Language typology as relational measurement

Michael Cysouw
WiP, April 2008

Measurement theory

Measurement theory

- Stevens (1946)
  - from a psychological background

Measurement theory

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

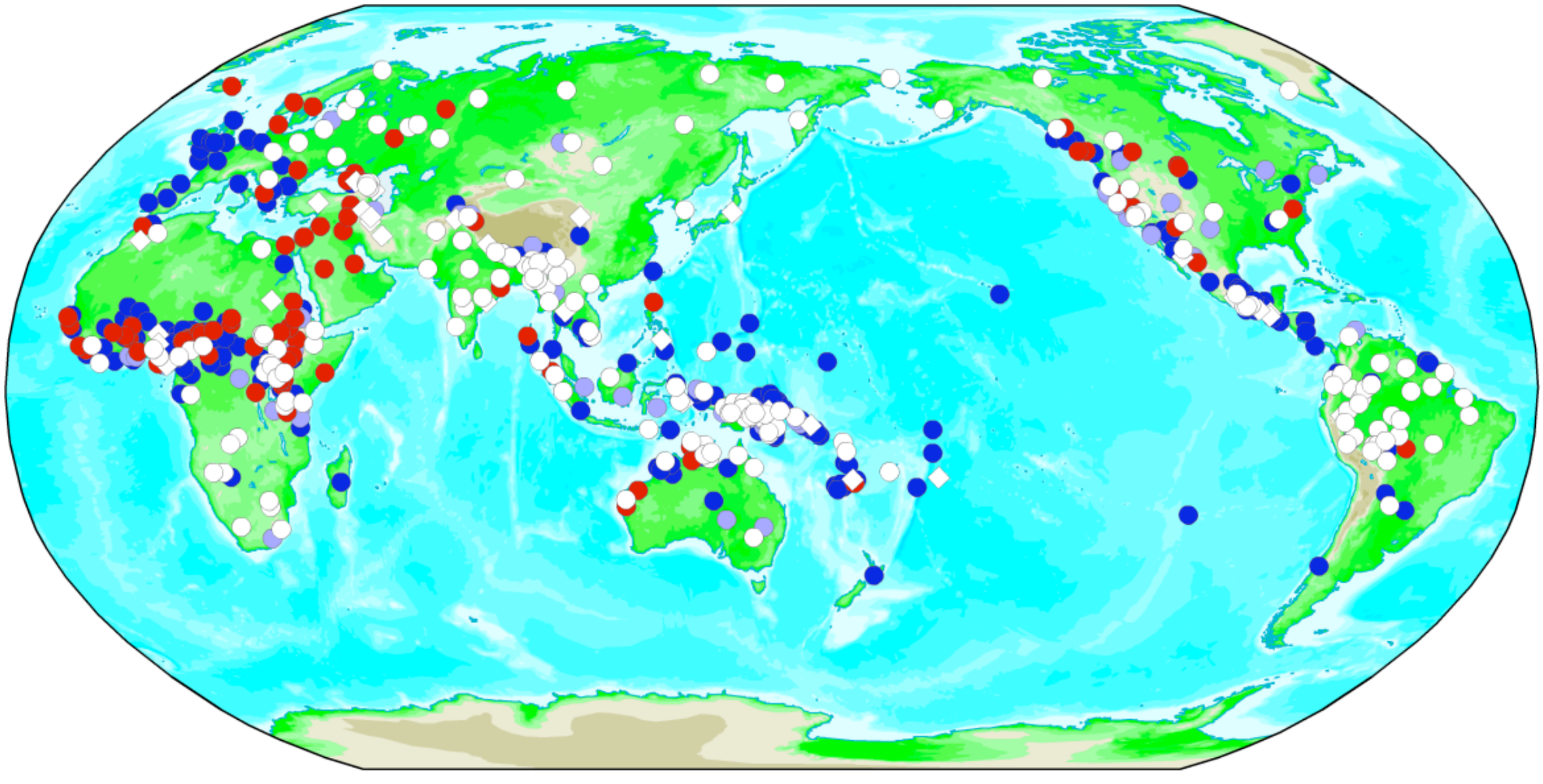
Measurement theory

- 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)

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.

Categorization
(nominal variable)

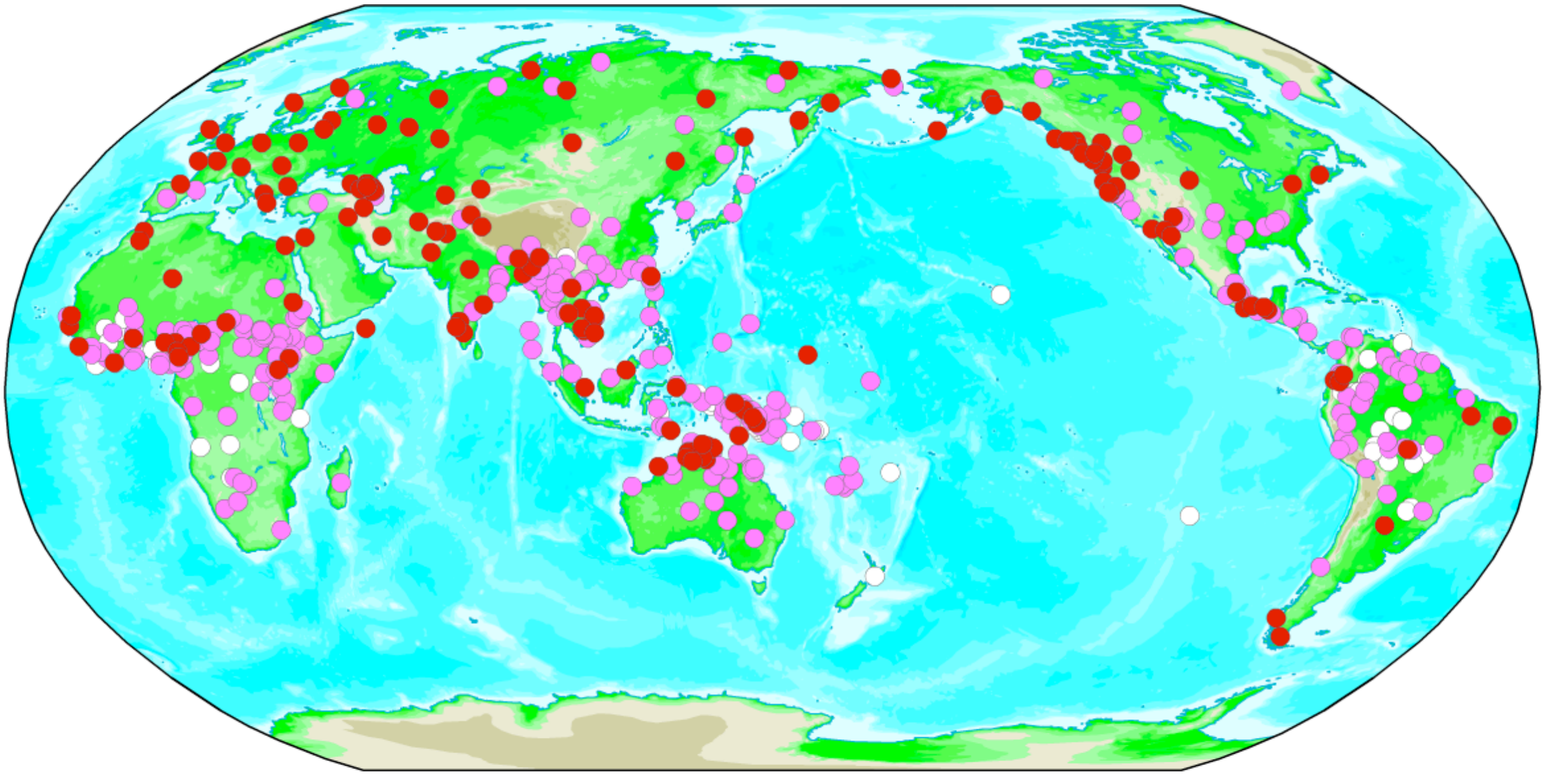
1. Definite word distinct from demonstrative
2. Demonstrative word used as definite article
3. Definite affix
4. No definite, but indefinite article
5. No definite or indefinite article

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)

