Typology and Corpora

An Eternal Golden Braid

Michael Cysouw
MPI-EVA

What is typology?

- The study of differences between languages
- Are there limits on the structure of language?
- How should such limits be explained?

Typology versus Universals

- Typology: language A is of type X
  - 'English is isolating'
  - 'English puts the possessor before the possessed object'
- Universals: for all languages, property P holds
  - 'All languages have negation'
  - 'If a language has Verb-Object order, then it has Possessor-Possessed order'

Why do typology?

- Look at correlations between types, not necessarily absolute/universal
- What is the reason for particular significant correlations?
  - human biological endowment (UG)
  - human cognitive structure
  - structure of communication
  - structure of society
  - incidents of history

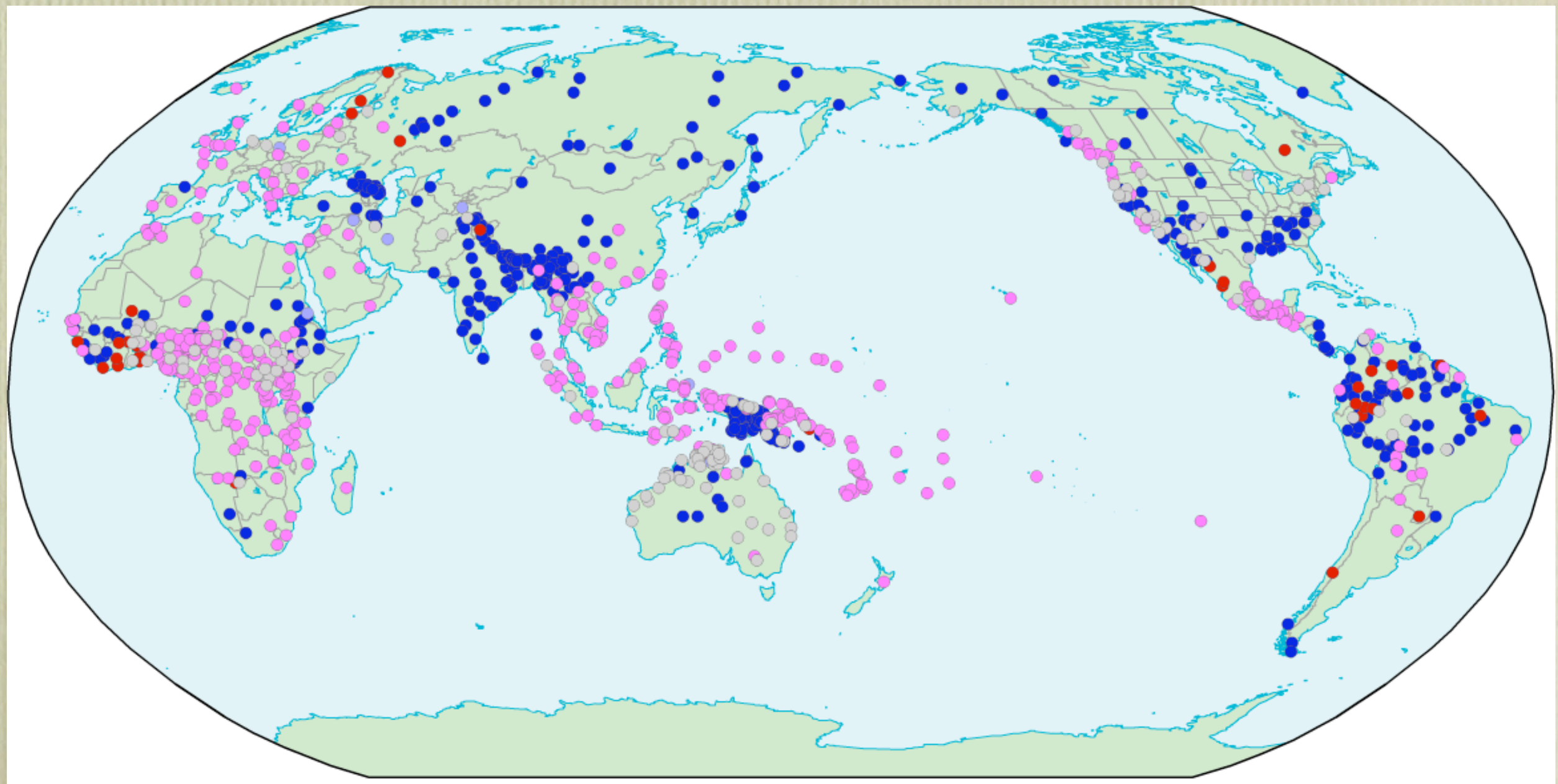
Word order correlations

	OV	VO
Prepositions	11	416
Postpositions	427	39

(140 language in between...)

