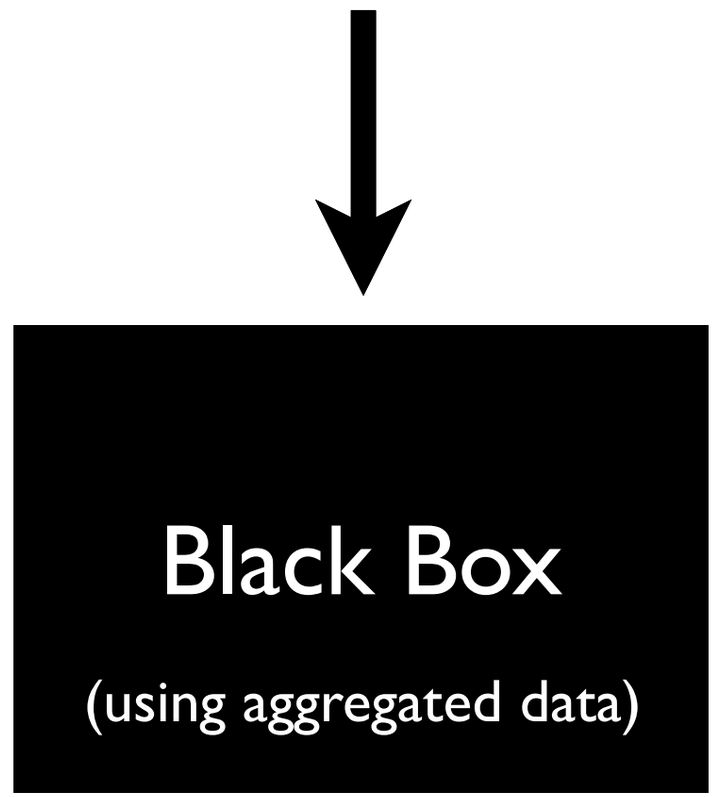
Beyond the black box

A plea for explicit models of language change

Michael Cysouw
Ludwig Maximilians University Munich

Data



Result

Lexical Data

Black Box

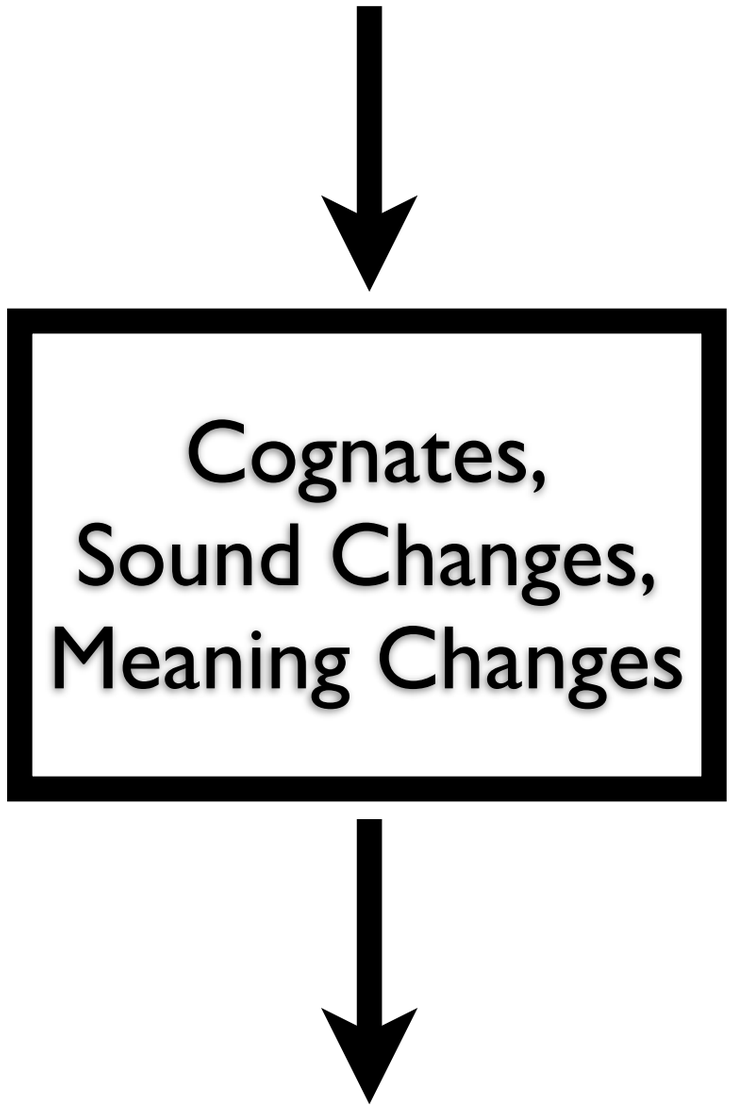
Tree of Languages

Grammatical Data

Black Box

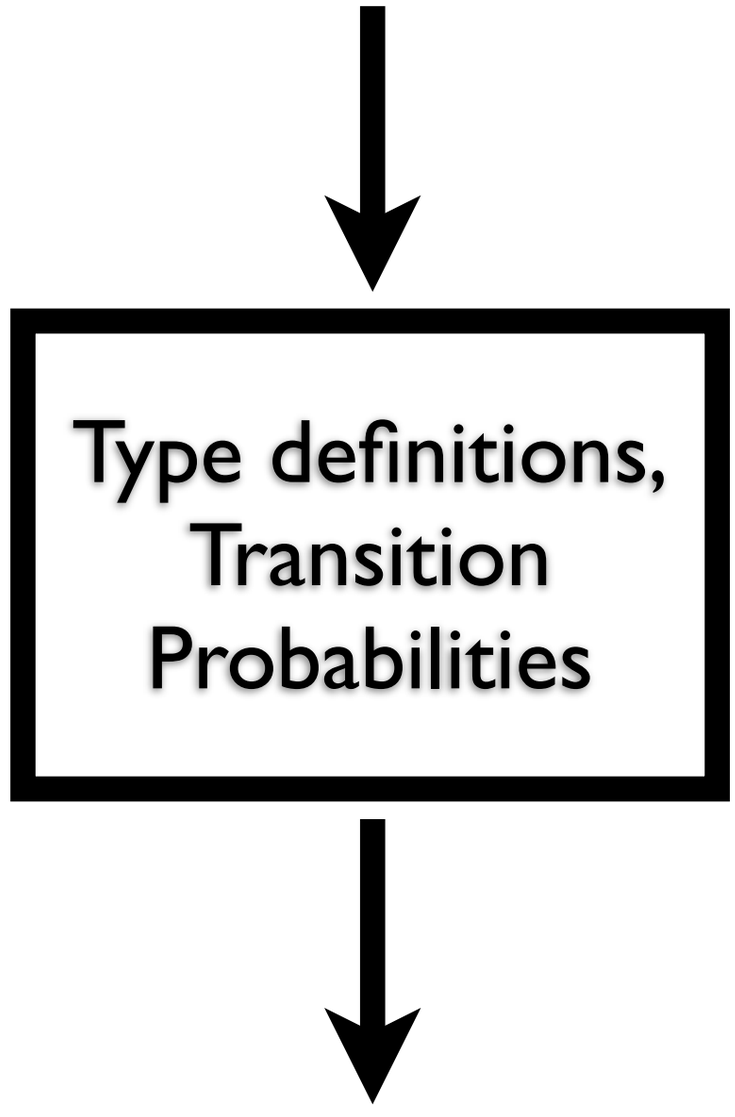
Typological Universals,
Language Clusters