Linearly ordered categorization

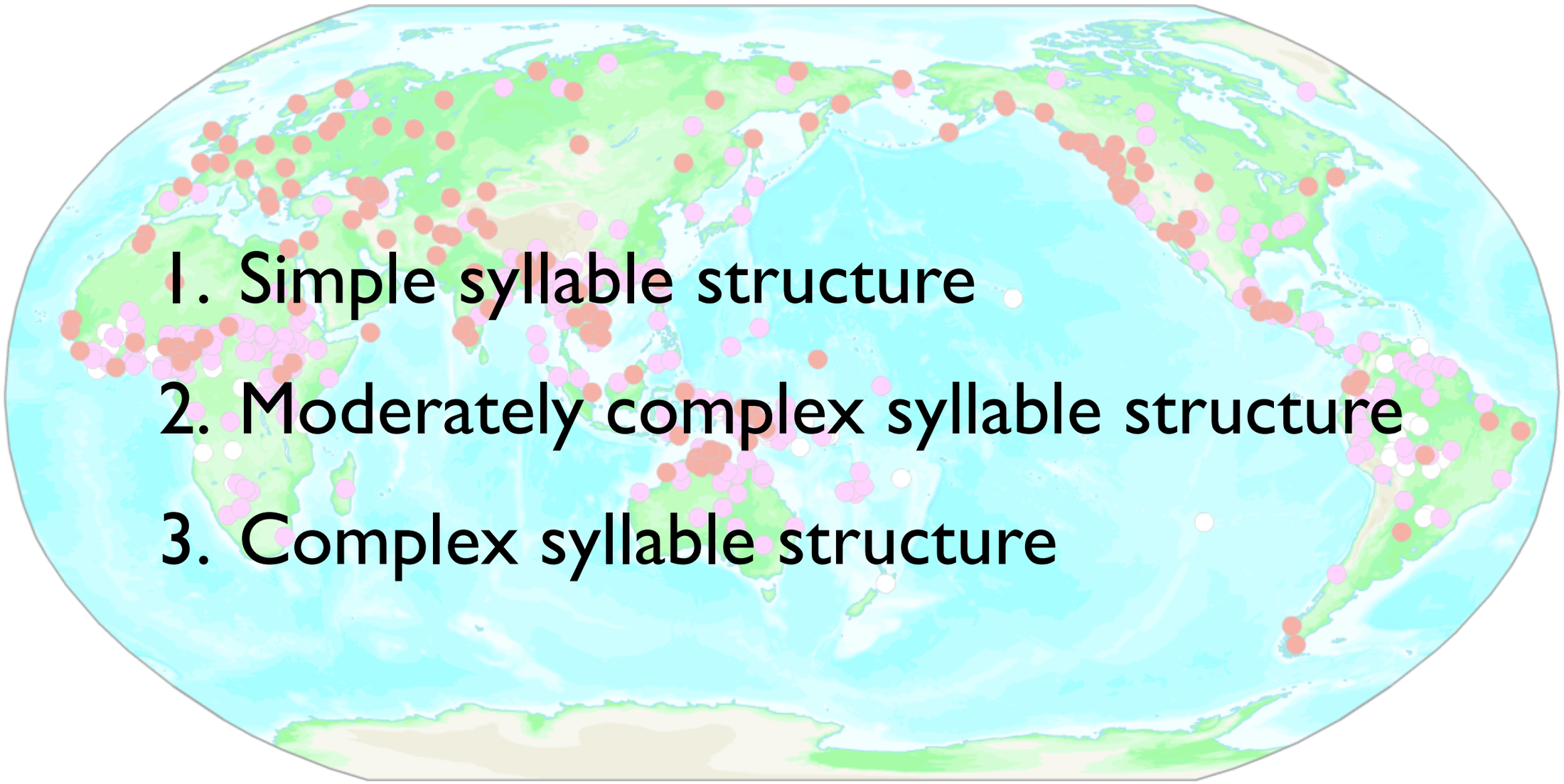
(interval variable)



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.

Linearly ordered categorization

(interval variable)

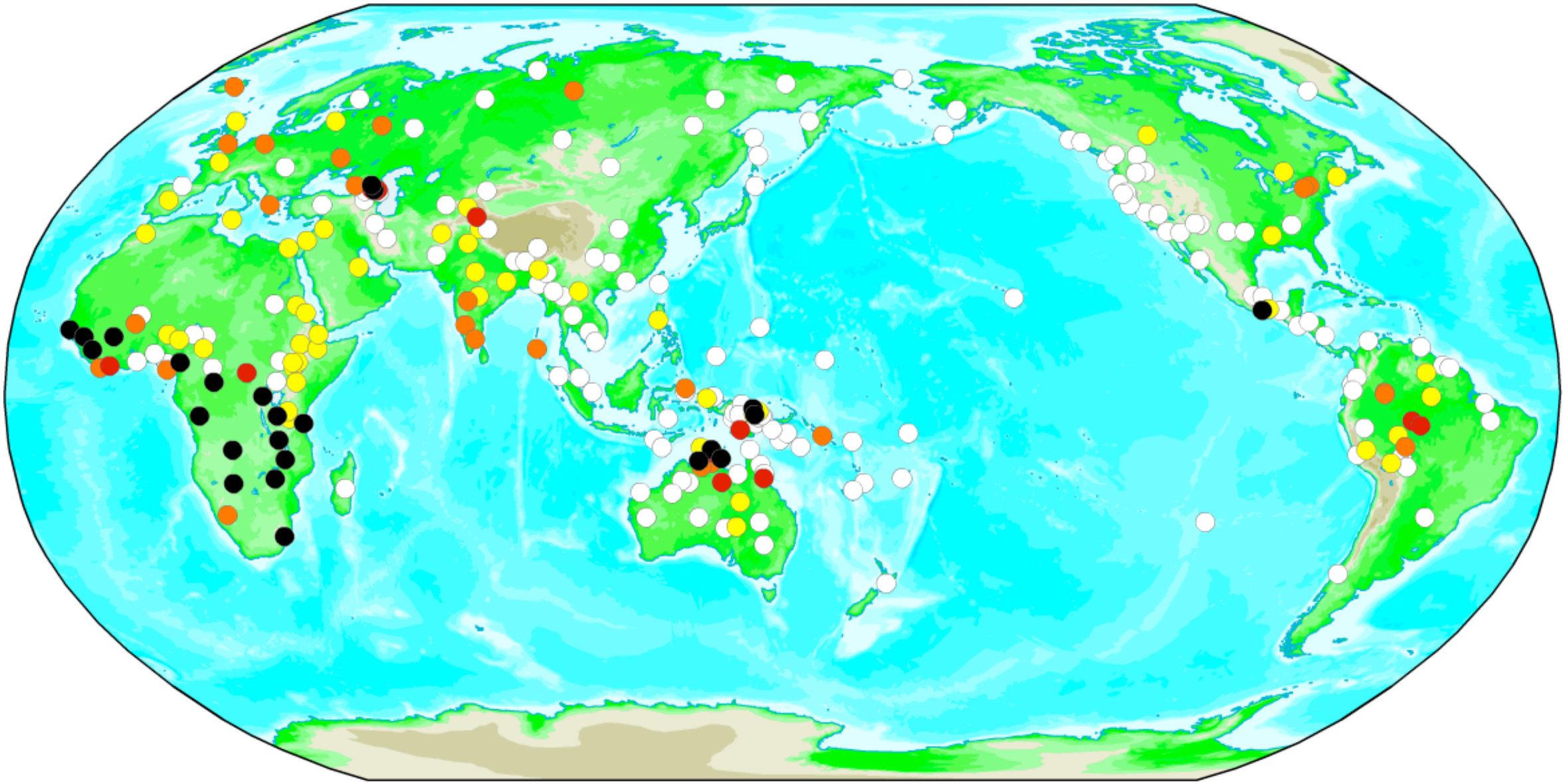


Maddieson, Ian (2005) 'Syllable structure' in: Martin Haspelmath, Matthew S. Dryer, David Gil, & Bernard Comrie (eds.) *World Atlas of Language Structures*. Oxford: Oxford University Press, 54-57.

