Universität Zürich[UZH]

Department of Comparative Linguistics

Moving beyond Pāṇini: causal theories in linguistics

Balthasar Bickel

a VERY brief history of linguistics

or: why linguistics has a problem with causal theories

The origin of grammatical analysis

Pāṇini's *Aṣṭādhyāyī* (*fl.* 4th c. BCE)

3,959 rules of Sanskrit

An example:

> "2.3.1 if not already expressed,
>
> > 2.3.2 for goal: case 2 (ACC)
> >
> > 2.3.46 for gender and number only (i.e. no role specs): case 1 (NOM)
> >
> > 3.4.69 for agent, goal or intransitive: *laḥ* (finite verb endings)"

We get can accusative on goals *because* it's the law.

Translation based on Kiparsky 2002, *On the Architecture of Panini's Grammar*

Pāṇini's *Aṣṭādhyāyī* (*fl.* 4th c. BCE)

२.१.१ समर्थः पदविधिः ।
२.१.२ सुबामन्त्रिते पराङ्गवत् स्वरे ।
२.१.३ प्राक् कडारात् समासः ।
२.१.४ सह सुपा ।
२.१.५ अव्ययीभावः ।
२.१.६ अव्ययं विभक्तिसमीपसमृद्धि-
व्यृद्ध्यर्थाभावात्ययासम्प्रति-
शब्दप्रादुर्भावपश्चाद्यथाऽऽनुपूर्व्ययौगपद्यसादृश्य-
सम्पत्तिसाकल्यान्तवचनेषु ।
२.१.७ यथाऽसादृये ।
२.१.८ यावदवधारणे ।
२.१.९ सुप्प्रतिना मात्राऽर्थे ।
२.१.१० अक्षशलाकासङ्ख्याः परिणा ।
२.१.११ विभाषा ।
२.१.१२ अपपरिबहिरञ्चवः पञ्चम्या ।
२.१.१३ आङ् मर्यादाऽभिविध्योः ।

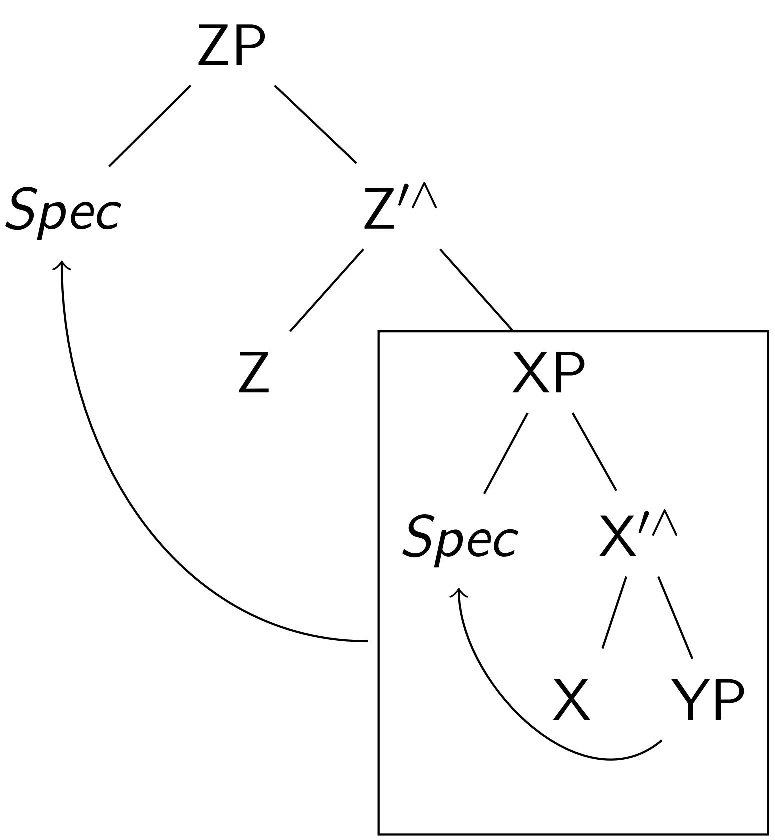
Linguistics is engineering, even now

- Formulate the most concise, most parsimonious, most elegant description, like Pāṇini!
- Mostly a goal in itself: “pure linguistics” (Lazard 2012*)
- But perhaps not so interesting for other disciplines:
  - The most elegant and concise description may not capture
    - the generalizations by which children learn
    - the components that fit with the phylogeny of language
    - the units that brains process
- Still, linguists adopt the Pāṇinian style even for cross-linguistic work...

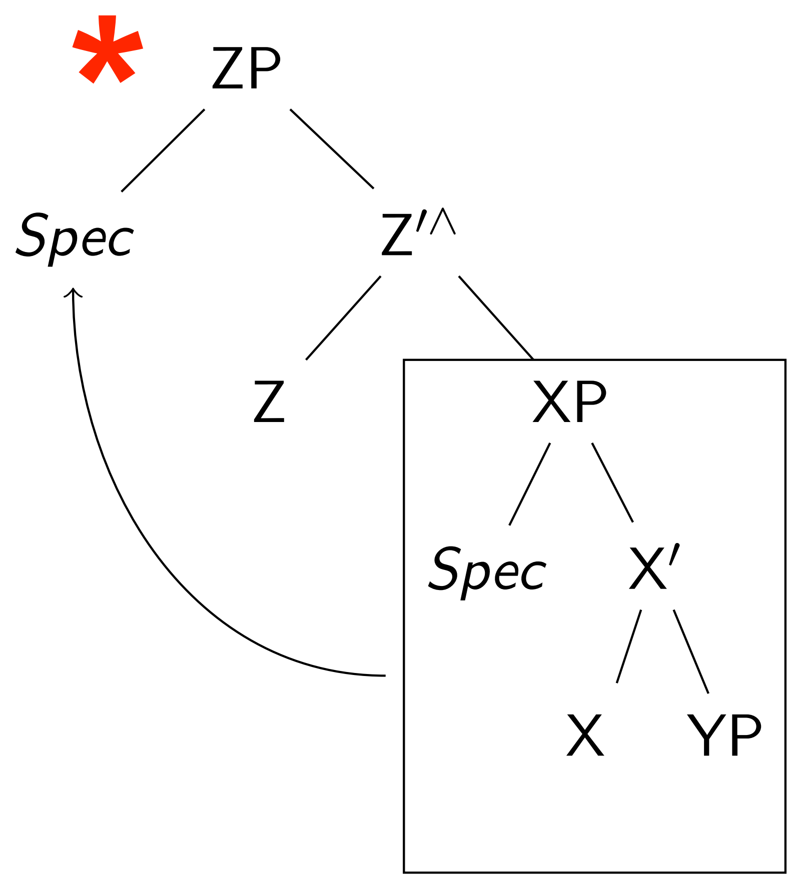
***Studies in Language*

Pāṇinian Thinking in Comparative Linguistics, Typology

- Fomulate a law and explain away any counter-examples!
- And so **the law causes the facts**!
- Illustration: The Final-Over-Final-Constraint (a modern version of Greenberg Universal #2; Biberauer et al. 2014*)



e.g. [$_{PP}$ [$_{NP}$ YP N] P]



e.g. *[$_{PP}$ [$_{NP}$ N YP] P]