Lexical Data



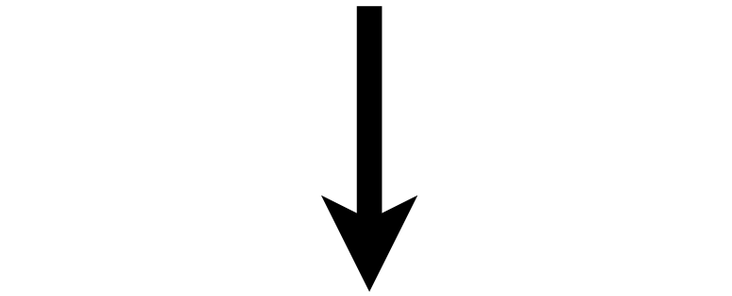
Tree of Languages

Grammatical Data



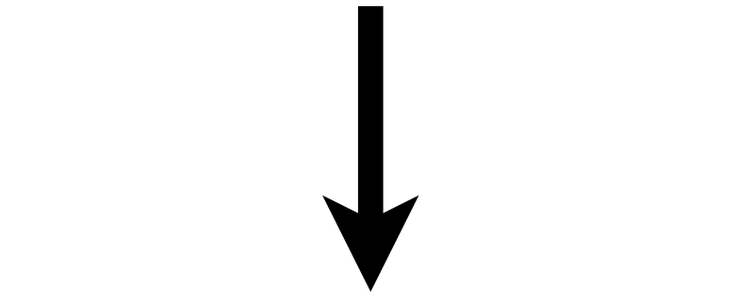
Typological Universals,
Language Clusters