Count
(ratio variable)

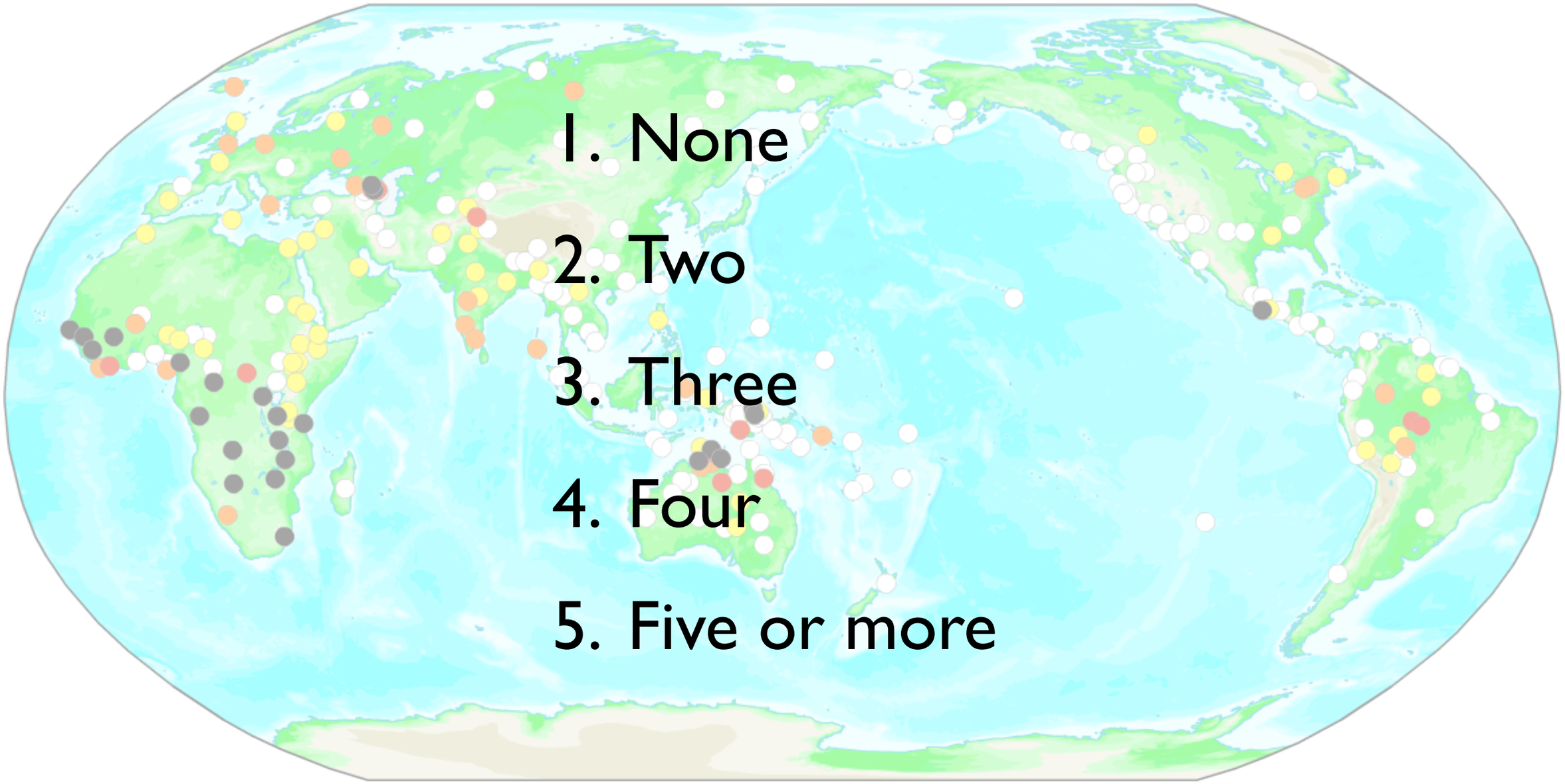
Count

(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.

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)

Continuum
(ratio variable)

- Not common in language comparison

Continuum
(ratio variable)

- Not common in language comparison
- Examples:
  - physical characteristics of speech
  - averages of corpus counts

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

Continuum
(ratio variable)

- Not common in language comparison
- Examples:
  - physical characteristics of speech
  - averages of corpus counts
- Watch out with the interpretation of combinations of various (*a priori*) independent counts (e.g. sum or fraction)

Problems

Problems

- More measurements wanted

Problems

- More measurements wanted
  - more specification in categorization

Problems

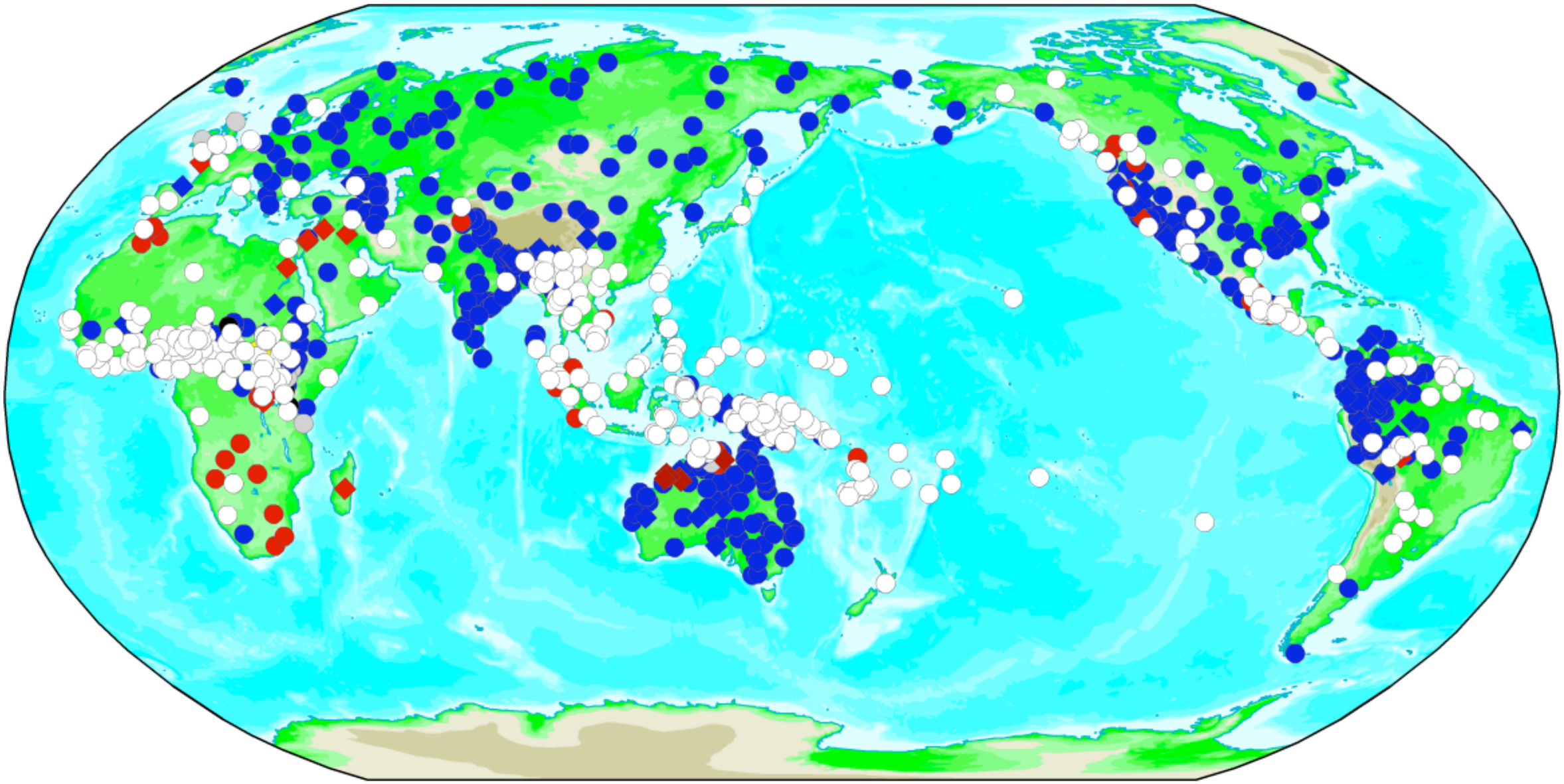
- More measurements wanted
  - more specification in categorization
  - full pairwise comparisons