*Linguistic Inquiry

Pāṇinian Thinking in Comparative Linguistics, Typology

- *Counterexample* in Harar Oromo (Kushitic, Owens 1985)

[$_{\text{PP}}$ [$_{\text{NP}}$ *maná*	[$_{\text{NP}}$ *obbolesá*	*xiyyá*]]	=*tt*]
house	brother	my	in
N	NP		P

- *Solution:* Explain the example away, e.g. limit the FOFC to complements with the same category features (Biberauer et al. 2014*) and argue that Oromo postpositions are [+V], or indeed not postposition at all.

**Linguistic Inquiry*

Why not?

- Nothing is guaranteed to be exceptionless, not even “exceptionless (p<.05)” (Piantadosi & Gibson 2014[*])
- No idea what survived the human population bottlenecks 20-60*kya*!
- So pick generalizations that are justified (Chomsky 1964ff), but this leaves us in the end perhaps only with very abstract generalizations like
    - simple composition (α & β), as shared with other species (e.g. mongooses, Janssen et al. 2012[+])
    - supra-regularity, as shared with other cognitive domains (e.g. action, Fitch 2014[†])
    - recursion, as shared with other species when limited to regular grammars (e.g. Tamarin monkeys; Fitch & Hauser 2004[‡])
    - asymmetry (categories), as shared with other species (e.g. Campbell monkeys; Ouattara et al. 2009[§])

[*] *Cogn Sci*, [+] *BMC Biol*, [†] *Phys Life Rev*, [‡] *Science*, [§] *PLOS One*

A cheap way out

- Plough through databases, find soft constraints (correlations). Then explain them *post hoc*...
- **but this is the very problem that brings us here!**
  - sample?
  - missing data
  - unclear stochastic process
  - **causality?**

Perhaps after nearly 2500 years,

it's time to move on!

A more expensive way out: a normal science approach

- **How is the (evolutionary, diachronic, ontogenetic) development of specific parts of languages *caused* by the natural and social ecology of language?**
- For this, we need:

(1) **Theories** on how natural and social conditions causes specific patterns in language evolution, change and development so that structures end up with the distributions we observe

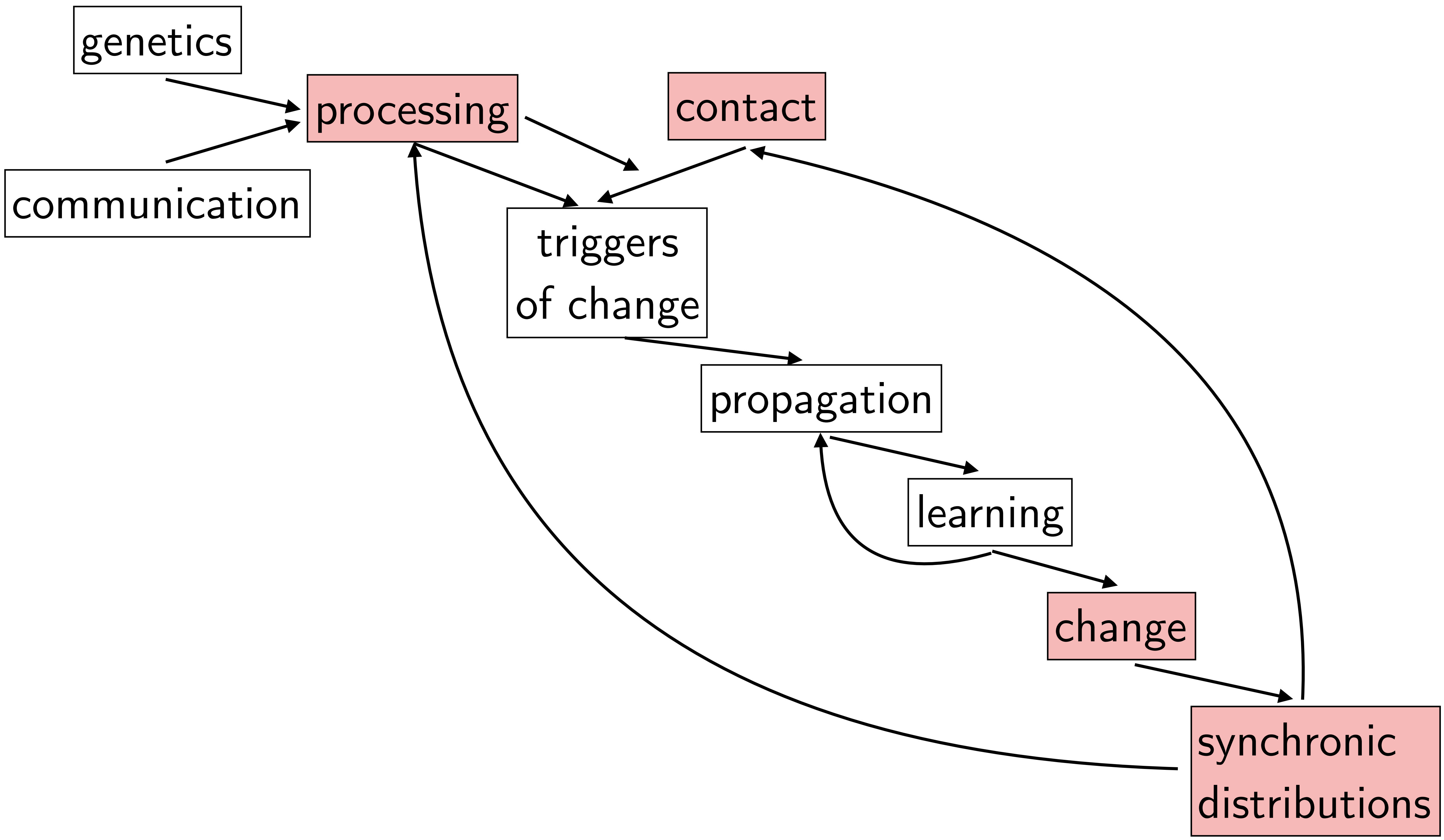
(2) Fine-grained variables for **measuring** these distributions. Adequate iff