Lexical Data



Tree of Languages

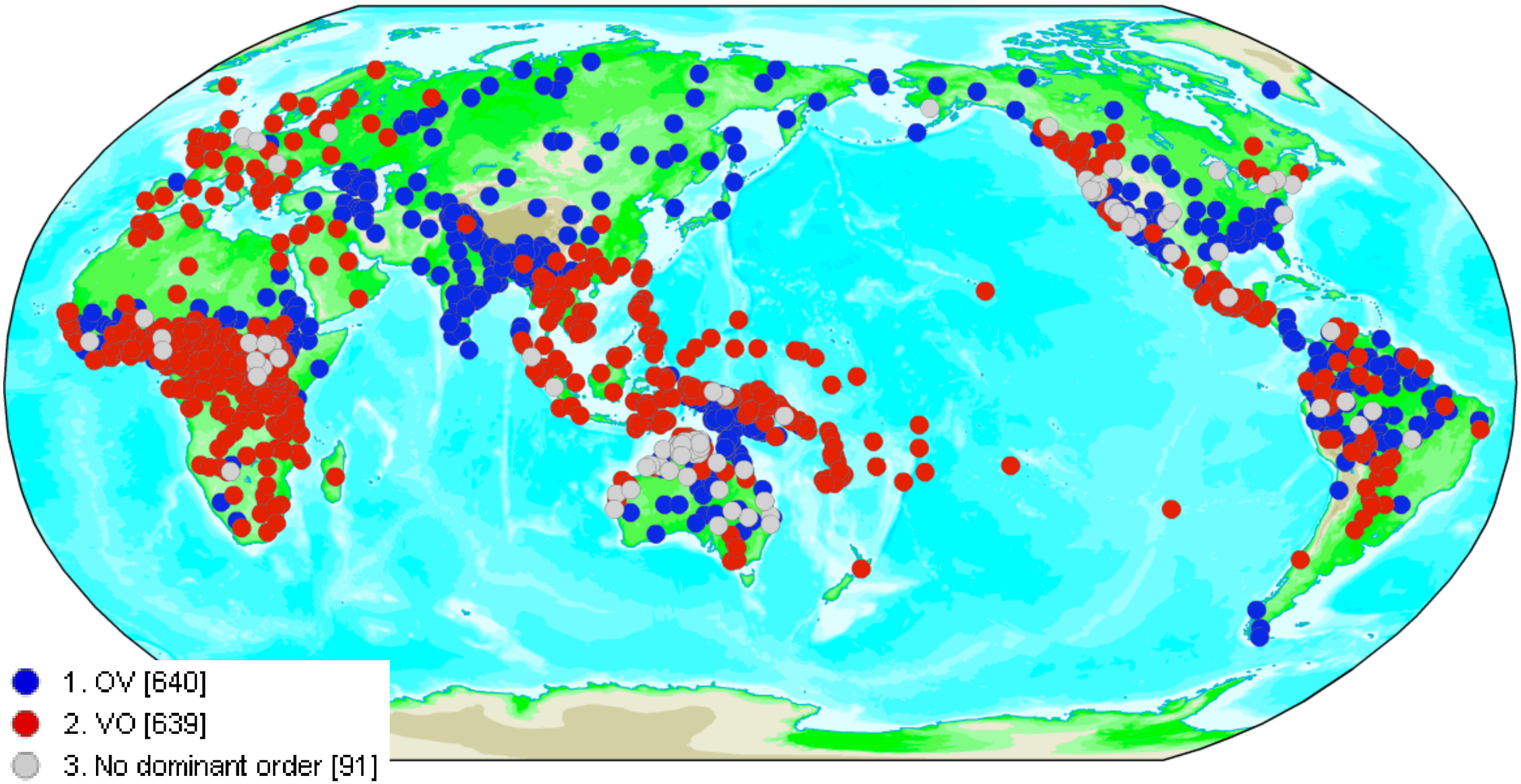
Grammatical Data



Typological Universals,
Language Clusters

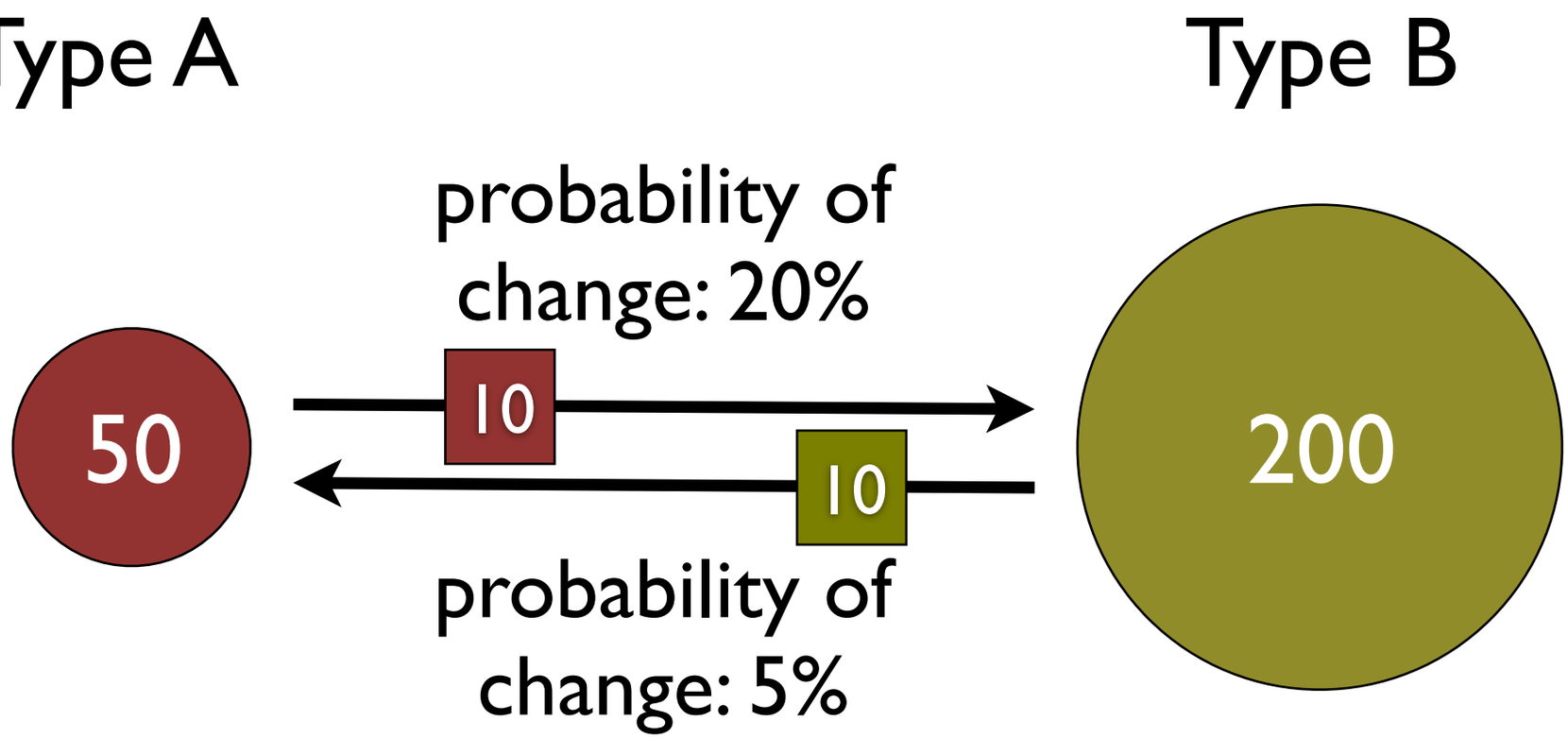
Grammatical Data

Order of Object and Verb

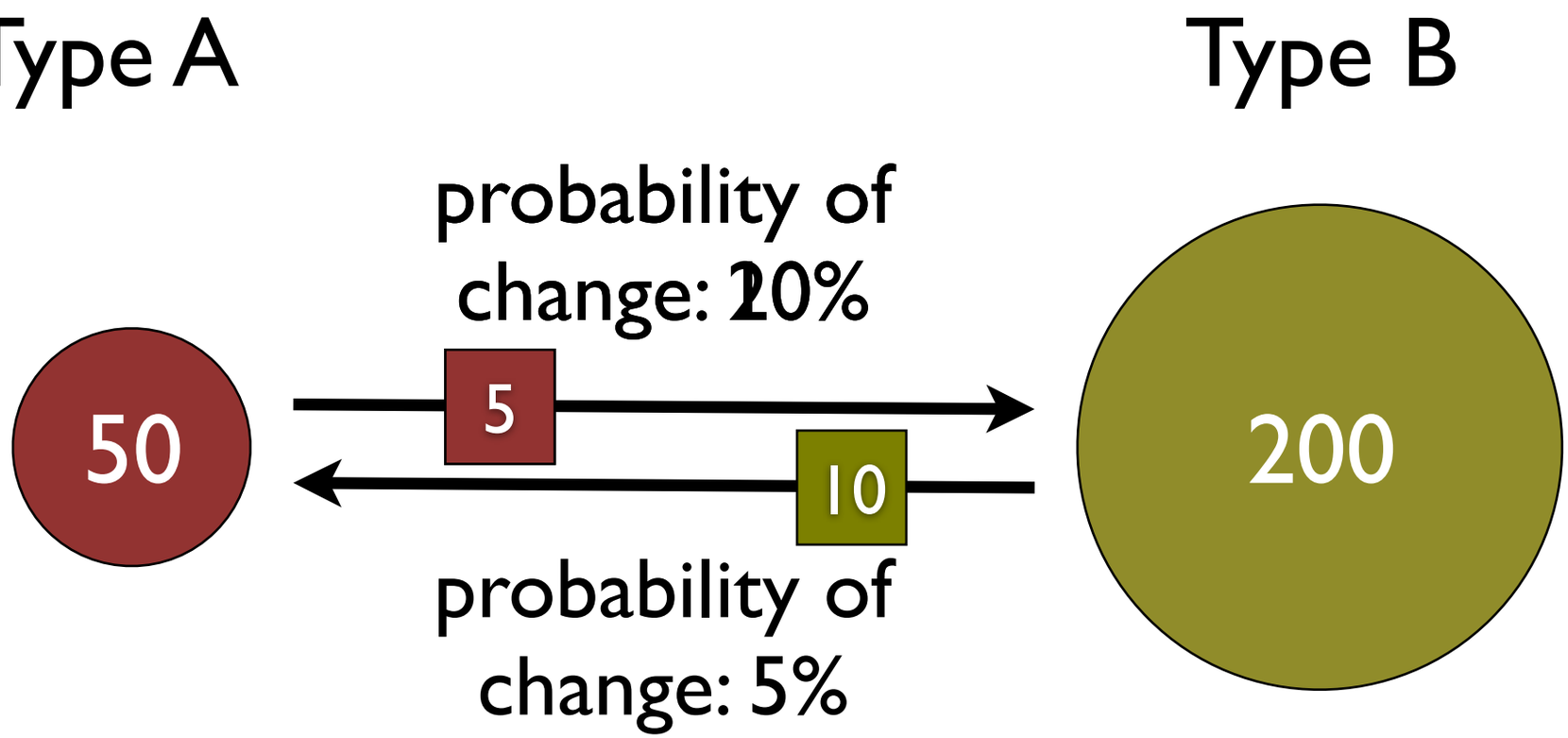


Dynamic Typology

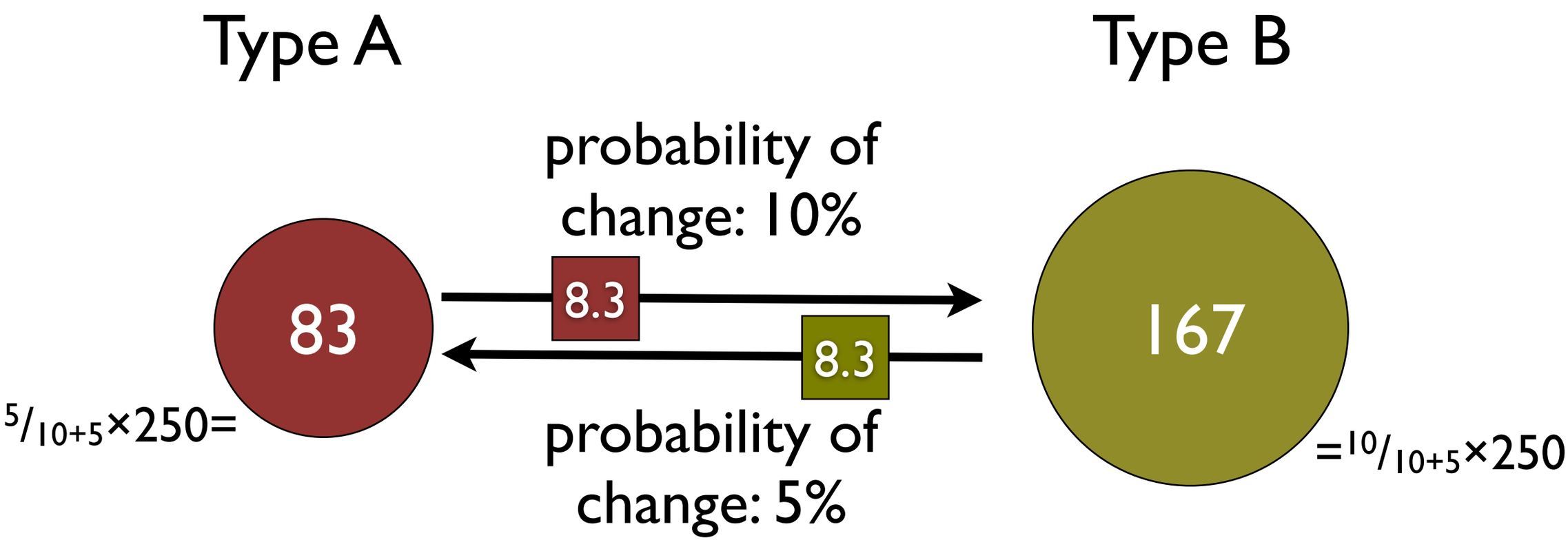
- It is not the **actual frequencies** that matter
- It is the **stable distribution** that matters
  - A stable distribution is a situation in which just as many languages change from **A to B** as change from **B to A**.
  - The extent to which the **actual is different from the stable situation** signals an effect of history