(Data from Dryer 2005, as published in the *World Atlas of Language Structure*)

Word order correlations



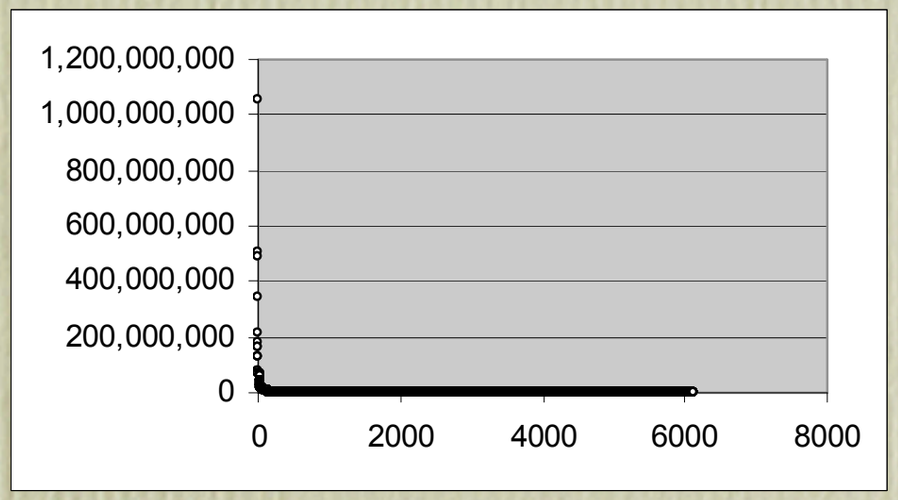
Establishing typological correlations

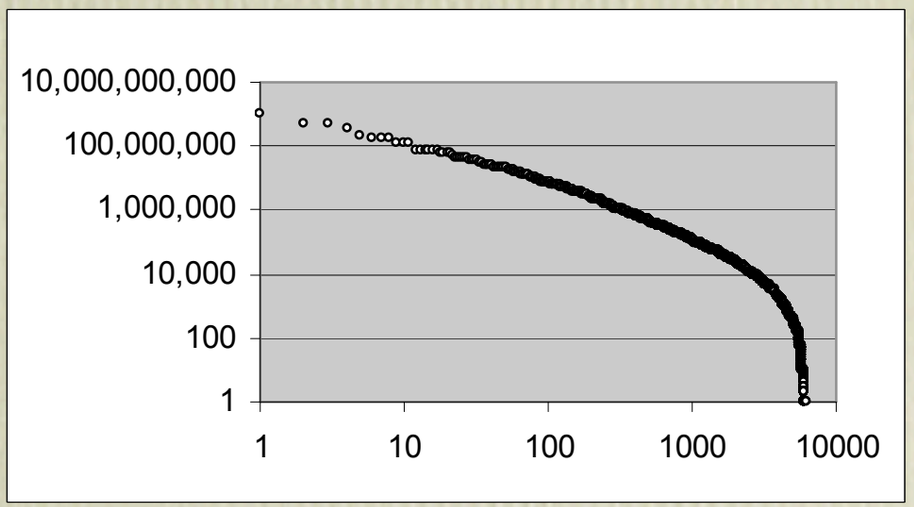
- Use many different languages
  - How many is many?
  - How different is different?
- Generalisations depend on sample used
- Our approach here at MPI-EVA:
  - How many: 100-400 (and up...)
  - How different: capture world-wide variation

The world's languages

Languages are vanishing!