- descriptively correct
- cross-linguistically applicable
- in sync with what we know about processing, acquisition

(3) Statistical **models** for testing (1) against (2)

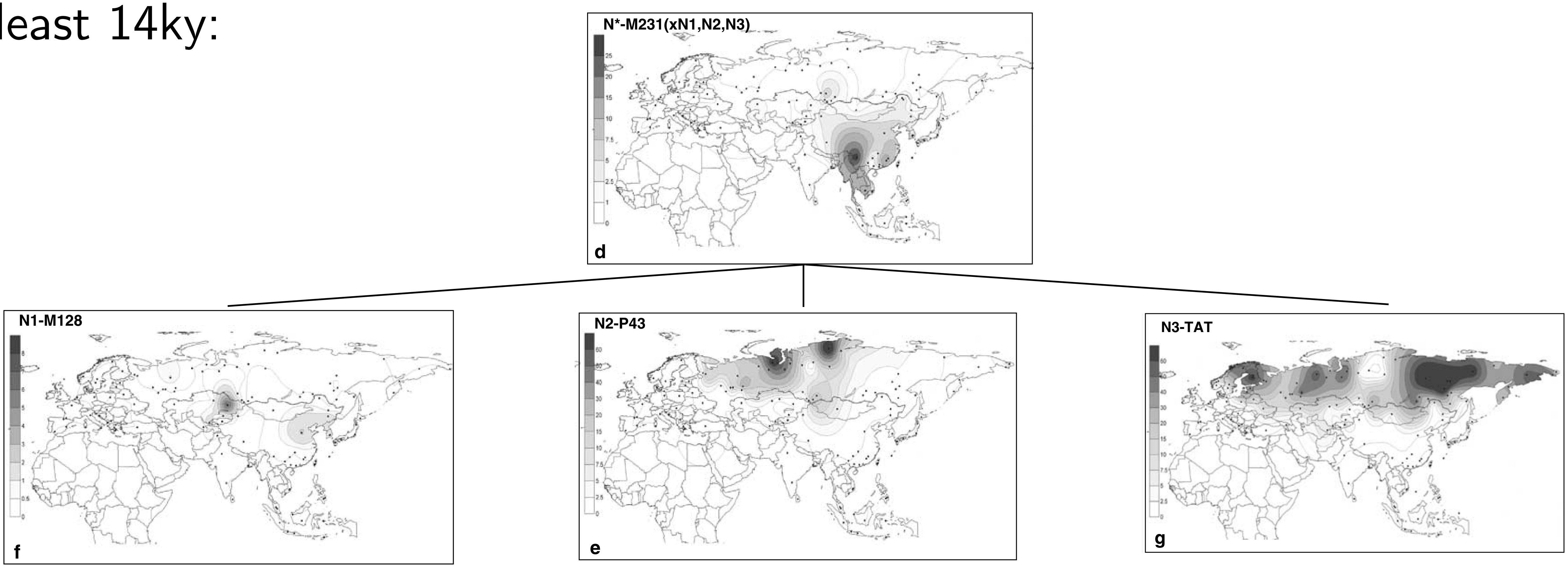
Theories

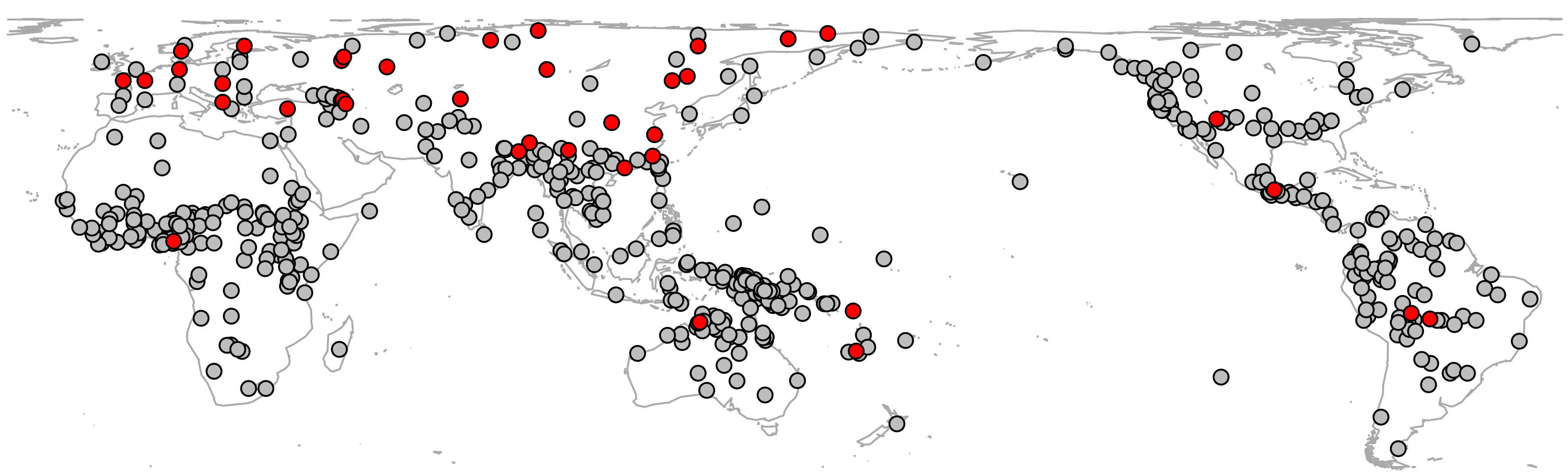
- General framework (cf. talks by Dan Dediu, Morten Christiansen, Florian Jaeger, Jasmeen Kanwal, Christian Bentz)



Causal theories — some examples

- **Event-based theories:** contact effects limited to concrete, *localized and historical* events, with no functional motivation, e.g. events in Eurasia in the least 14ky:





/y/

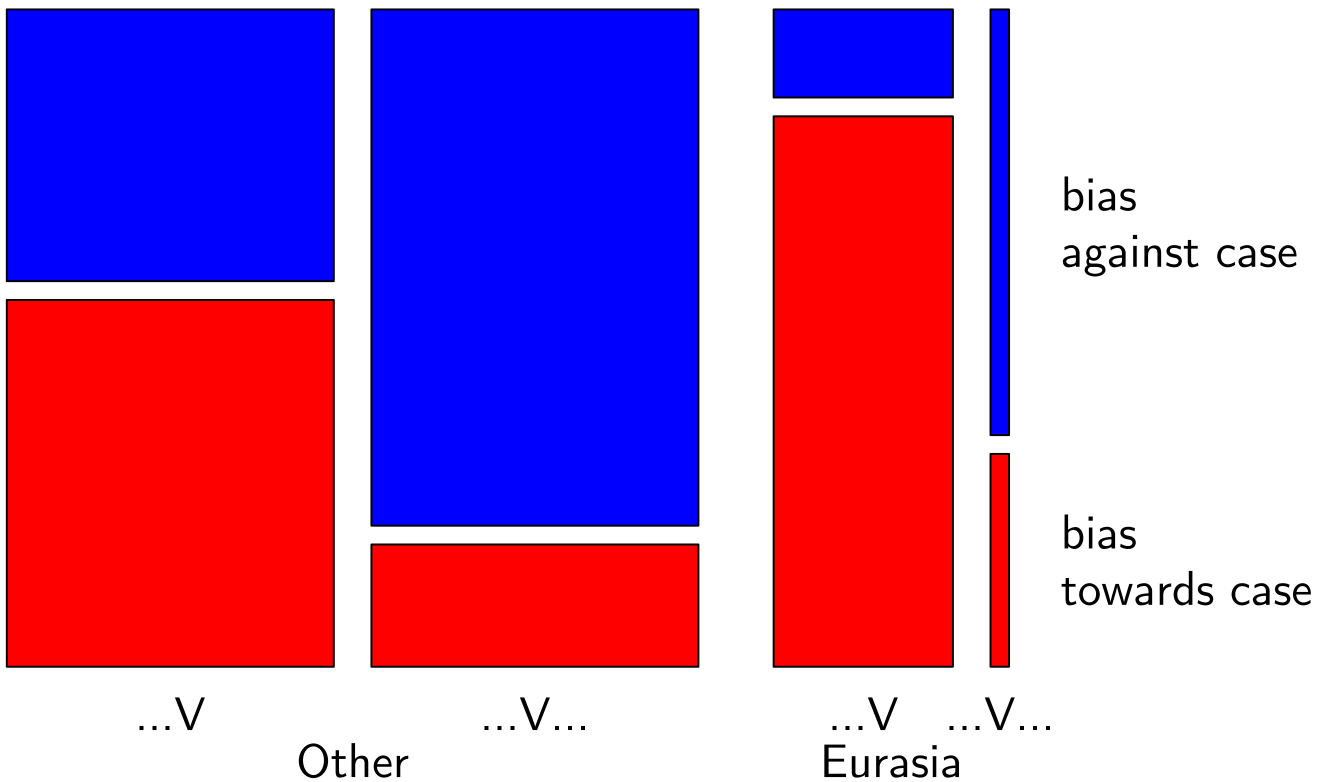
Causal theories — some examples

- **Functional theories:** processing and communication principles cause certain directions in language change, e.g.
  - High cost of voicing in word-final position favors development and maintenance of final devoicing (Blevins 2004[*])
  - Low humidity disfavors development and maintenance of rich tonal distinctions (Everett et al. 2015[+]; also Coupé's talk)
  - Signal transmission in verb-final structures is safer with case makers (Hall et al. 2013[†], Gibson et al. 2013[‡])
  - Informative communication prefers certain lexical patterns (Regier's talk)
  - Priming trends cause differences in NP frequency (Bickel 2003[§])
  - *Perhaps*: supra-regular computation favors the development and maintenance of embedded phrase structures ("*Dendrophilia*", Fitch 2014[#])

[*]*Evol Phon*, [+]*PNAS*, [†]*Cognition*, [‡]*Psych Sci*, [§]*Language*, [#]*Phys of Life Rev*

Signals may be weak

- Causes trigger *possible* change, but actualization requires many opportunities for change (many speakers, many generations) because:
  - uncertainty of social propagation (but once there, we get amplification through feedback loop in the next generation; cf Dediu's talk)
  - competing forces: e.g. contact events can enhance or suppress a principled trigger of change



- In fact, a causal trigger must not be too strong: it might harm communication and acquisition!

Methodological challenge

- must pick up signals of change: diachronic transition probabilities (Maslova 2000 etc.)
- even when languages don't belong to a family (44-47% of all families have only 1 known member*)

*AUTOTYP, GLOTTOLOG

Traditional approaches

- Family relations are a confound (Galton's Problem, Simpson's Paradox), so control for them by...:
  - strategic sampling (Dryer 1989[*]), or re-sampling (Everett et al. 2015[+])
  - modeling them as fixed (Dediu & Ladd 2007[†], Bickel et al. 2009[‡]) or random (Jaeger et al. 2011[§], Bentz & Winter 2013[#]) factors

- but...
  - even after controling for confounds,
  - synchronic frequency estimates $\nRightarrow$ transition probabilities:
    - the process may not have reached stationarity (Maslova 2000[¶])
    - indeed sometimes has not reached stationarity (Cysouw 2011[‖]),
    - especially when it is driven by local contact events!

[*]*Stud. Lang*, [+]*PNAS*, [†]*PNAS*, [‡]*Phon. Domains*, [§]*Ling Typ*, [#]*Lang Dyn Change*, [¶]*Ling Typ*, [‖]*Ling Typ*

and more problems..

- also, shared inheritance or parallel development within a family can be the very signal we seek to pick up!
- E.g. DOM in Romance (e.g. Spanish a, Romanian pe) or Indo-Iranian (e.g. Hindi -ko, Nepali -$l\bar{a}i$, Persian $r\hat{a}$)

The Family Bias Method (or the Family of Family Bias Methods)

Core ideas:

1. Families are not a confound but demonstrated families are the very basis on which we can estimate transition probabilities (Greenberg 1978[*], Maslova 2000[+] etc.)