Stable distribution



Instable distribution



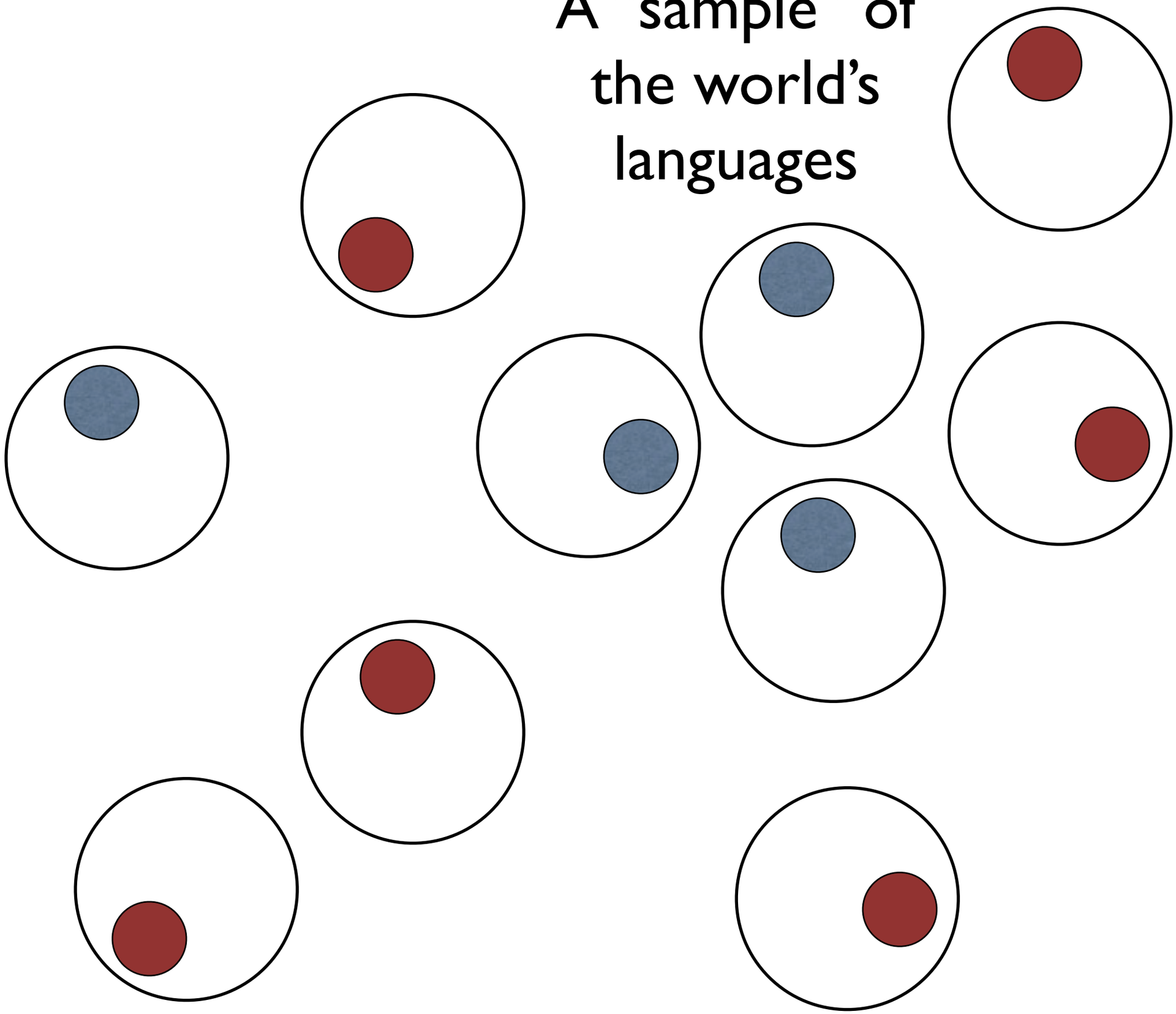
Expected stable distribution

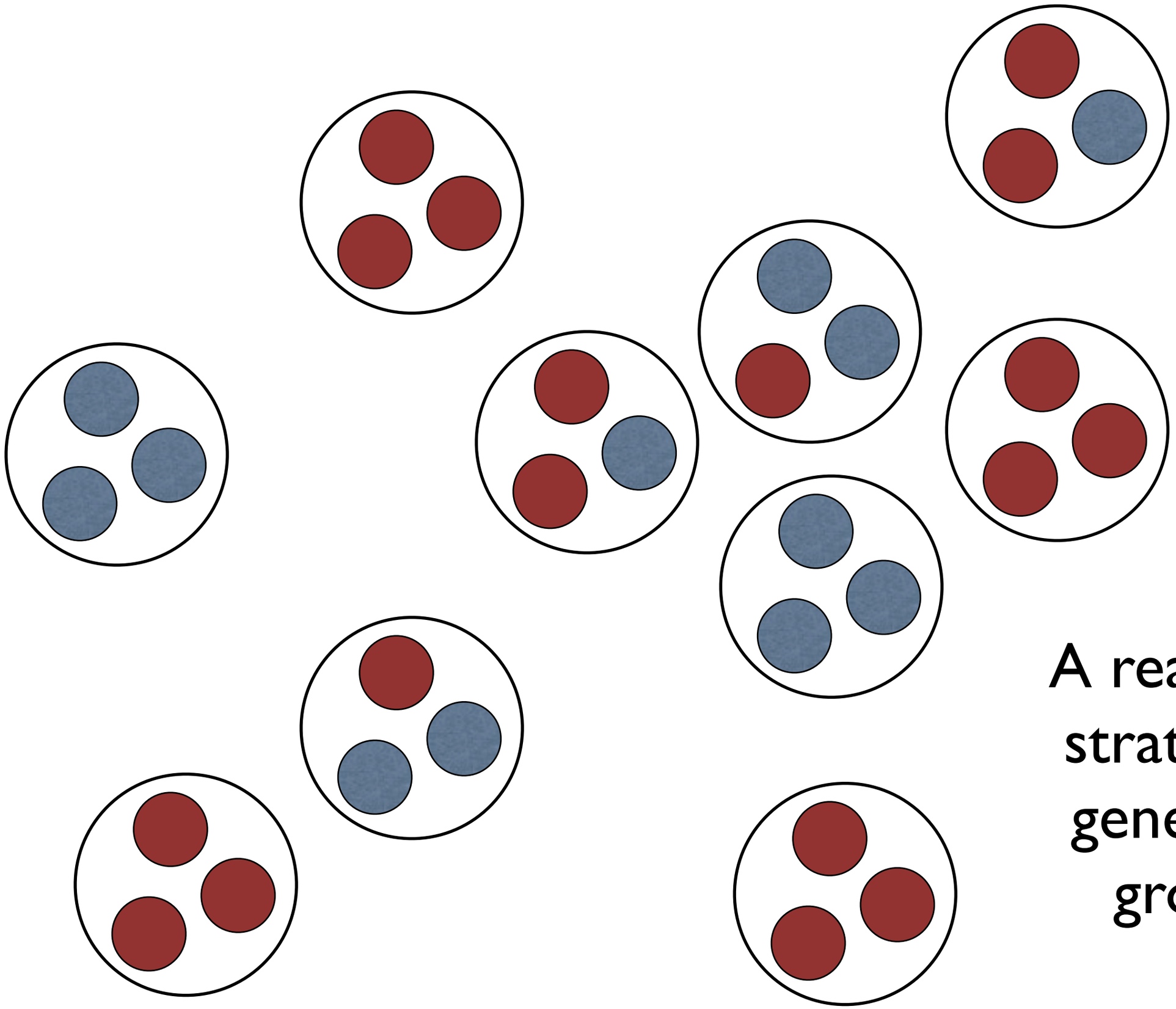
Estimating Transition Probabilities

- Are transitions probabilities **measurable** ?
- If yes:
  use **group internal variation** of groups

Groups of related languages

A "sample" of the world's languages





A real sample stratified for genealogical grouping

Elena Maslova's proposal

probability of **any** change happening = $\alpha \cdot$ frequency (blue) + β

