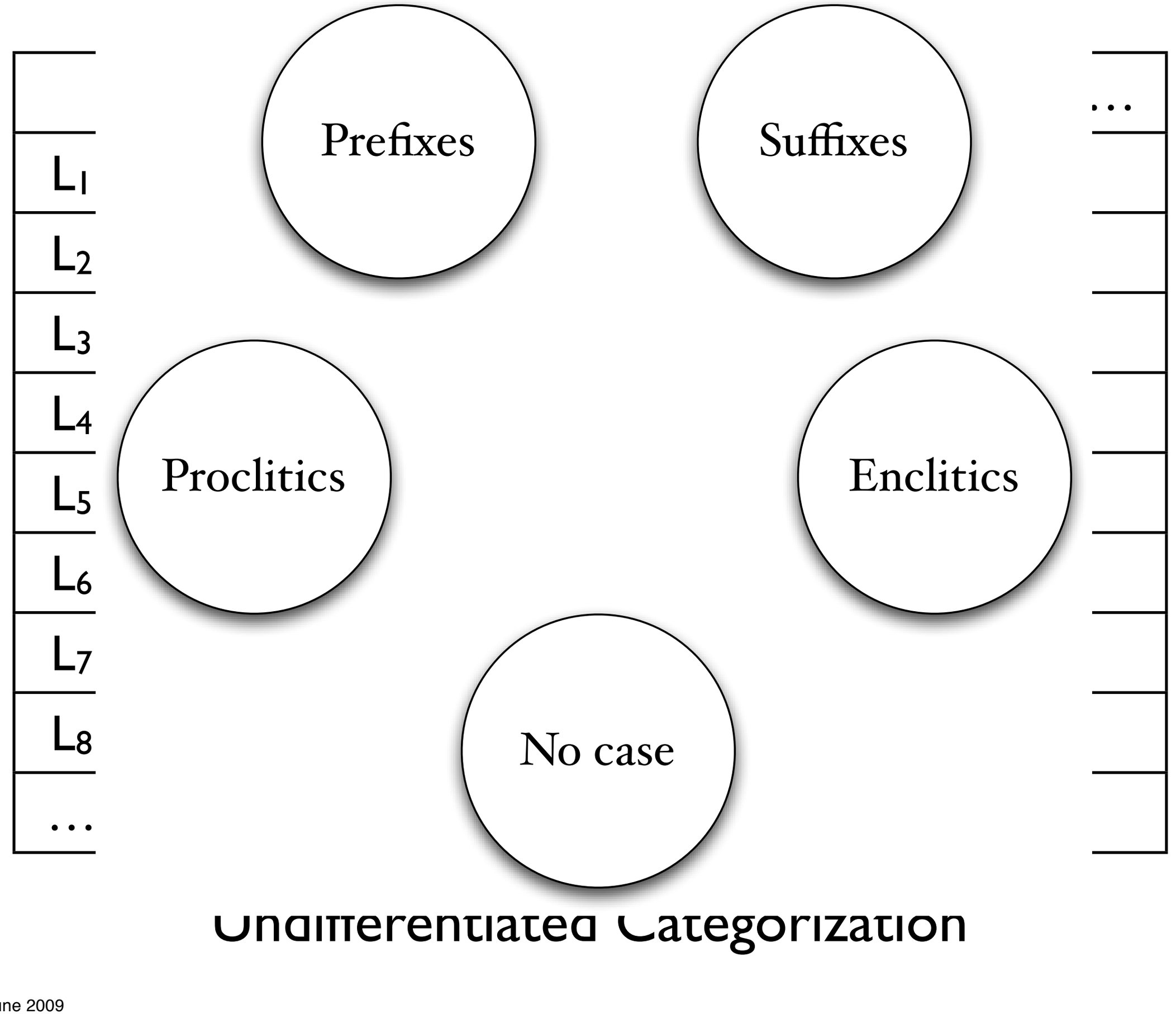
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1									
L_2									
L_3									
L_4									
L_5									
L_6									
L_7									
L_8									
…									