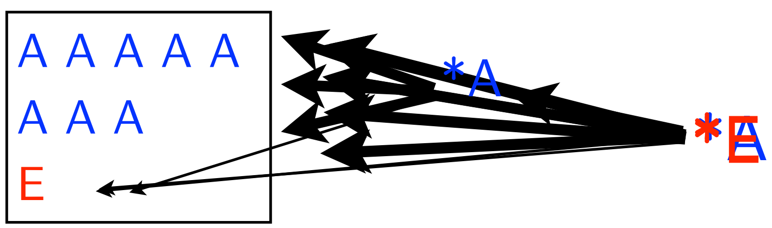
   → estimate difference in transition probabilities, eg. $P(\text{A>B}) > P(\text{A<B})$: **"family biases"**

2. We can estimate family biases even for isolates and small families via extrapolation (Bickel 2013[§])

[*]*Univ. of Lang.* 1, [+]*Ling Typ*,, [§]*Ling Typ Hist Cont*

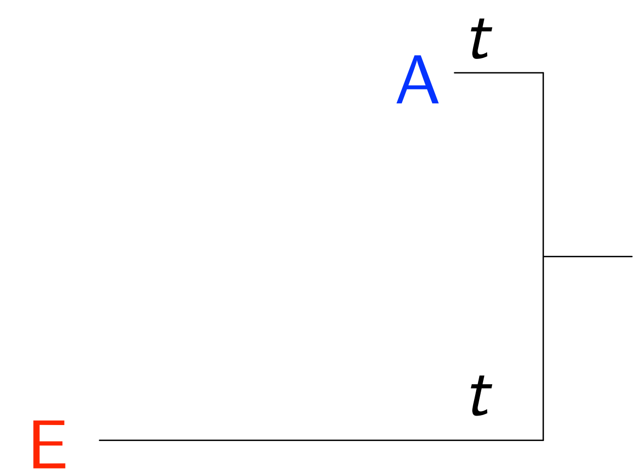
Step 1: estimating family biases in sufficiently large families

Set-based approach:

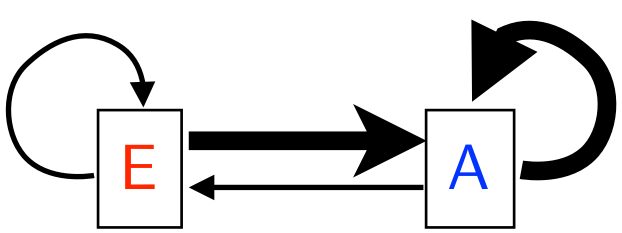


- Infer a family bias if A "dominates", using e.g. a binomial test. (If nothing dominates, we don't know.)

Tree-based approach:



- Estimate the best-fitting transition rate matrix Q in a Continous-Time Markov chain
- Infer a family bias if qAE$\neq q$EA fits the data better than qAE$=q$EA (LR or BF)



Step 1: estimating family biases in sufficiently large families

Assumptions

	set-based	*tree-based*
family model	tree, wave, linkage, network	tree (strict)
stochastic process of diachronic event	independent multinomial trial	Continuous-Time Markov or Wiener process
data requirement	none	non-constant
family requirement	none	topology; branch lengths*

*e.g. length 1 between each node, assuming that anagenetic change in, say, the lexicon, is irrelevant for type change, especially if caused by contact (Thomason & Kaufman 1988)

Step 2: estimate bias probabilities behind small families and isolates

- Use the mean probability of bias in large families for estimating the *probability that a small family is what survives of a large family with a bias* (in whatever direction). E.g. Laplace estimates on biases with 95%CI:

Africa	Eurasia	Pacific	N/C America	S America
.92 (.75,1)	.75 (.48, .94)	.5 (.27,.73)	.88 (.59,1)	.5 (.15,.85)

- if estimated to be biased, estimate direction of bias value (e.g. E) based on what they have, allowing for deviations with a probability based on deviations in large families, and resolving ties at random, e.g.

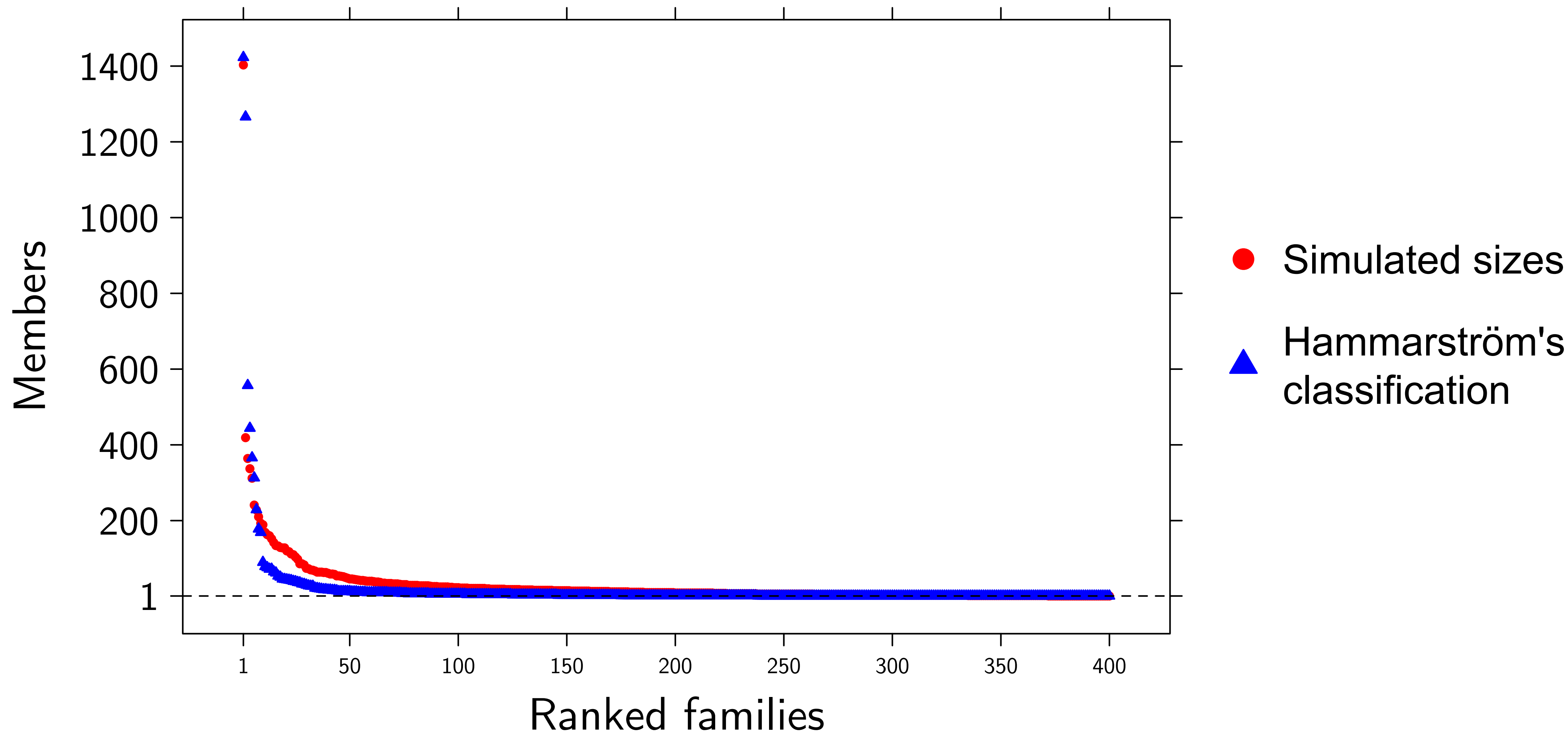
	Africa	Eurasia	Pacific	N/C America	S America
AUTOTYP	.0	.027	.034	.0002	0.01

- take the mean across many extrapolations (e.g. 10,000)

Performance of methods in simulations (preliminary!)