Problems

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

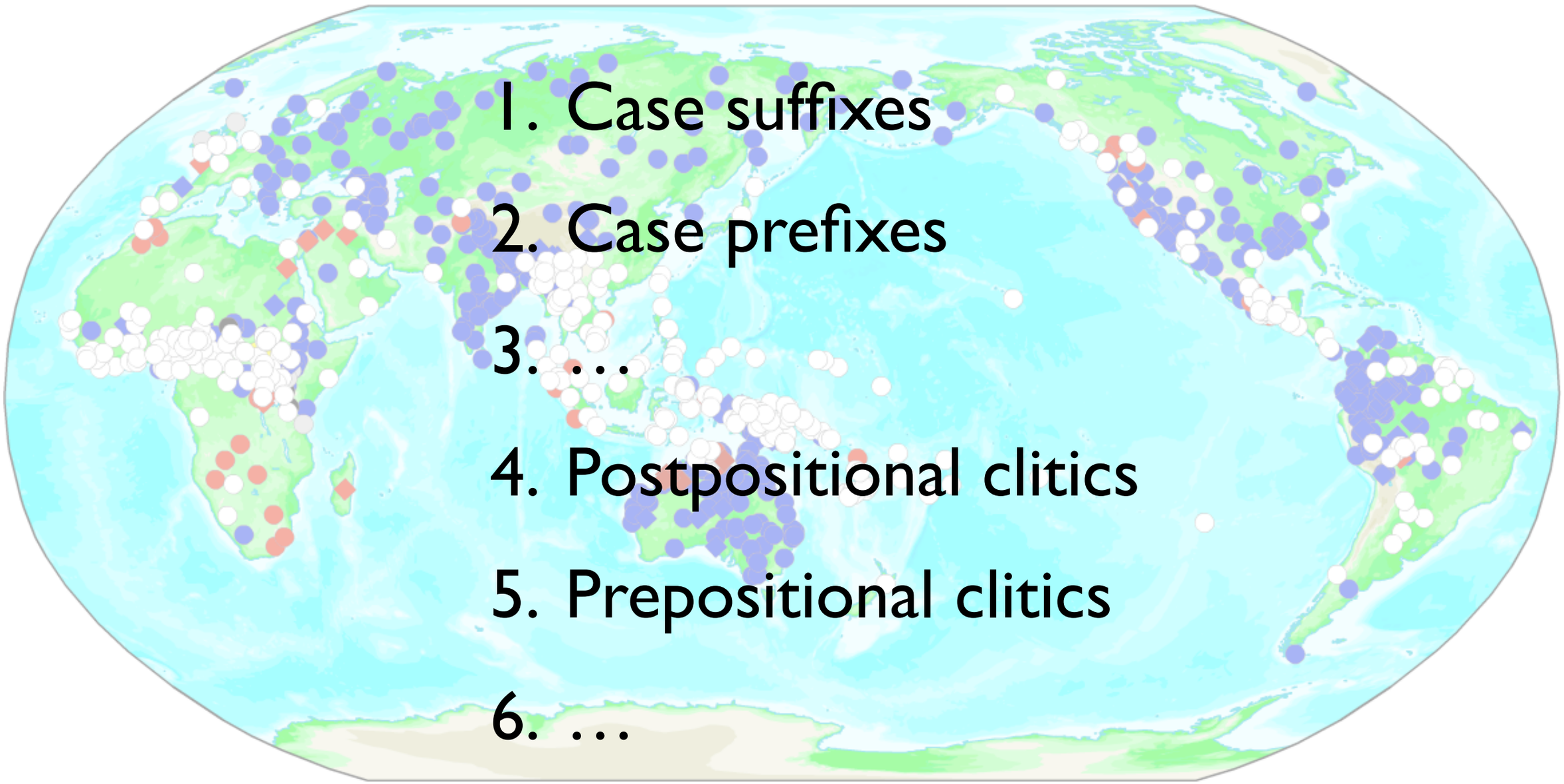
More specification for categorizations

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.

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.