For groups of three languages:

$\alpha = 3 \cdot (p_{\text{blue}\to\text{red}} - p_{\text{red}\to\text{blue}})$

$\beta = 3 \cdot p_{\text{red}\to\text{blue}} \cdot (1 - p_{\text{blue}\to\text{red}})$

LETTER

doi:10.1038/nature09923

Evolved structure of language shows lineage-specific trends in word-order universals

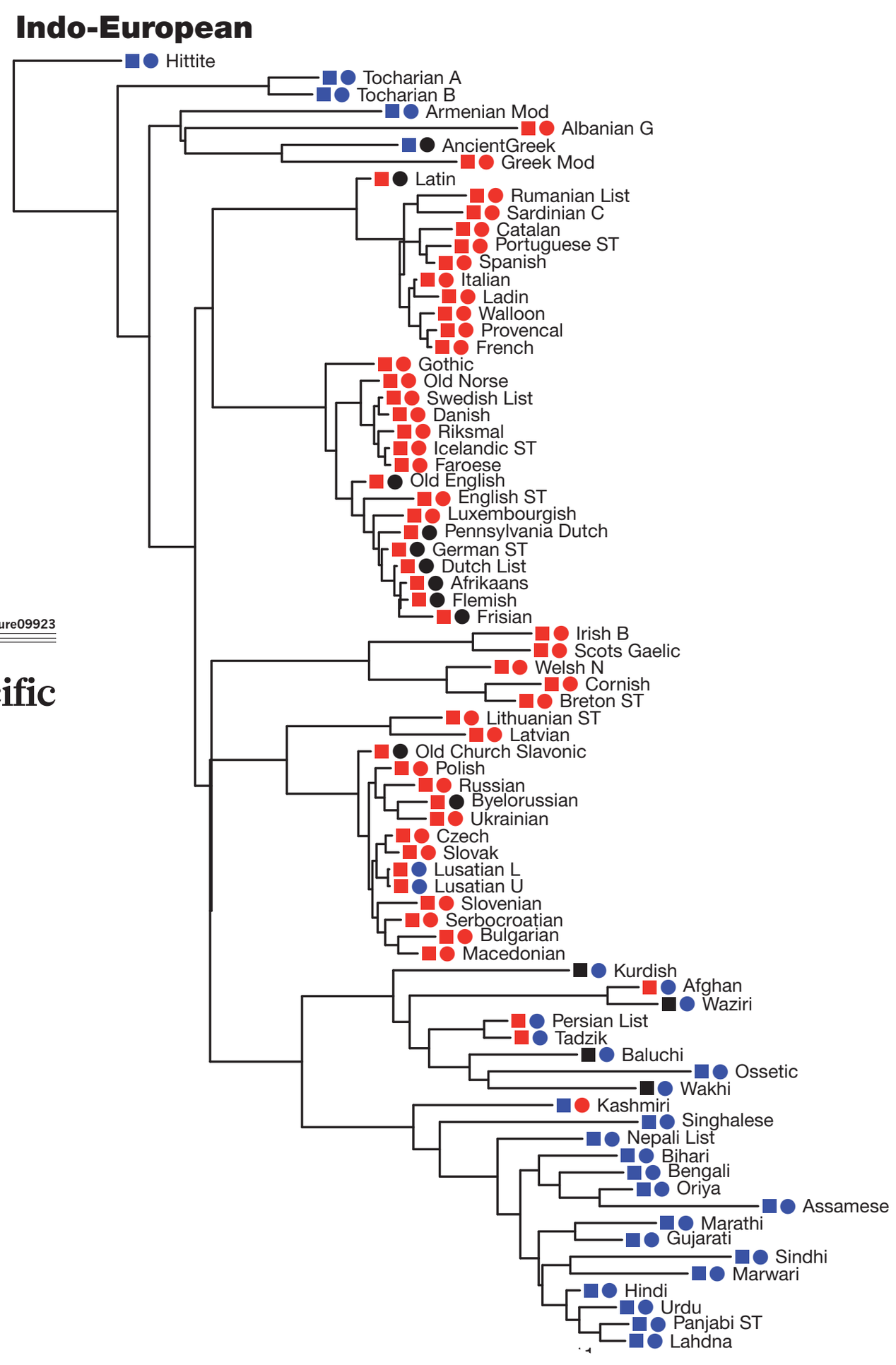
Michael Dunn[1,2], Simon J. Greenhill[3,4], Stephen C. Levinson[1,2] & Russell D. Gray[3]