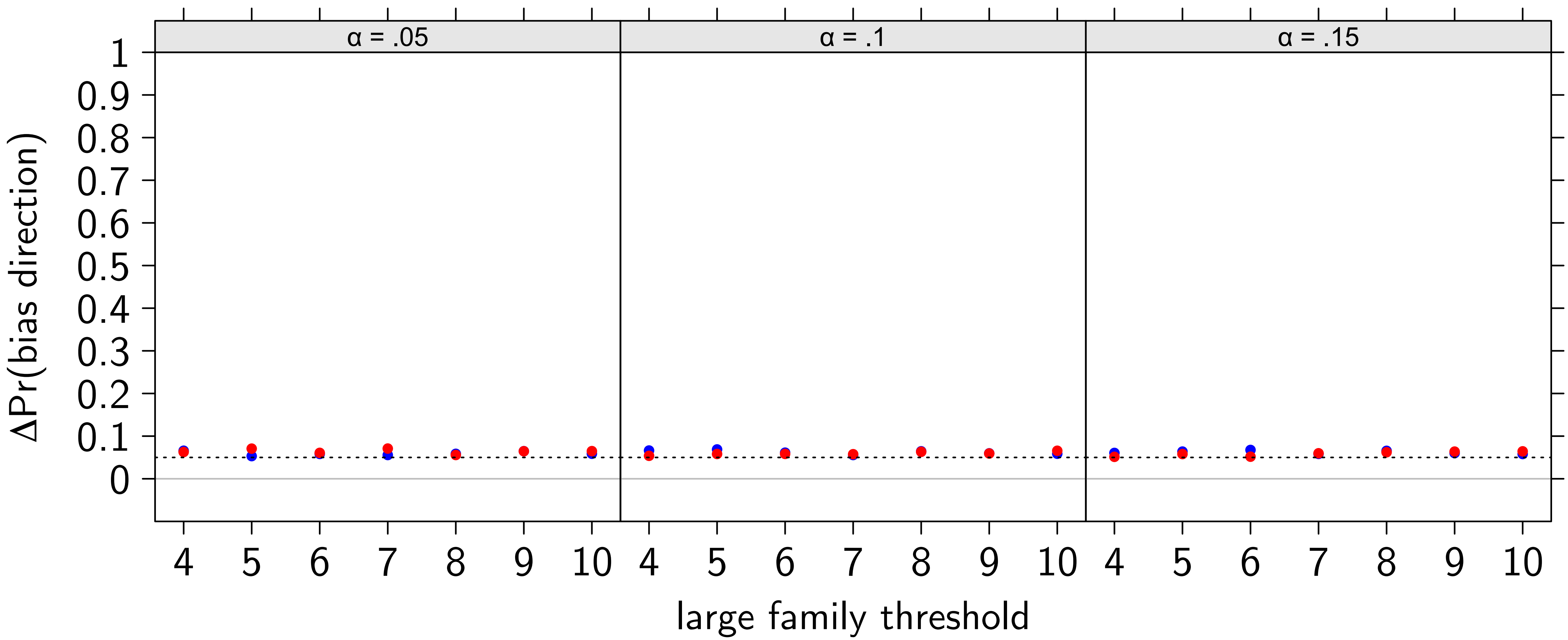
Simulation of a discrete-time Markov process, where language varieties can (within steps of ca. 100 years $\sim$ 3 generations)

- *give birth*: Poisson process with birth rate $\lambda = [.7, .8]$
- *die or stay alive:* Bernoulli process with survival prob. $\pi = [.1, .2]$



Performance of methods in simulations (preliminary!)

- add a binomial variable with a family bias
- and see what we can recover, varying the definition of 'small family' and the rejection level of binomial test for inferring a bias in a family:



Mean Pr(bias direction) estimated lower than built in

Mean Pr(bias direction) estimated higher than built in

So, we have framework and a method

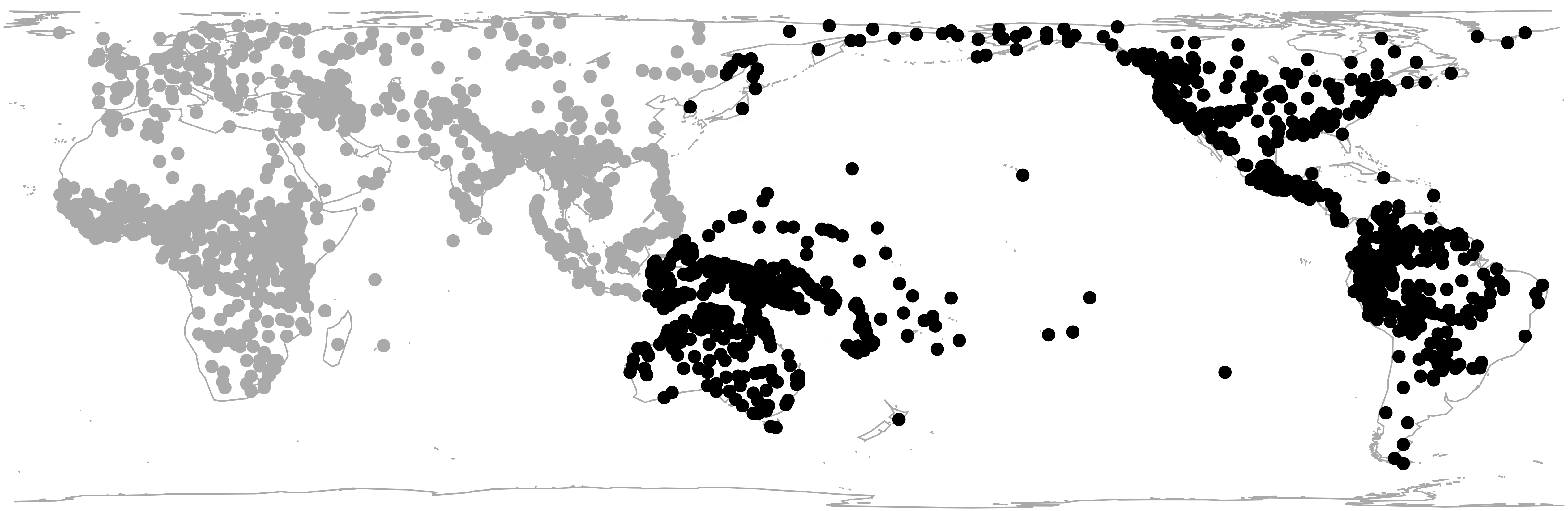
$\rightarrow$ apply in two case studies

focusing on methods

Case Study #1: the Trans-Pacific Hypothesis

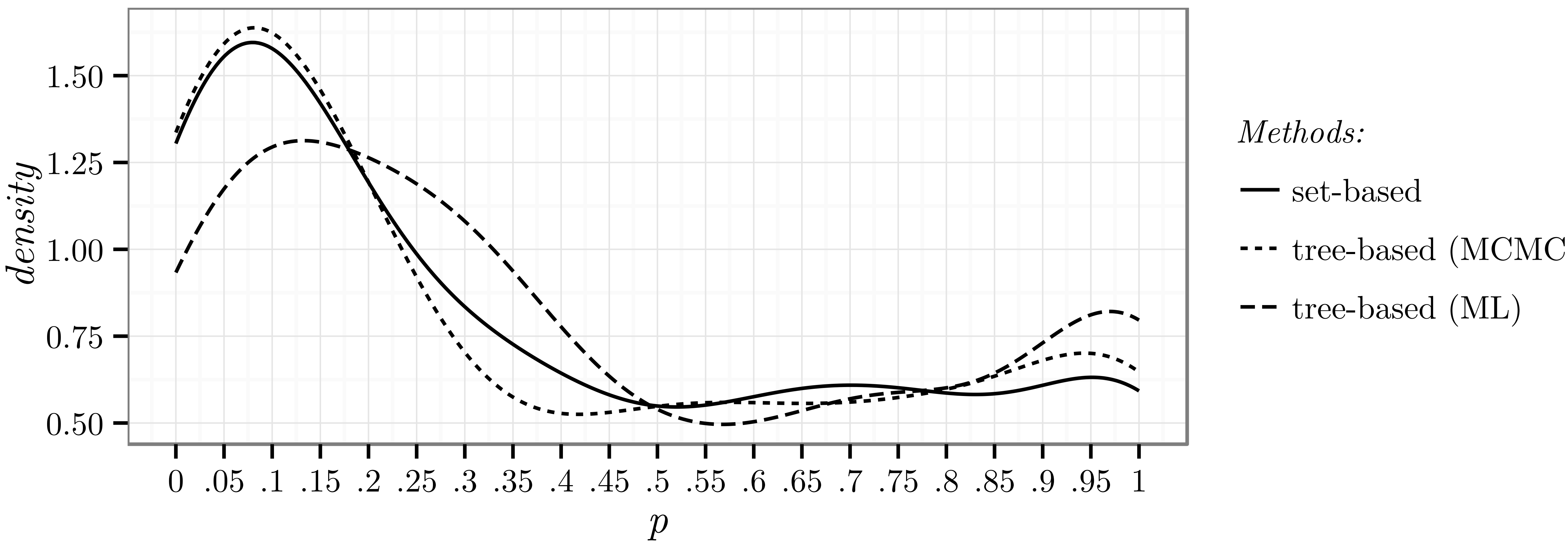
- Causal theory grounded in the peopling of the Pacific and the Americas vs. the younger spreads in Eurasia 20-1kya and Africa in the past 2ky: contact triggers change towards similar properties

- **Hypothesis:** families show different diachronic biases in the Trans-Pacific area vs. elsewhere, keeping many diverse properties that were swept away through contact elsewhere



Case Study #1: the Trans-Pacific Hypothesis

- Data from AUTOTYP and (re-coded) WALS, $N \geq 250$, $k < 10$
- 354 multinomial variables coded for N=[250, 1370] languages
- Set-based family bias estimates of large ($N \geq 5$) families with, α=.1
- Tree-based family bias estimates of non-constant large families, BF>2
- Extrapolations, then Fisher Exact Test of MEAN BIASES IN VARIABLE $\times$ AREA



Case Study #1: the Trans-Pacific Hypothesis

- False Discovery Rate (q) estimates (using Dabney & Storey's 2014 bootstrap method):

	Significant at $\alpha < .05$	q at that level	Significant at $q < .1$
Set-based	73	0.16	32
Tree-based (MCMC)	71	0.15	26
Tree-based (ML)	43	0.27	17

- From this, subtract variants of variables, e.g re voicing distinctions in WALS:
  - MADVOI: {none, in_plos_&_fric, in_plos_only, in_fric_only}
  - MADVOI2: {none, some}

→ **30 true discoveries** (mean, set-based and MCMC-based estimates)

Case Study #1: the Trans-Pacific Hypothesis

- Top 15:

Variable	*Source*	N(lgs)	p (sets)	p (MCMC)	p (ML)	*Trans–Pacific*	*Other*	*Variant of*
MADVOI2	WALS	565	0.0000	0.0000	0.0001	–voicing	+voicing	
DRYPOS	WALS	794	0.0000	0.0007	0.0069	+poss pref	–poss pref; +poss suff	
MADVOI	WALS	565	0.0000	0.0018	0.0079	–voicing in plos/fric	+voicing in plos/fric	MADVOI2
DRYPOS0	WALS	591	0.0000	0.0003	0.0000	+poss pref;–poss suff	–poss pref; +poss suff; –both	DRYPOS0
MADLAT2	WALS	565	0.0001	0.0002	0.0002	–laterals	+laterals	
BAKADP2	WALS	377	0.0002	0.0002	0.0009	–adp	+adp	
DRYGEN	WALS	1102	0.0002	0.0024	0.0009	–NGen	+NGen	
MADLAT	WALS	565	0.0002	0.0031	0.0046	–non-obstr lat	+non-obstr lat	MADLAT2
DRYGEN0	WALS	1020	0.0002	0.0002	0.0001	–Nnp	–npN; +Nnp	DRYGEN
POLYAGR	AUTOTYP	331	0.0004	0.0001	0.0018	–without;+POLYAGR	+without; –POLYAGR	
DRYDEM0	WALS	1011	0.0004	0.0004	0.0017	+DemN;–NDem	–DemN; +NDem	
MADPRS	WALS	565	0.0006	0.0000	0.0019		+Labial–velars	
LOCUS.POSS	AUTOTYP	270	0.0008	0.0376	0.3543		–H	
MADTON02	WALS	525	0.0008	0.0009	0.0029	+atonal;–tonal	–atonal; +tonal	
HASWAN03	WALS	269	0.0011	0.0011	0.0055	+desid aff	+implicit subj; –desid aff	
LOCUS.POSS.S	AUTOTYP	276	0.0013	0.0025	0.3346		–H	LOCUS.POSS