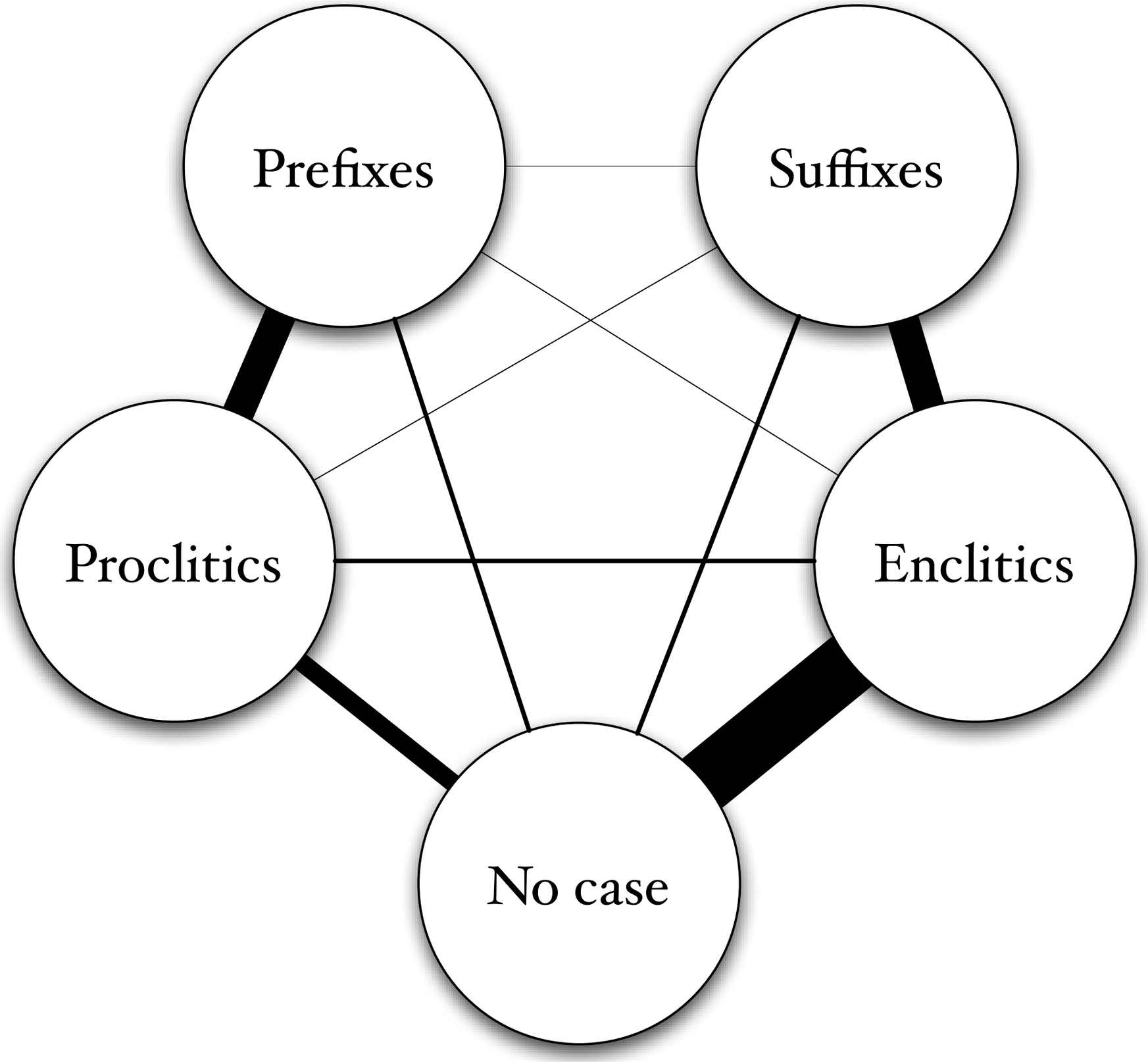
Prefixes

Suffixes

Proclitics

Enclitics

No case


Prefixes
Suffixes
Proclitics
Enclitics
No case

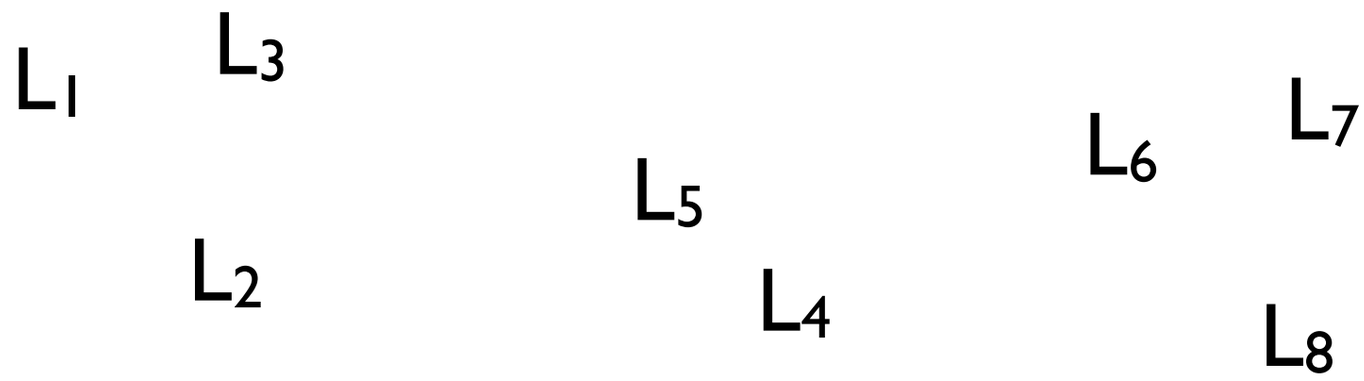
Relational metaphor of measurement

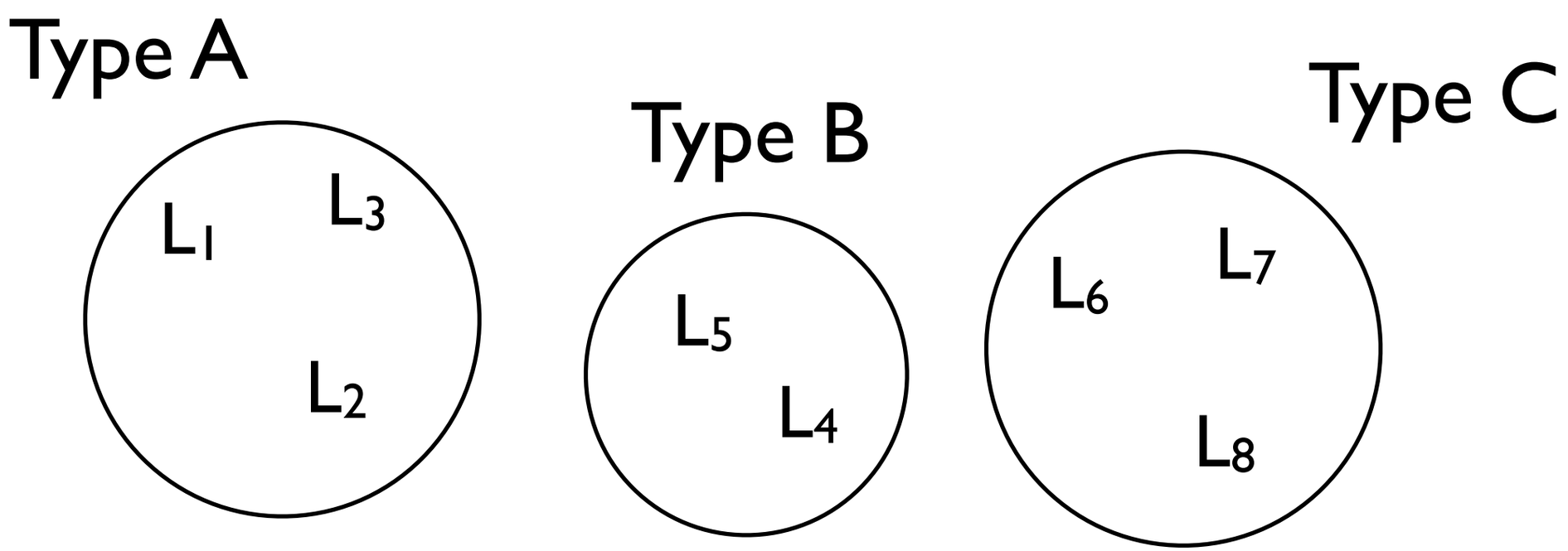
Relational metaphor of measurement

- Express typology as pairwise language-to-language similarities

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


L1
L3
L5
L6
L7
L2
L4
L8


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	
…									

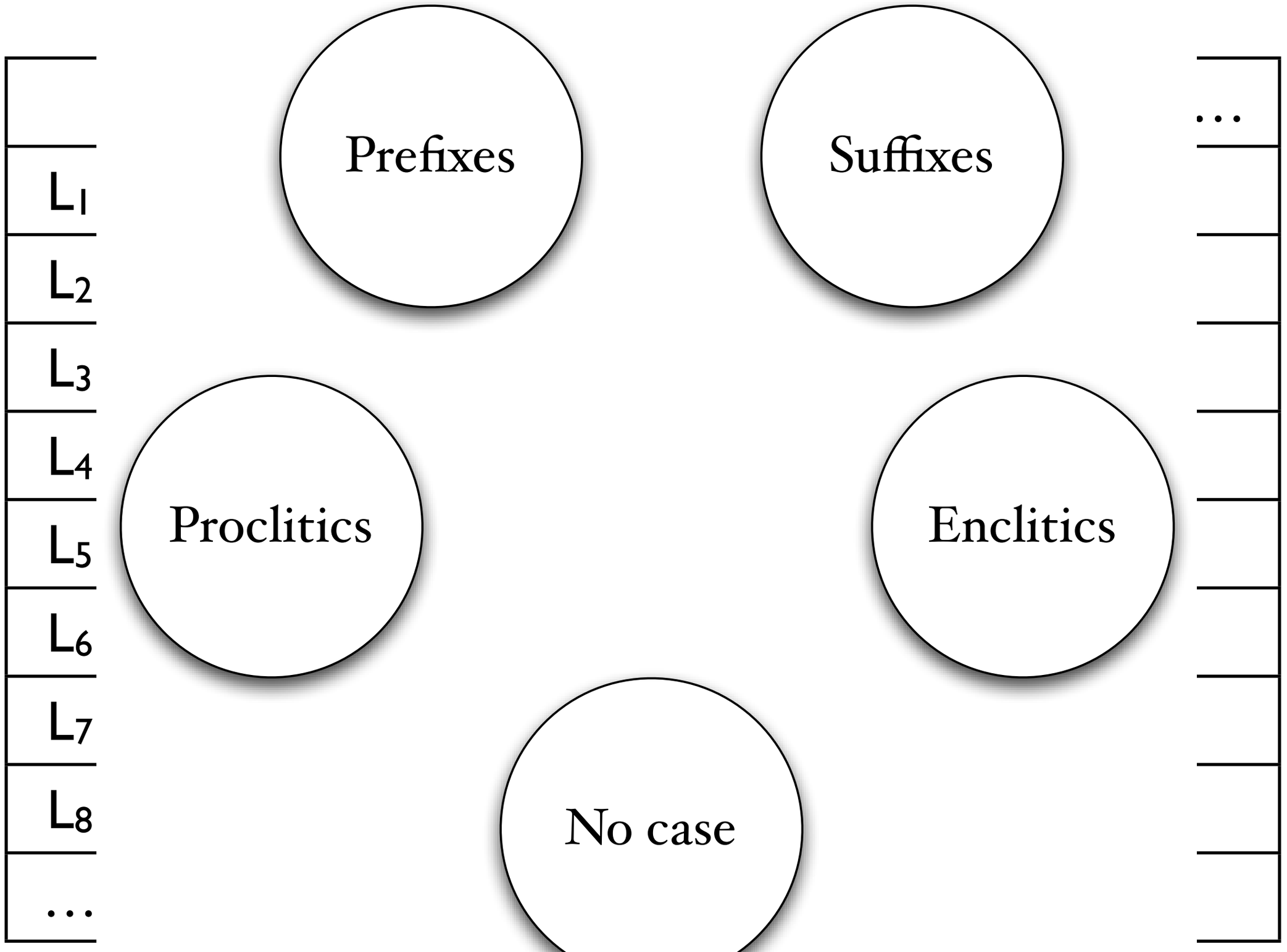
	Type A			Type B		Type C			
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	I	I	I						
L_2	I	I	I						
L_3	I	I	I						
L_4				I	I				
L_5				I	I				
L_6						I	I	I	
L_7						I	I	I	
L_8						I	I	I	
…									

Type A Type B Type C

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

Undifferentiated Categorization

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	0	0	
L_2	1	1	1	0	0	0	0	0	
L_3	1	1	1	0	0	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	
…									

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	
…									

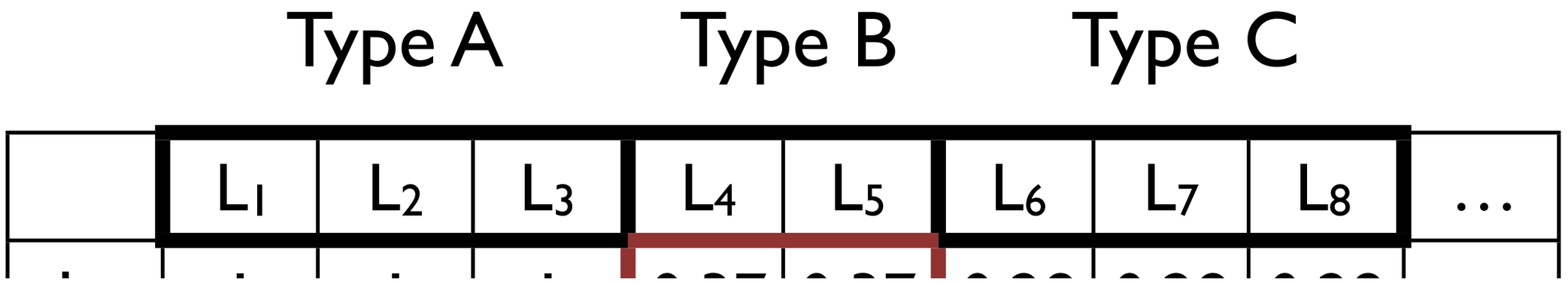
Type A Type B Type C

	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	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	
…									

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	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	
…									