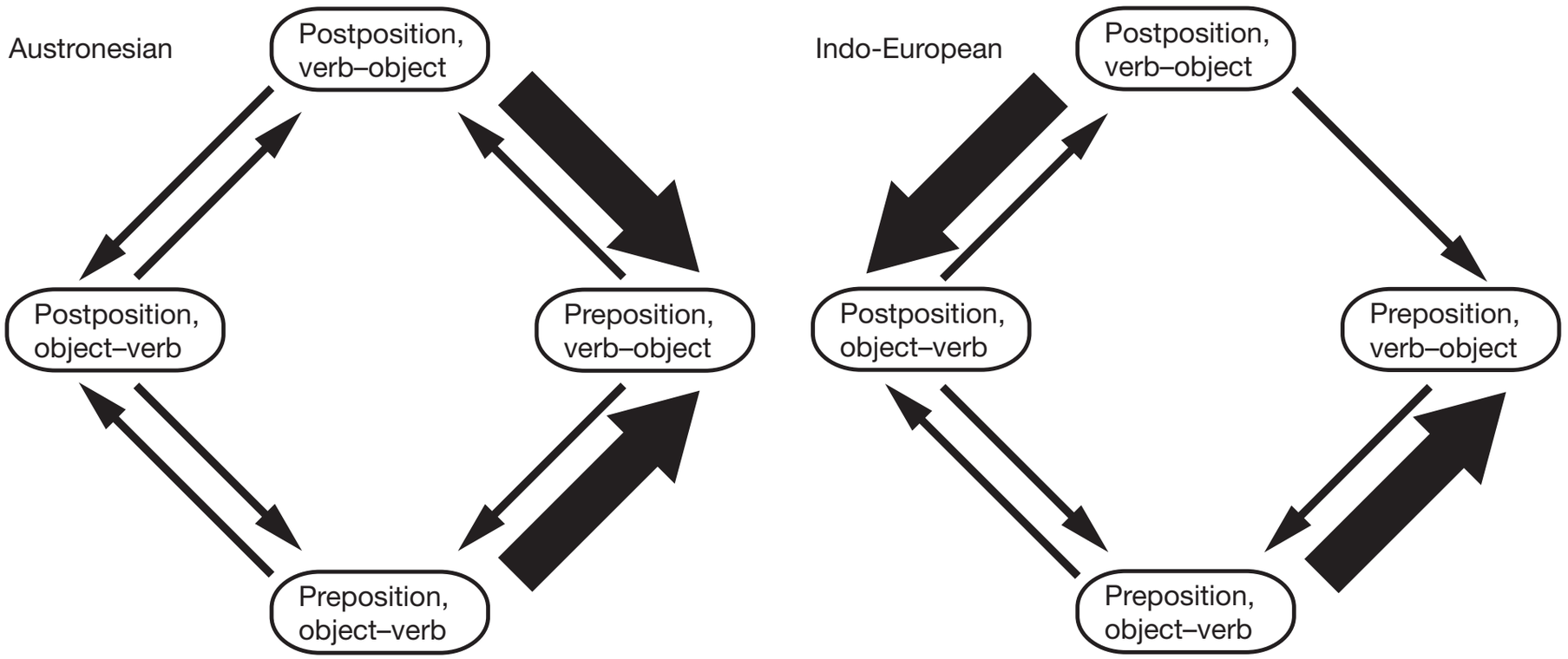
LETTER

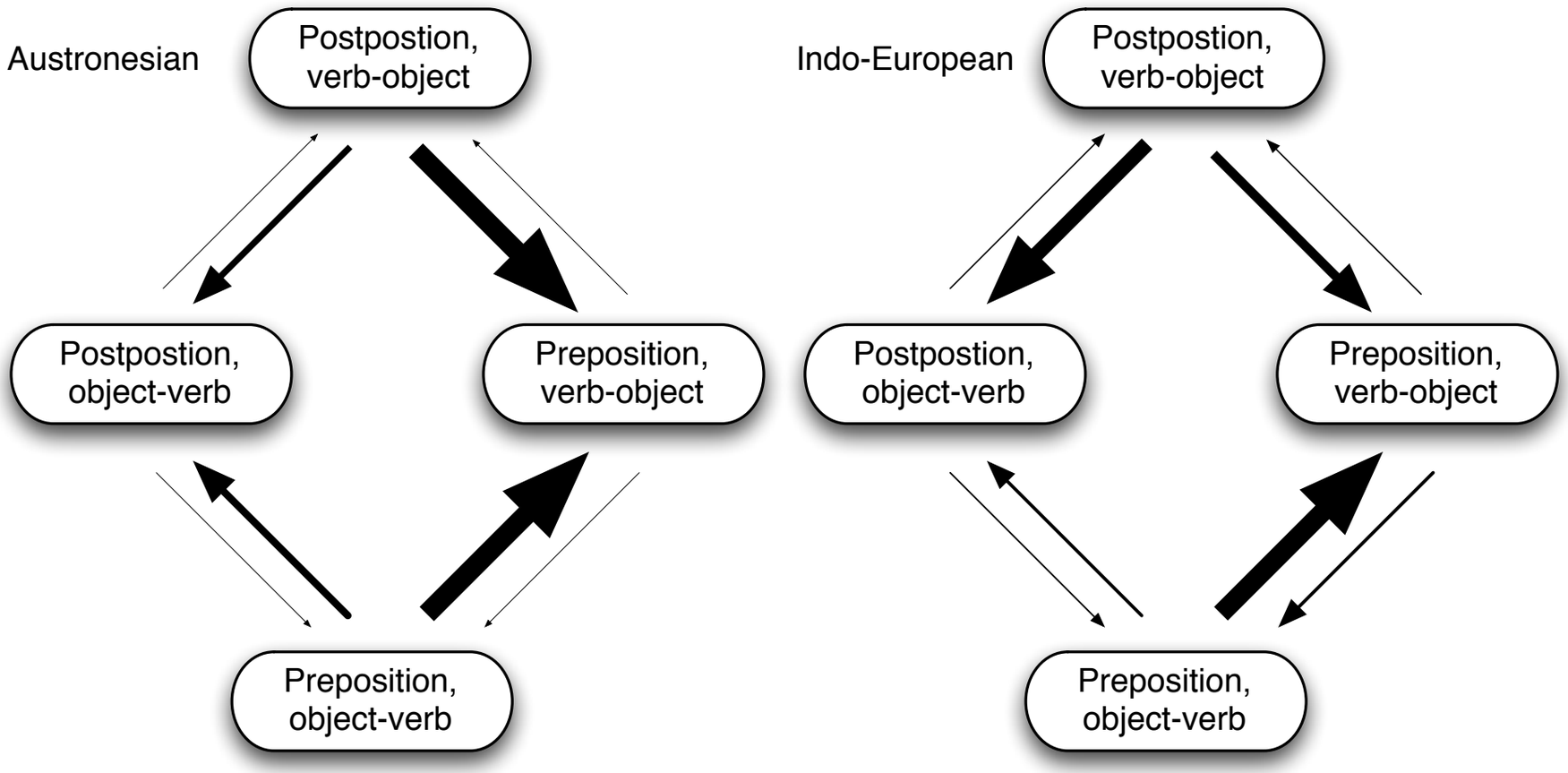
doi:10.1038/nature09923

Evolved structure of language shows lineage-specific trends in word-order universals

Michael Dunn[1,2], Simon J. Greenhill[3,4], Stephen C. Levinson[1,2] & Russell D. Gray[3]




Austronesian
Postposition, verb–object
Postposition, object–verb
Preposition, verb–object
Preposition, object–verb
Indo-European
Postposition, verb–object
Postposition, object–verb
Preposition, verb–object
Preposition, object–verb


Austronesian
Postpostion, verb-object
Postpostion, object-verb
Preposition, verb-object
Preposition, object-verb
Indo-European
Postpostion, verb-object
Postpostion, object-verb
Preposition, verb-object
Preposition, object-verb

Peeking inside the typological black box

- Typological parameters are not static, it is actually possible to approach them dynamically
- The real scholarly debate should be about the validity of estimates of transition probabilities

Lexical Data

What is the goal of lexical comparison ?

- Producing **new trees** or other classifications is only of **limited interest**
- There are only two possible reactions:
  - "We knew that all along"
  - "That tree is wrong"
- More productive are explicit proposals of
  - cognacy
  - sound change
  - meaning change

What is the goal of lexical comparison ?

- Producing **new trees** or other classifications is only of **limited interest**
- There are only two possible reactions:
  - "We knew that all along"
  - "That tree is wrong"
- More productive are explicit proposals of
  - cognacy
  - **sound change**
  - meaning change

Modeling Sound Similarities

- Manually specified
  (Kondrak 2002; Heeringa 2004)
- Hidden Markov Models
  (Ristad & Yianilos 1997; Bhargava & Kondrak 2009)
- Regular multi-alignment
  (Prokić 2010; Steiner, Stadler & Cysouw 2011)
- Bayesian inference
  (Prokić 2010)
- Investigating almost identical words
  (Holman, Brown & Wichmann 2011)

Graphemic Normalization

- Widespread idea:
  "Convert everything into IPA"
- IPA is just another orthography !
  (only approximation of sound)
- Still: sound-based normalization is practical
  (but there are strong differences !)
- But: can we do without ?

Graphemic parsing

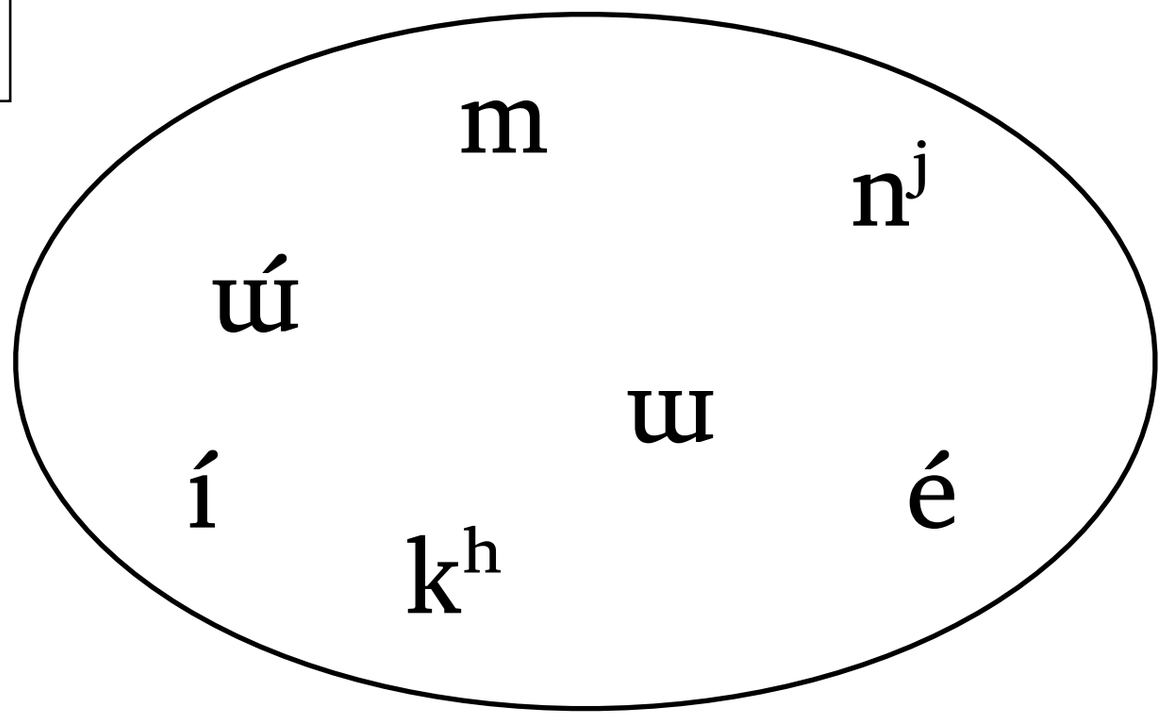
- **Unicode normalization**
  ő vs. o ˜ ´
- **Orthographic parsing**
  (separate orthographic units as used in the source: “graphemes”)
- **Orthographic normalization**
  (research specific!)