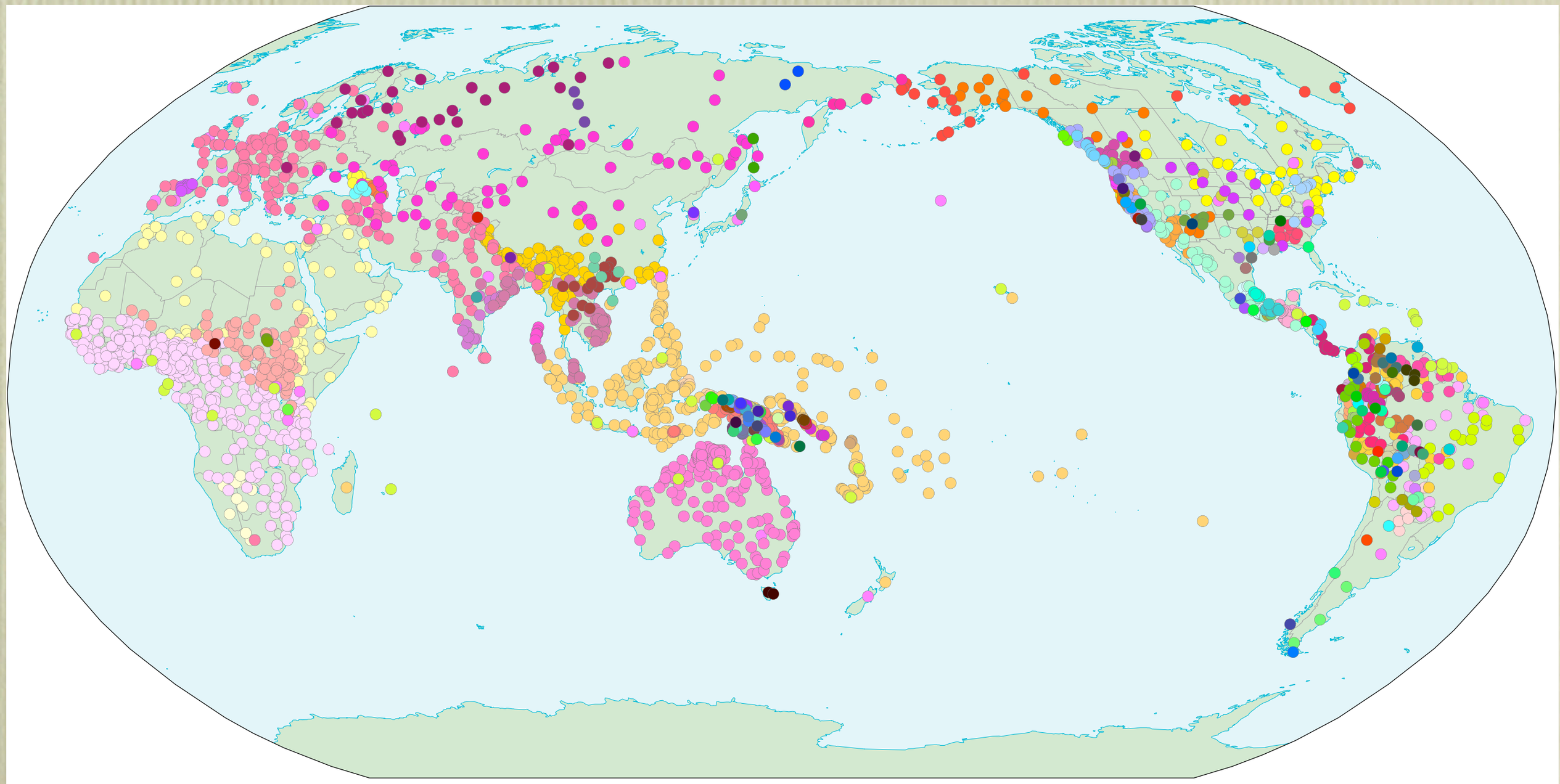
Number of speakers per languages
(rank-ordered, aka 'Zipfian order')