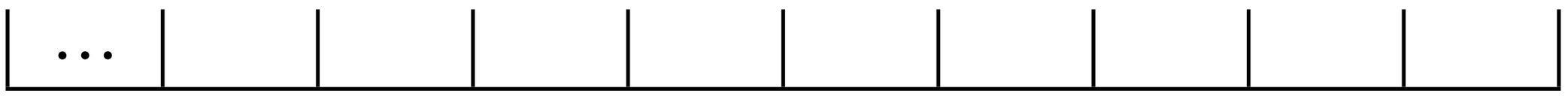
Differentiated Categorization



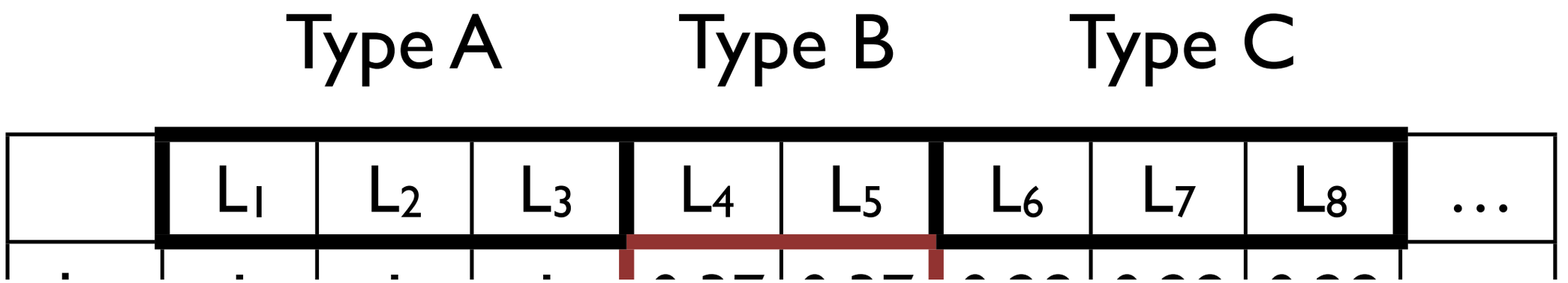
A. Simple syllable structure

B. Moderately complex syllable structure

C. Complex syllable structure



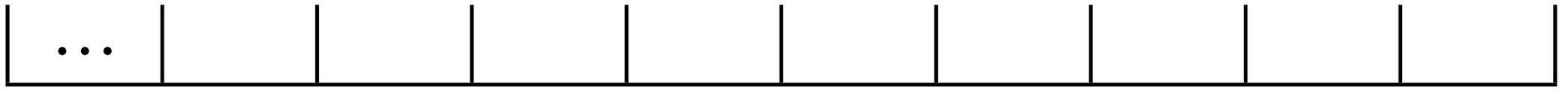
Differentiated Categorization



A. Simple syllable structure

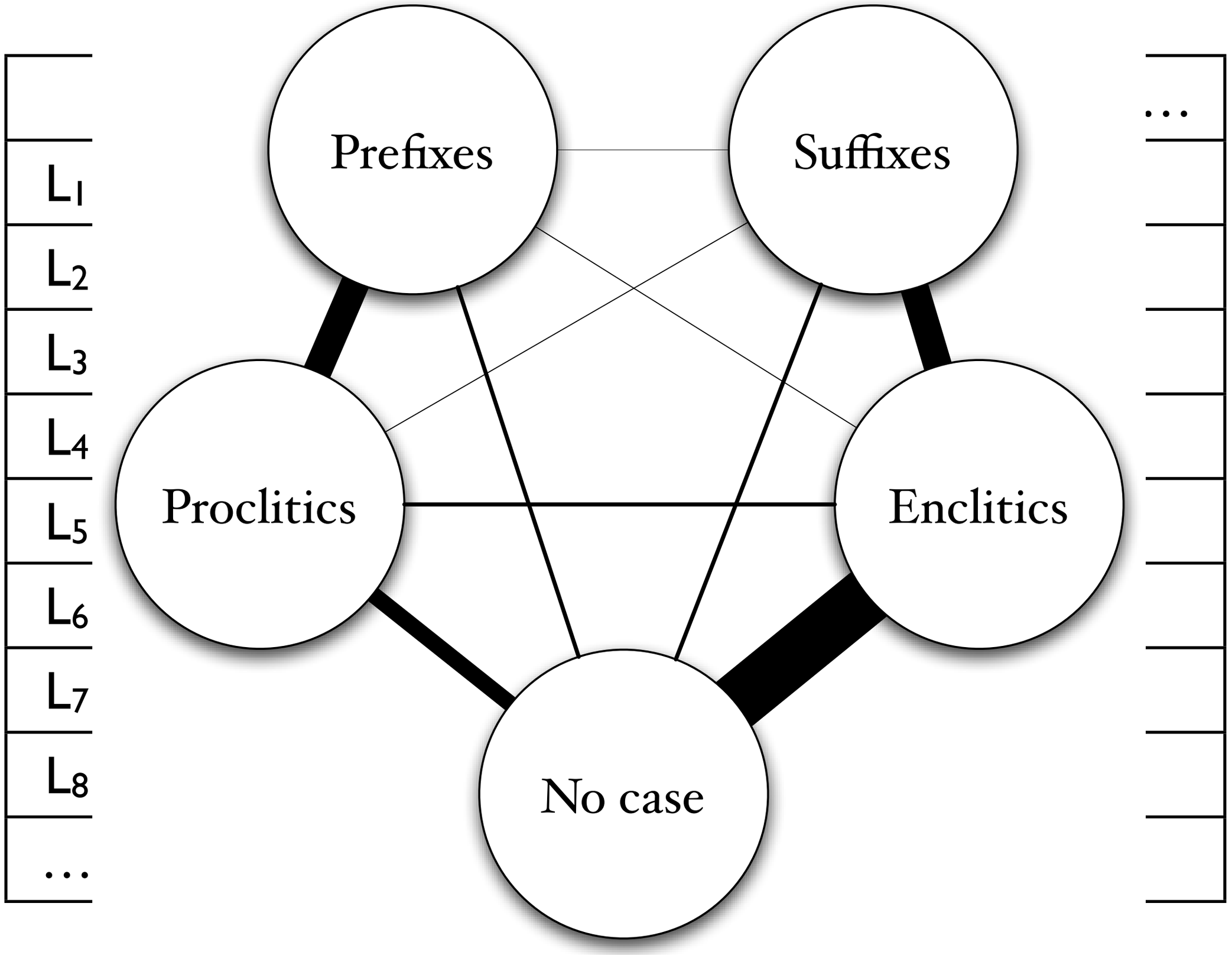
B. Moderately complex syllable structure

C. Complex syllable structure



Differentiated Categorization

Type A Type B Type C



Differentiated 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	0.37	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	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

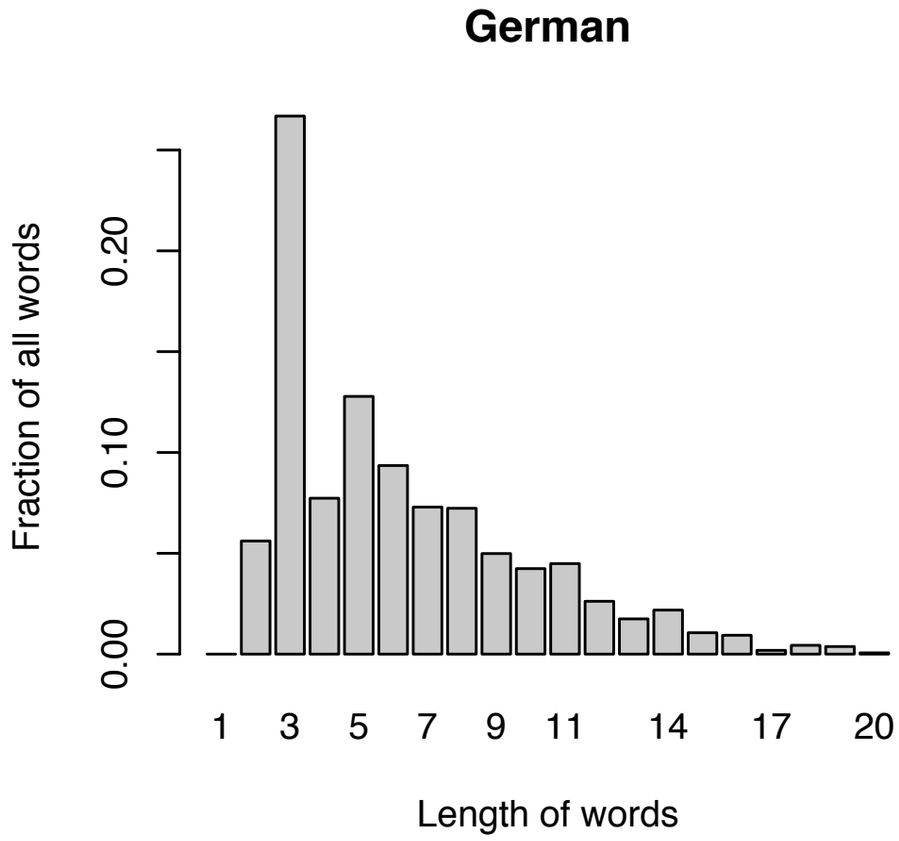
	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8	…
L_1	1	0.55	0.72	0.31	0.70	0.61	0.50	0.58	
L_2	0.55	1	0.55	0.31	0.40	0.44	0.31	0.48	
L_3	0.72	0.55	1	0.29	0.53	0.51	0.48	0.60	
L_4	0.31	0.31	0.29	1	0.38	0.36	0.26	0.27	
L_5	0.70	0.40	0.53	0.38	1	0.64	0.51	0.46	
L_6	0.61	0.44	0.51	0.36	0.64	1	0.57	0.43	
L_7	0.50	0.31	0.48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