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	I								
L_2		I							
L_3			I						
L_4				I					
L_5					I				
L_6						I			
L_7							I		
L_8								I	
…									

	Type A			Type B		Type C			
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	I								
L_2		I							
L_3			I						
L_4				I					
L_5					I				
L_6						I			
L_7							I		
L_8								I	
…									

	Type A			Type B		Type C			
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	I	I	I						
L_2	I	I	I						
L_3	I	I	I						
L_4				I	I				
L_5				I	I				
L_6						I	I	I	
L_7						I	I	I	
L_8						I	I	I	
…									

Type A Type B Type C



Undifferentiated Categorization

Type A Type B Type C

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	1	1	1	?	?	0	0	0	
L_2	1	1	1	?	?	0	0	0	
L_3	1	1	1	?	?	0	0	0	
L_4	0	0	0	1	1	0	0	0	
L_5	0	0	0	1	1	0	0	0	
L_6	0	0	0	0	0	1	1	1	
L_7	0	0	0	0	0	1	1	1	
L_8	0	0	0	0	0	1	1	1	
…									

Type A Type B Type C

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	1	1	1	0.37	0.37	0.28	0.28	0.28	
L_2	1	1	1	0.37	0.37	0.28	0.28	0.28	
L_3	1	1	1	0.37	0.37	0.28	0.28	0.28	
L_4	0.37	0.37	0.37	1	1	0.51	0.51	0.51	
L_5	0.37	0.37	0.37	1	1	0.51	0.51	0.51	
L_6	0.28	0.28	0.28	0.51	0.51	1	1		
L_7	0.28	0.28	0.28	0.51	0.51	1	1	1	
L_8	0.28	0.28	0.28	0.51	0.51	1	1	1	
…									