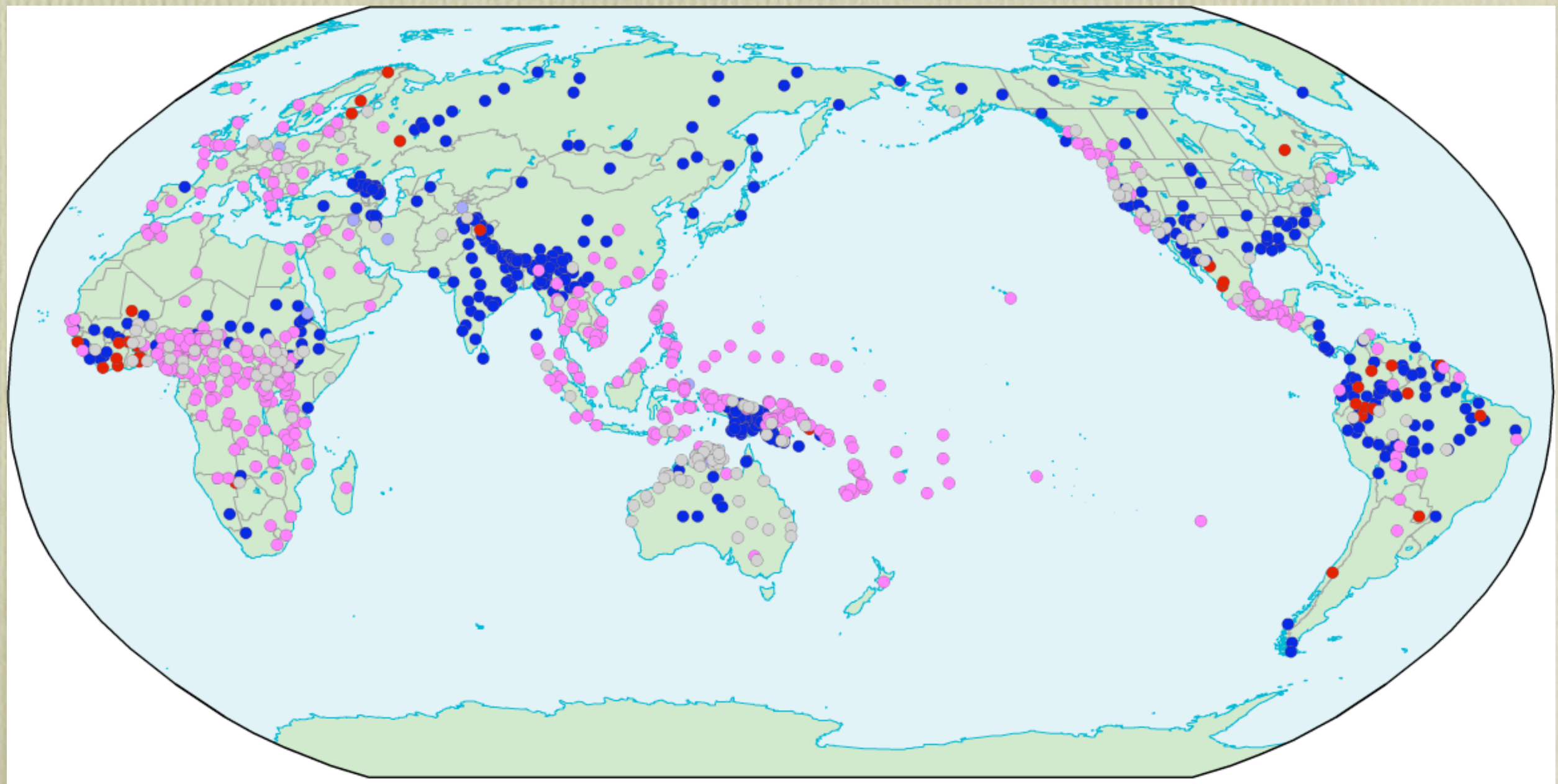


(Graphs from Wichmann 2005, using data from the *Ethnologue*, Grimes 2004)