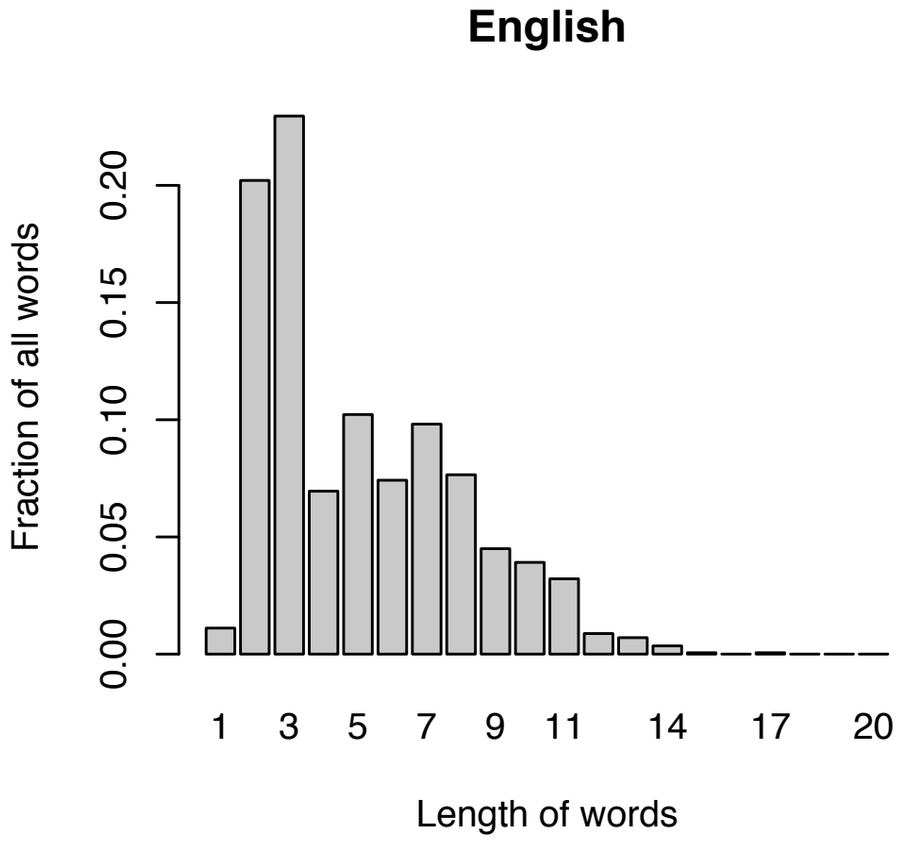
‘Deconstructed’ Typology

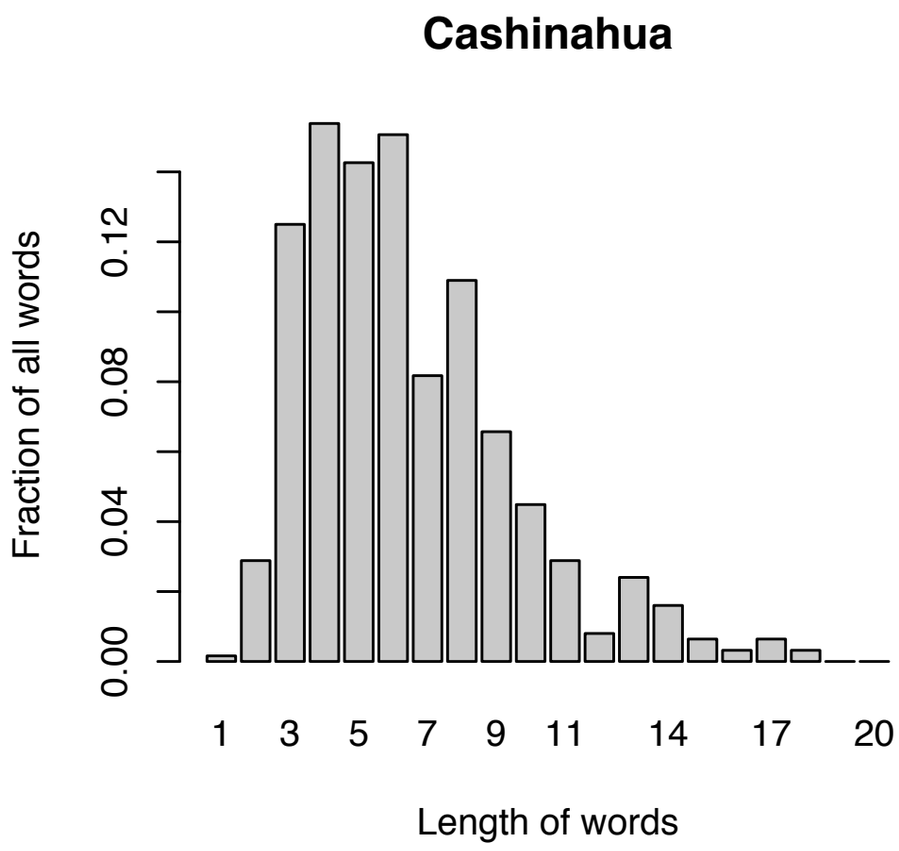
Pairwise Comparison

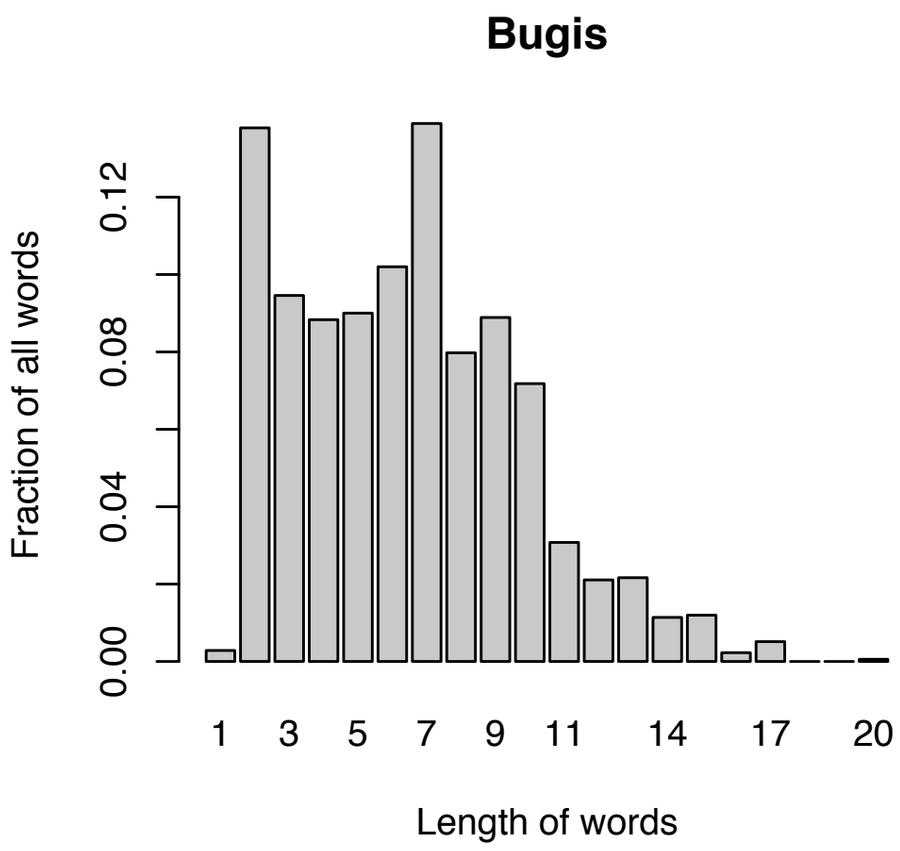
Pairwise Comparison

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

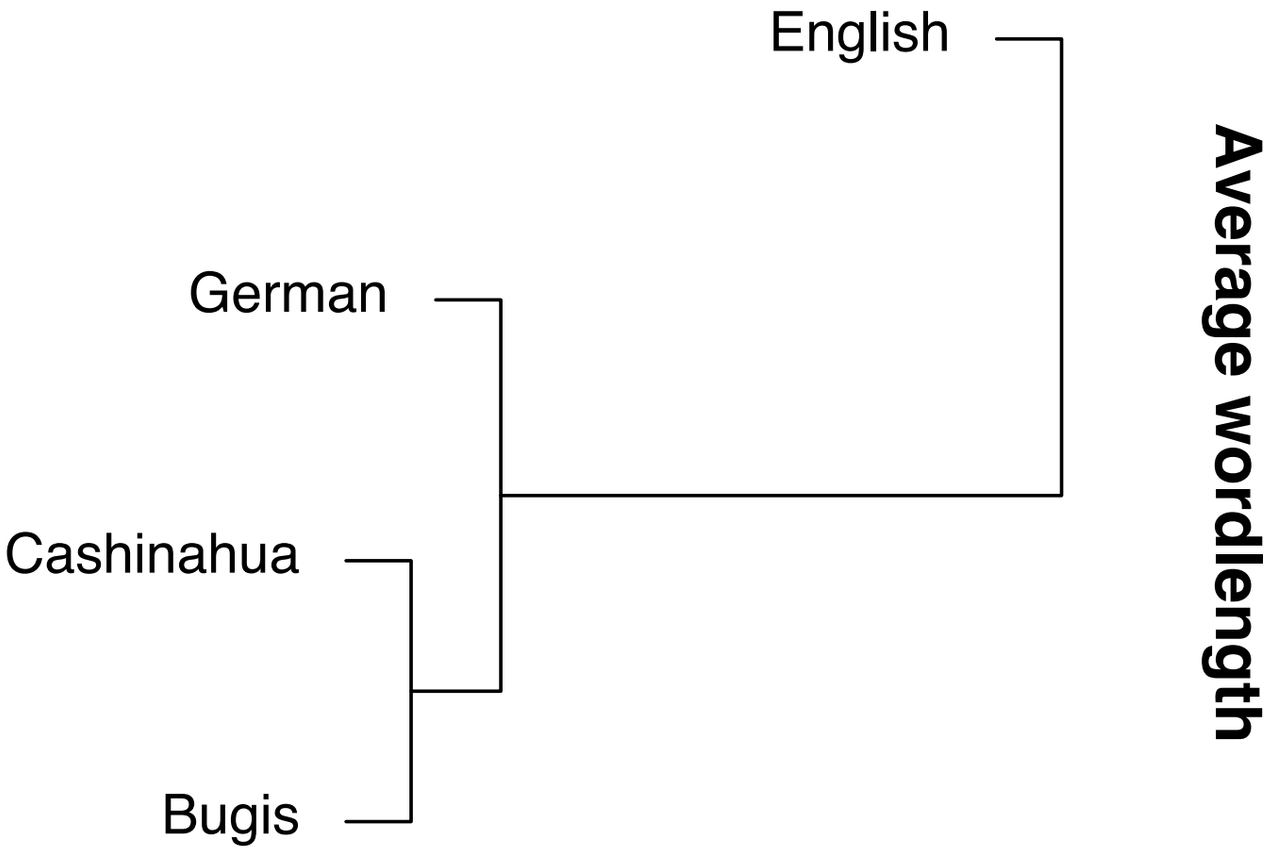

German
Fraction of all words
0.00
0.10
0.20
1 3 5 7 9 11 14 17 20
Length of words

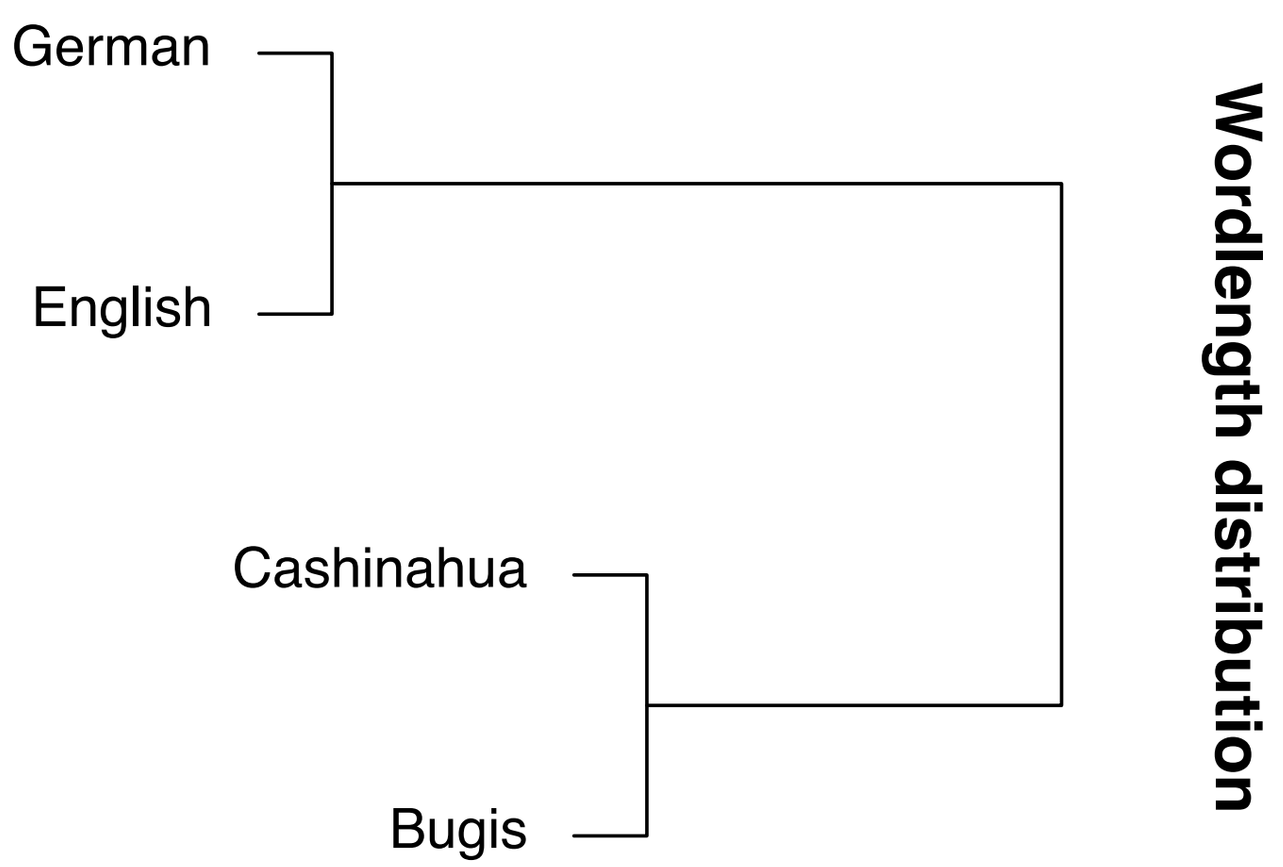

English
Fraction of all words
0.00
0.05
0.10
0.15
0.20
1 3 5 7 9 11 14 17 20
Length of words


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

Hmong Nua	0	0.60	0.53	0.58	0.74	I
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	I	0.74	0.66	0.70	0.68	0


English
German
Cashinahua
Bugis
Average wordlength


German
English
Cashinahua
Bugis
Wordlength distribution

	Data	Similarity

	Data	Similarity
undifferentiated categorization	types of languages	implicit (all types are equally dissimilar to each other)

	Data	Similarity
undifferentiated categorization	types of languages	implicit (all types are equally dissimilar to each other)
count, continuum	values measured	implicit similarity function (typically absolute difference)

	Data	Similarity
undifferentiated categorization	types of languages	implicit (all types are equally dissimilar to each other)
count, continuum	values measured	implicit similarity function (typically absolute difference)
differentiated categorization	types of languages	specify similarity between types (empirically or not)

	Data	Similarity
undifferentiated categorization	types of languages	implicit (all types are equally dissimilar to each other)
count, continuum	values measured	implicit similarity function (typically absolute difference)
differentiated categorization	types of languages	specify similarity between types (empirically or not)
count, continuum	values measured	specific similarity function (e.g. stressing low differences)

	Data	Similarity
undifferentiated categorization	types of languages	implicit (all types are equally dissimilar to each other)
count, continuum	values measured	implicit similarity function (typically absolute difference)
differentiated categorization	types of languages	specify similarity between types (empirically or not)
count, continuum	values measured	specific similarity function (e.g. stressing low differences)
'deconstructed' typology	whatever (e.g. collection of word lengths)	whatever (e.g. histogram similarity)

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

Typology as language similarities

Typology as language similarities

- Separating data and similarity is both a curse and a blessing

Typology as language similarities

- Separating data and similarity is both a curse and a blessing
- **Curse**: it is necessary to be much more precise in what it takes for two languages to be similar

Typology as language similarities

- Separating data and similarity is both a curse and a blessing
- **Curse**: it is necessary to be much more precise in what it takes for two languages to be similar
- **Blessing**: Such precision results in much more consistent and fine-grained language typologies