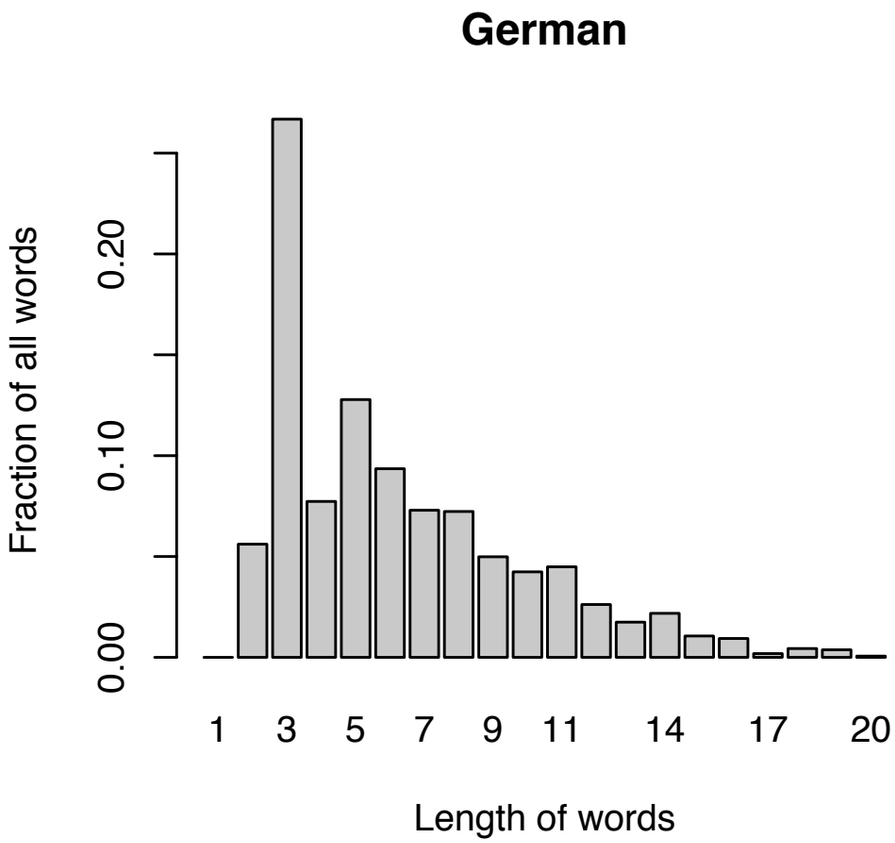
Differentiated Categorization

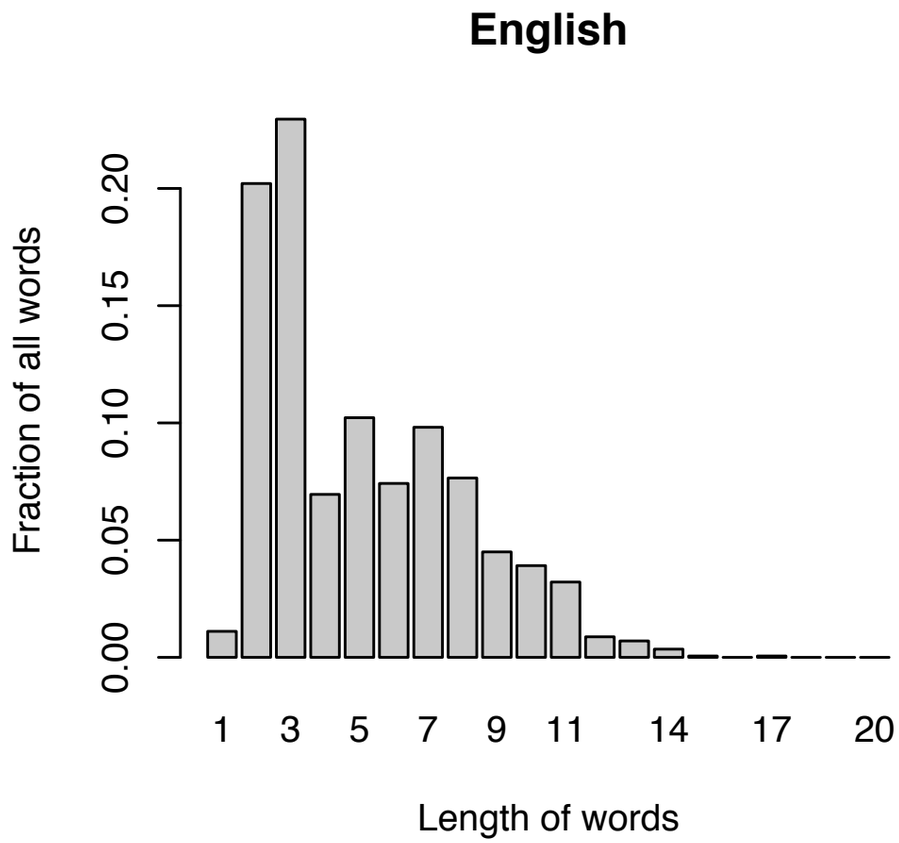
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	1	0.55	0.72	0.31	0.70	0.61	0.50	0.58	
L_2	0.55	1	0.55	0.31	0.40	0.44	0.31	0.48	
L_3	0.72	0.55	1	0.29	0.53	0.51	0.48	0.60	
L_4	0.31	0.31	0.29	1	0.38	0.36	0.26	0.27	
L_5	0.70	0.40	0.53	0.38	1	0.64	0.51	0.46	
L_6	0.61	0.44	0.51	0.36	0.64	1	0.57	0.43	
L_7	0.50	0.31	0.48	0.26	0.51	0.57	1	0.47	
L_8	0.58	0.48	0.60	0.27	0.46	0.43	0.47	1	
…									