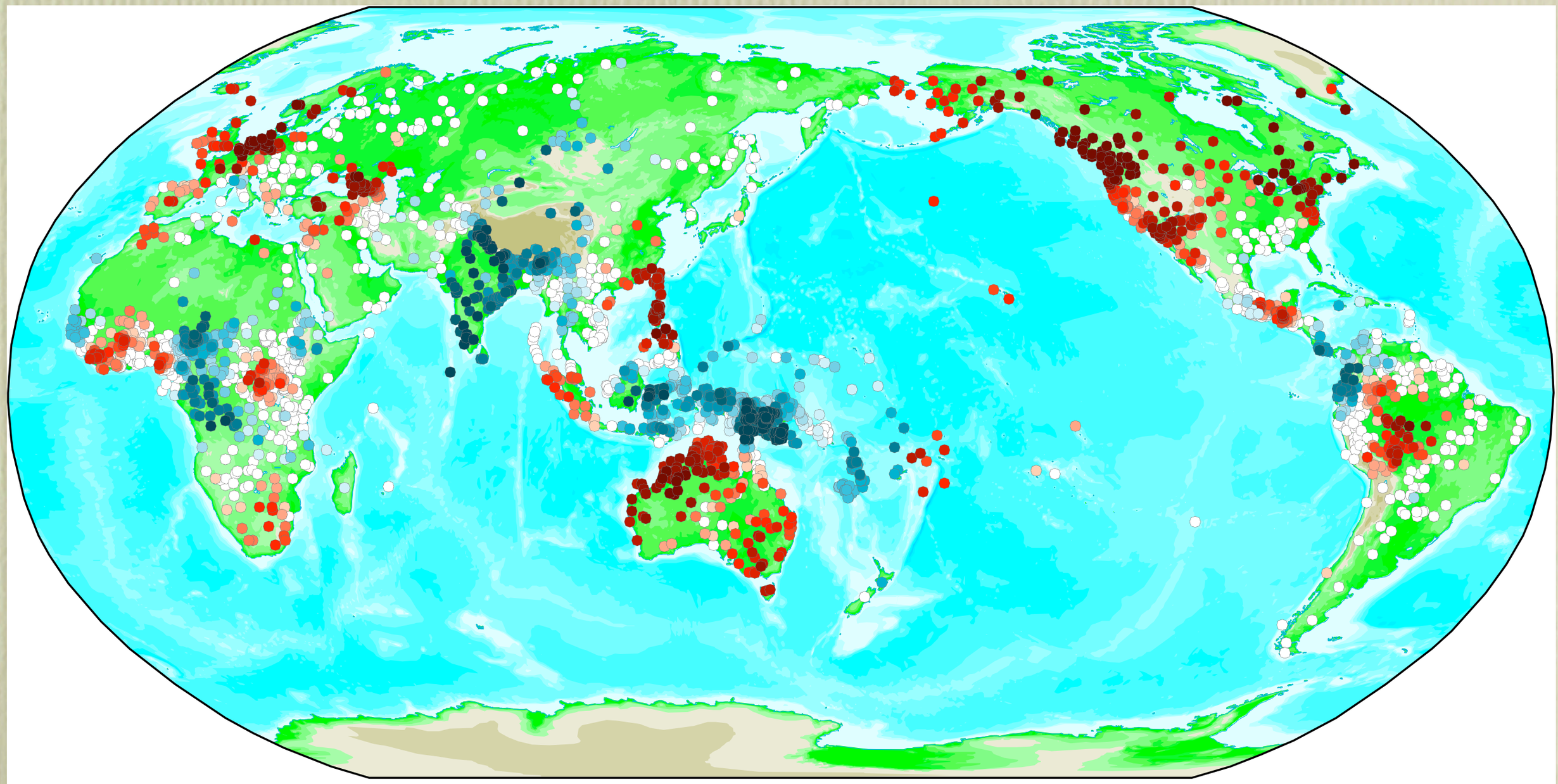
Linguistic families



Areal patterns



Areal patterns



How can corpora help?

- Using continuous (or more fine grained) types, instead of categorical ‘yes/no’-types
- Automatically establishing types on the basis of values extracted from a corpus
- Establishing new kinds of types
- Help disentangle possible reasons for correlations by looking at usage (performance)

Continuous types

- Traditionally, typologists used discrete types
- All languages are forced into a type
- Boundaries between types are problematic
- Better: use frequency instead of yes/no answer

Continuous types

Word Order in Hanis Coos

SVO	6 (38%)	SV	30 (23 %)	OV	17 (30%)
VOS	4 (25%)	VS	98 (77%)	VO	39 (70%)
VSO	3 (19%)				
OVS	3 (19%)				

(Data from Dryer 1997: 81)