Graphemic parsing

- **Code points** (7) t s ʰ o ̃ ́ :
- **Characters** (4) t s^h ő :
- **Graphemes** (2) ts^h ő:

‘bag of symbol” approach

mínʲéékʰɯɯ́



(1-grams)

	Bora	**Muinane**
down	ʧʃɨ́nʲe, paári	báari, gíino
bee	íímɯ́ʔóexpʰi, téʔtsʰipa	níɨ́bɨri, míɨbɨriʔɨ
sharp	tsʰɯ́ʔxɨβáne	síɨ́xéβano
…	…	…

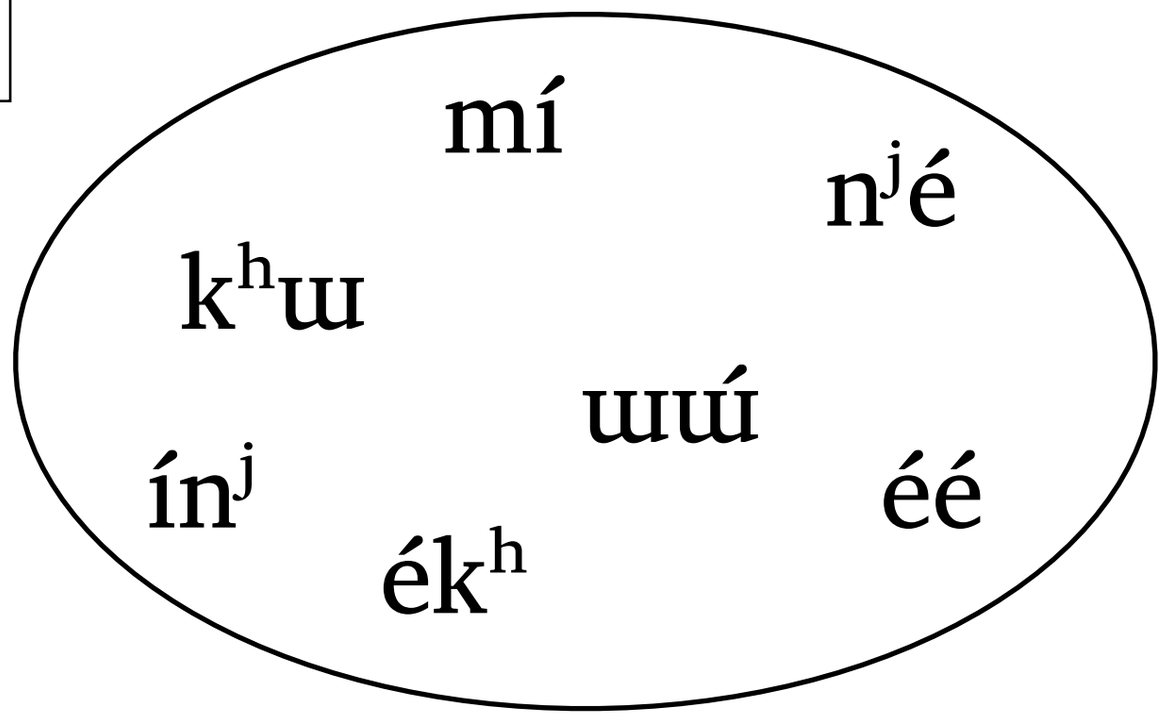
Cross-script mapping

E	R	freq	dice
r	р	184	0.88874745
n	н	115	0.8461936
l	л	104	0.79646295
s	с	114	0.7927922
t	т	165	0.7701921
m	м	47	0.7699933
o	о	184	0.7510106
k	ть	21	0.74458015
p	п	50	0.7388723
i	и	102	0.7034591
a	а	221	0.6866478
u	у	40	0.6449104
c	к	77	0.6251676
e	е	219	0.59066784
b	б	32	0.525643
w	в	46	0.46787763
d	д	42	0.381996
⋮	⋮	⋮	⋮

Cysouw, Michael & Hagen Jung. 2007. Cognate identification and alignment using practical orthographies. Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology, 109-116.

‘bag of symbol” approach

mínʲéékʰɯɯ́



(2-grams)

Bora	Muinane	Bora	Muinane	Bora	Muinane
#k	#k^h	#i	#i	#n	#n
kɨ	k^hɯ	#a	#a	#m	#m
se	ts^hɨ	di	ti	mɨ	mɯ
xe	xɨ	du	to	nɨ	nɯ
ga	k^wa	#d	#t	us	ts^hɨ
ba	pa	#s	#ts^h	#t	#t^h
#b	#p	gi	ʧi	ɨg	ɯk^w
e#	ɨ#	ni	ni	#ɸ	#p^h

Using bigram matching

- Bora ‘two’: mínʲéékʰɯɯ́
- Muinane ‘two’: míínokɨ

	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m	22	3	2	2	2	2	2	2
mi	4	12	2	2	5	1	1	1
in^j	2	1	5	9	3	1	1	2
n^je	1	1	5	5	4	1	1	2
ee	3	3	3	3	6	2	2	2
ek^h	1	2	1	1	4	2	3	2
$k^hɯ$	2	2	2	2	2	1	23	2
ɯɯ	2	2	3	3	2	2	4	4
ɯ#	2	2	3	2	3	1	3	4

	#m	mi	ii	in	no	ok	kɨ	ɨ#
#m	22	3	2	2	2	2	2	2
mi	4	12	2	2	5	1	1	1
in^{j}	2	1	5	9	3	1	1	2
n^{j}e	1	1	5	5	4	1	1	2
ee	3	3	3	3	6	2	2	2
ek^{h}	1	2	1	1	4	2	3	2
k^{h}ɯ	2	2	2	2	2	1	23	2
ɯɯ	2	2	3	3	2	2	4	4
ɯ#	2	2	3	2	3	1	3	4

	Bora	**Muinane**
down	ʧʃɨ́nʲe, paári	báari, gíino
bee	íímɯ́ʔóexpʰi, téʔtsʰipa	níɨ́bɨri, míɨbɨriʔɨ
sharp	tsʰɯ́ʔxɨβáne	síɨ́xéβano
…	…	…

	Bora	**Muinane**
down	tʃíɲʲe, **paári**	**báari**, gíino
bee	íímɯ́ʔóexpʰi, téʔtsʰipa	nɨ́ɨ́biɨri, mɨ́ɨbiɨriʔɨ
sharp	**tsʰɯ́ʔxɨβáne**	**sɨ́ɨ́xéβano**
…	…	…

Inside the lexical black box

- Grapheme correspondences are relatively easy to approximate
- Use them to propose hypotheses about cognacy and sound correspondences
- Cognacy and sound correspondences can be fruitfully discussed