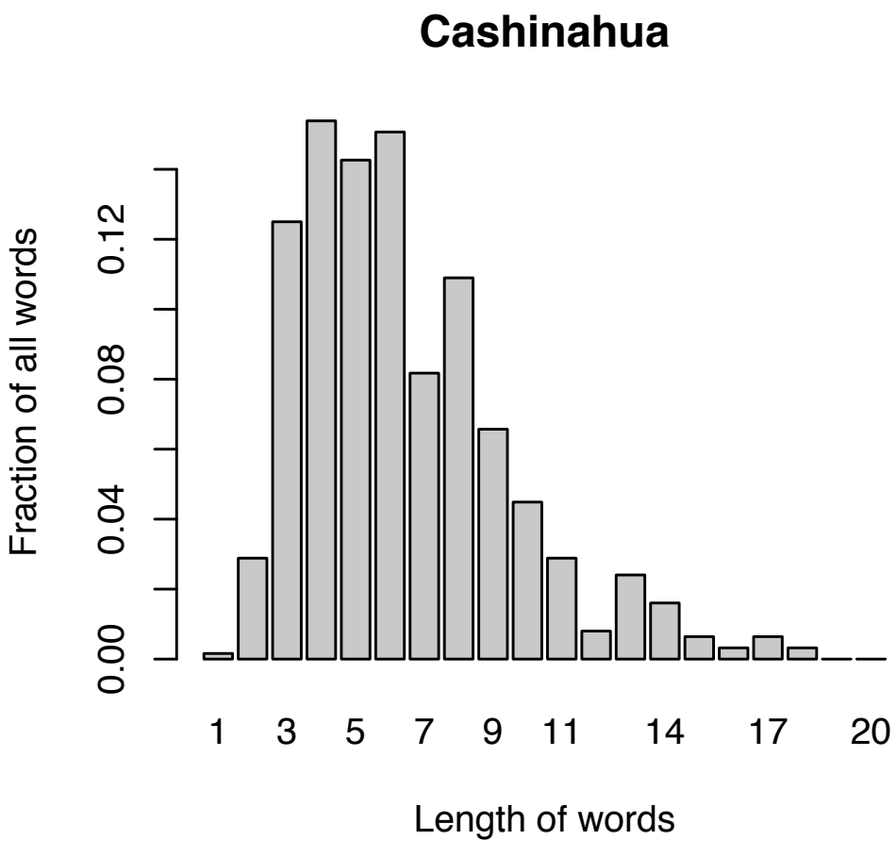
‘Deconstructed’ Typology

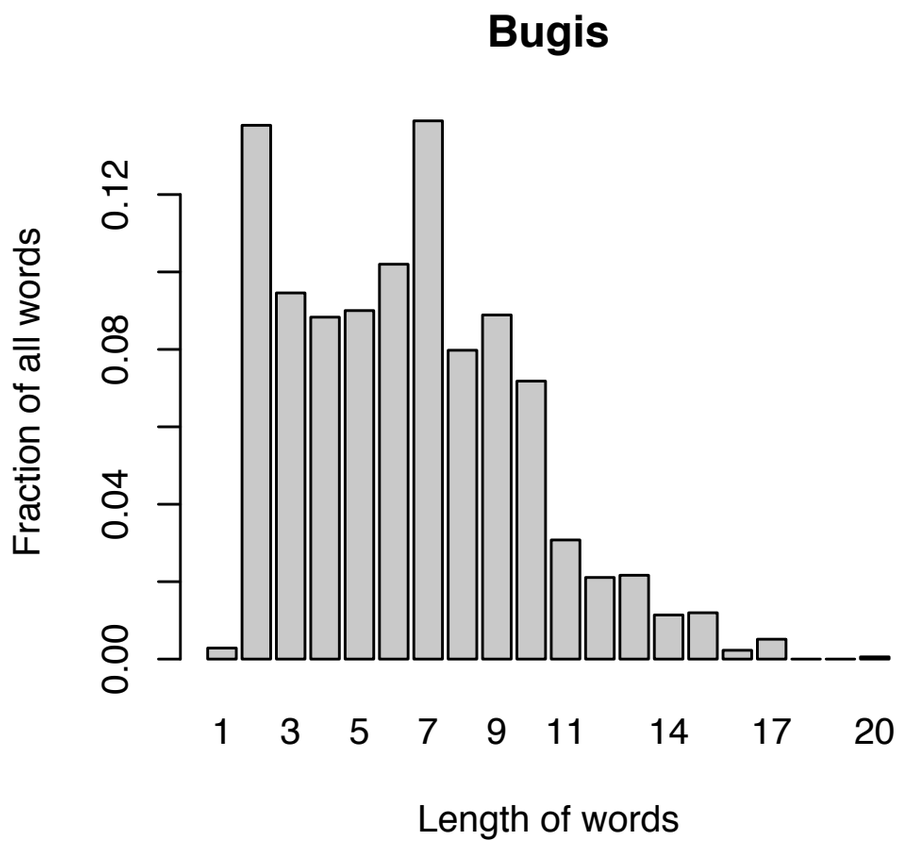
Pairwise Comparison

Language	Average wordlength
Hmong Nua	3.72
English	5.05
German	6.23
Cashinahua	6.42
Bugis	6.45