Automatically extract types

- Automatically extract a type out of a corpus
  - number of words per sentence
  - number of letters/phonemes per word
  - number of morphemes per word
- Such measures are claimed to converge extremely quick, but that has to be tested

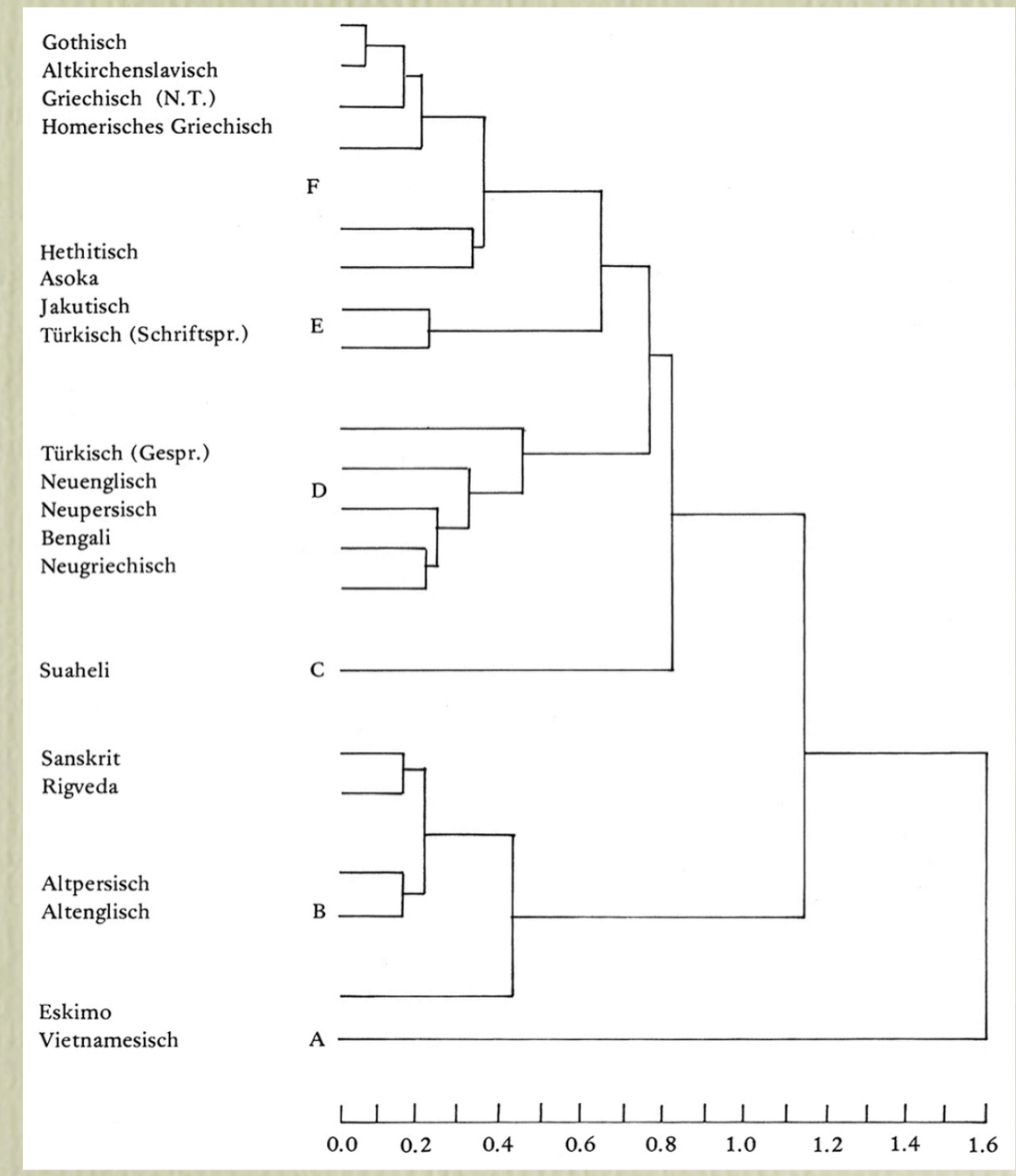
Automatically extract types

Tabelle I: Die Indizes von Greenberg-Krupa für 20 Sprachen

	W/M	A/J	W/R	D/M	I/M	P/M	S/M	O/N	Pi/N	Co/N
Sanskrit	0.39	0.09	0.88	0.24	0.32	0.06	0.46	0.16	0.46	0.38
Bengali	0.53	0.46	0.92	0.15	0.28	0.01	0.42	0.57	0.29	0.14
Altpersisch	0.41	0.20	0.98	0.17	0.41	0.08	0.50	0.23	0.39	0.38
Neupersisch	0.66	0.34	0.97	0.07	0.26	0.01	0.32	0.52	0.29	0.19
Griechisch (Homer)	0.48	0.10	0.99	0.10	0.41	0.03	0.48	0.48	0.27	0.26
Neugriechisch	0.55	0.40	0.98	0.07	0.37	0.02	0.42	0.53	0.21	0.26
Altenglisch	0.47	0.11	1.00	0.09	0.42	0.03	0.48	0.15	0.47	0.38
Neuenglisch	0.60	0.30	1.00	0.09	0.32	0.02	0.38	0.75	0.14	0.11
Jakutisch	0.46	0.51	0.98	0.16	0.38	0.00	0.53	0.29	0.59	0.12
Suaheli	0.39	0.67	1.00	0.03	0.31	0.45	0.16	0.40	0.19	0.41
Vietnamesisch	0.94	–	0.93	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Eskimo	0.27	0.03	1.00	0.34	0.47	0.00	0.73	0.02	0.46	0.38
Türkisch (Schriftspr.)	0.43	0.60	1.00	0.11	0.43	0.00	0.54	0.43	0.67	0.20
Türkisch (Gespr. Spr.)	0.57	0.67	0.96	0.06	0.38	0.00	0.44	0.69	0.16	0.03
Gotisch	0.43	0.19	0.97	0.13	0.42	0.04	0.52	0.37	0.34	0.29
Altkirchenslavisch	0.44	0.20	1.00	0.15	0.41	0.05	0.51	0.41	0.33	0.26
Hethitisch	0.51	0.42	1.00	0.12	0.36	0.01	0.48	0.35	0.32	0.33
Rigveda	0.39	0.08	0.91	0.19	0.38	0.07	0.48	0.26	0.48	0.27
Asoka	0.40	0.26	0.82	0.17	0.34	0.03	0.49	0.40	0.18	0.42
Griechisch (N.T.)	0.41	0.12	0.97	0.11	0.47	0.07	0.51	0.34	0.32	0.34

(Figures from Altmann/Lehfeldt 1973: 40-1)