Inuktitut	14.99


German
Fraction of all words
0.00
0.10
0.20
1 3 5 7 9 11 14 17 20
Length of words

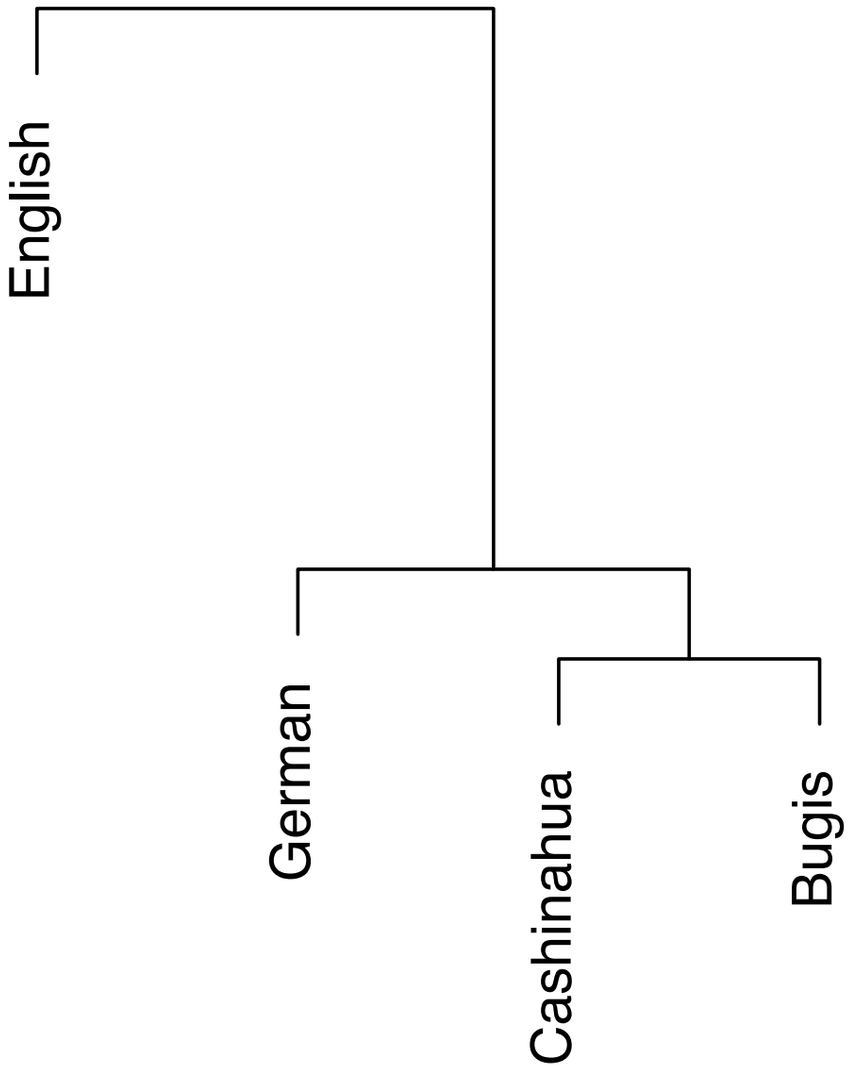

English
Fraction of all words
0.00
0.05
0.10
0.15
0.20
1 3 5 7 9 11 14 17 20
Length of words


Cashinahua
Fraction of all words
0.00
0.04
0.08
0.12
1 3 5 7 9 11 14 17 20
Length of words

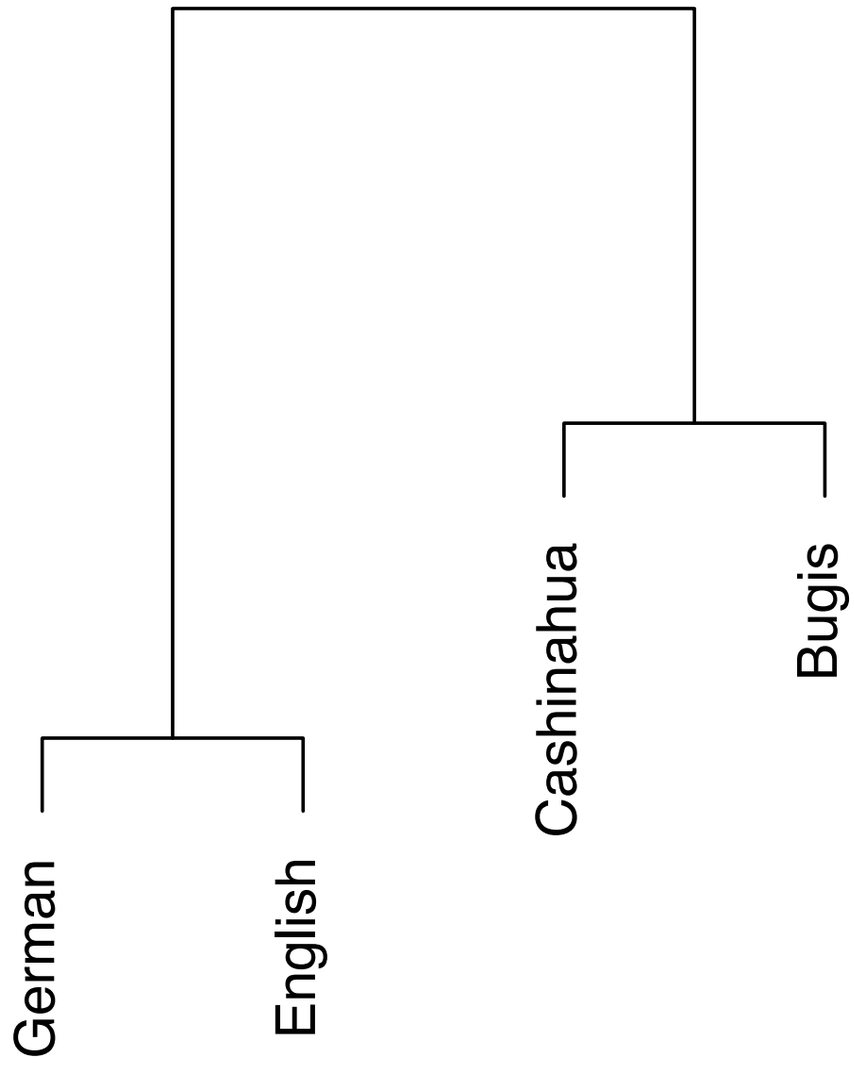

Bugis
Fraction of all words
0.00
0.04
0.08
0.12
1 3 5 7 9 11 14 17 20
Length of words

	H.N.	Eng.	Ger.	Cash.	Bug.	Inu.
Hmong Nua	0	0.60	0.53	0.58	0.74	1
English	0.60	0	0.19	0.32	0.23	0.74
German	0.53	0.19	0	0.23	0.27	0.66
Cashinahua	0.58	0.32	0.23	0	0.25	0.70
Bugis	0.74	0.23	0.27	0.25	0	0.68
Inuktitut	1	0.74	0.66	0.70	0.68	0

Average wordlength



Wordlength distribution



Language similarities ?!

- Similarities between languages do not follow automatically from the data !
- It has to be explicitly stated how the similarities are arrived at
- Different kinds of similarities are possible with the same data