Automatically extract types



(Figures from Altmann/Lehfeldt 1973: 40-1)

New kinds of types

- New measures can be used, according to the kind of data a corpus gives
  - e.g. mean entropy of a word as based on the overall phoneme frequency
  - Distribution of high-frequency words in the sentence
  - etcetera...

Explain structure by frequency

- Grammatical structure reflects frequency distribution of utterances
- Zipfian effects: frequent things get smaller
- or maybe anti-Zipfian effects: smaller things get more frequent?

Explain structure by frequency

- cross-linguistic generalisation: if number is marked, then singular will be smaller than plural
- e.g. *Buch* vs. *Bücher*
- Zipfian explanation: in German singular nouns are more frequent than plural ones
- For cross-linguistic generalisations, such counts should be made for many languages

Explain structure by frequency

English

Difference in wordlength of Prepositional Phrases

Order of PP's

	1	2-4	5-6	7+
small - large	60%	86%	94%	99%
large - small	40%	14%	6%	1%

(Data from Hawkins 2001: 4)

Explain structure by frequency

Japanese

Difference in wordlength of Postpositional Phrases

Order of PP's

	1-2	3-4	5-8	9+
small - large	34%	28%	17%	9%
large - small	66%	72%	83%	91%

(Data from Hawkins 2001: 8)

Explain structure by frequency

But where does frequency come from???

Current plans

- Collect/construct unannotated corpora of as many as possible languages
  - frequency counts
  - insert morphological boundaries
  - insert constituent brackets

Current plans

- Collect (smaller) annotated corpora
- Align parallel corpora:
  - Bible, Declaration of Human rights
  - Lenin, Marx
  - *Le petit prince*, Harry Potter

The End