- Pearson Residual Analysis:
  - 83% positive for outside Trans-Pacific (mean across methods)
  - 28% positive inside Trans-Pacific (mean across methods)

Case Study #2: The Anti-Ergative Hypothesis

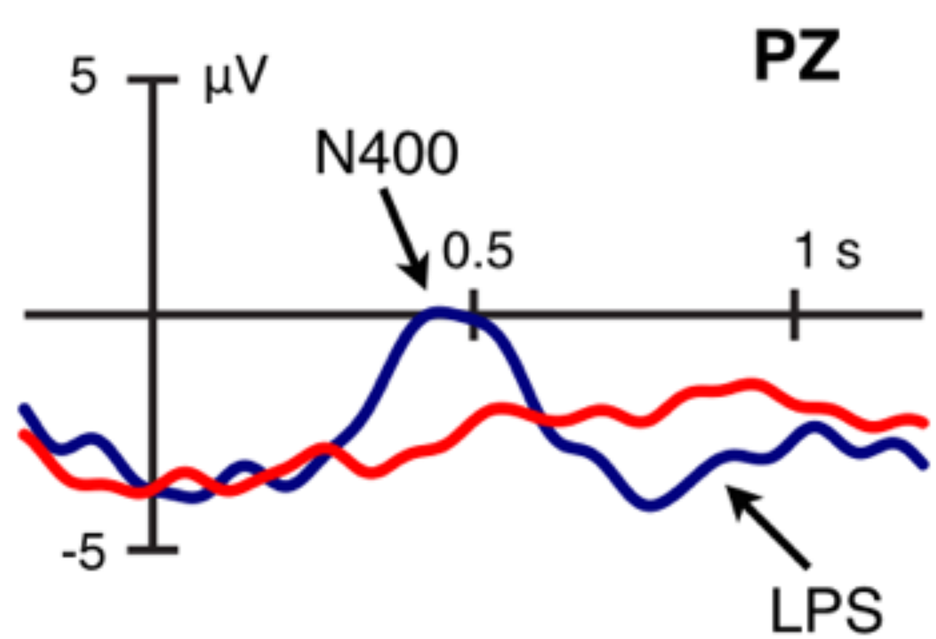
- **Primacy of A arguments in processing:**

dass *Peter* *Lehrerinnen*
that Peter: ~~S~~/A/P? teachers: A/P?

- ***mag*** **[NP1 was A!]**
  likes
- ***mögen*** **[NP1 was P!]**
  like

- The comprehension system tends to first assume that an unmarked initial NP is S or A, but not P
- If this NP later turns out to be P, this triggers an N400 (+ LPS):

  $\rightarrow$ ERP effect ("Anti-Ergative Effect")



Case Study #2: The Anti-Ergative Hypothesis

The Anti-Ergative Effect is independent of:

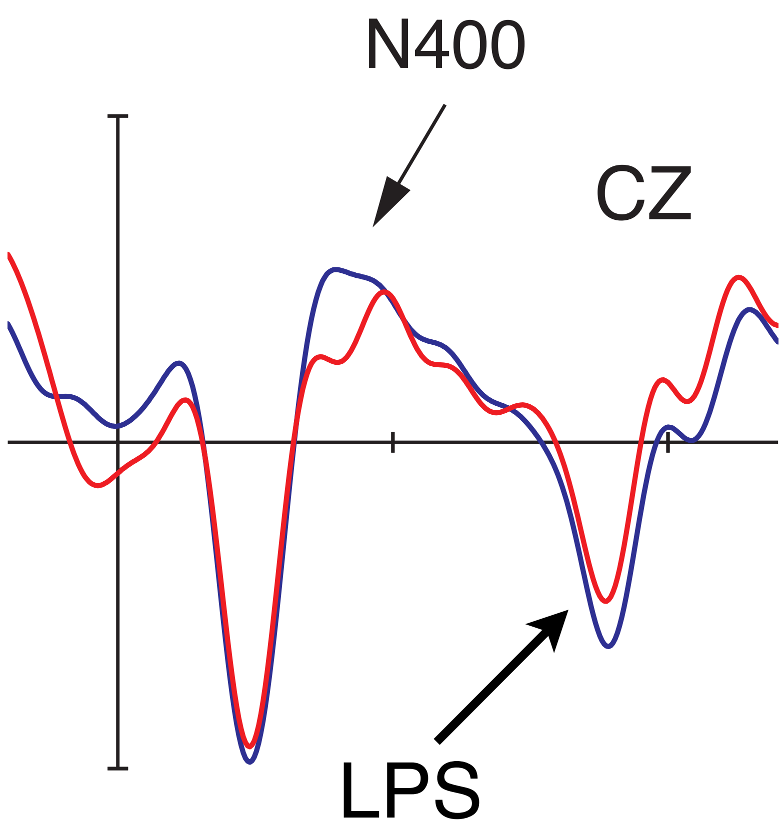
- *Frequency:* because of frequent A drop, initial NPs in Turkish tend to be P arguments, but the effect is still there (Demiral et al. 2008[*])
- *Animacy:* initial NPs in Turkish tend to be inanimate, but the effect is still there (Demiral et al. 2008[*])
- *Topicality:* initial NPs in Chinese show the effect regardless of whether the context makes them topical or not (Wang et al. 2010[+])
- *The role played by {S,A} vs {P} alignment in grammar:* very restricted relevance in Chinese but the effect is there nevertheless (Wang et al. 2009[#])

[*] *Cognition*, [+] *Lang Cogn Proc*, [#] *CUNY Conf Hum Sent Proc*

Case Study #2: The Anti-Ergative Hypothesis

And it even shows up in languages with ergative case, such as Hindi:

kitāb	*bec-ī*	*(Rām-ne)*
book(FEM)[NOM]	sell-PP.FEM	Ram-ERG
kitāb-ko	*bec-ā*	*(Rām-ne)*
book(FEM)-ACC	sell-PP.MASC	R-ERG



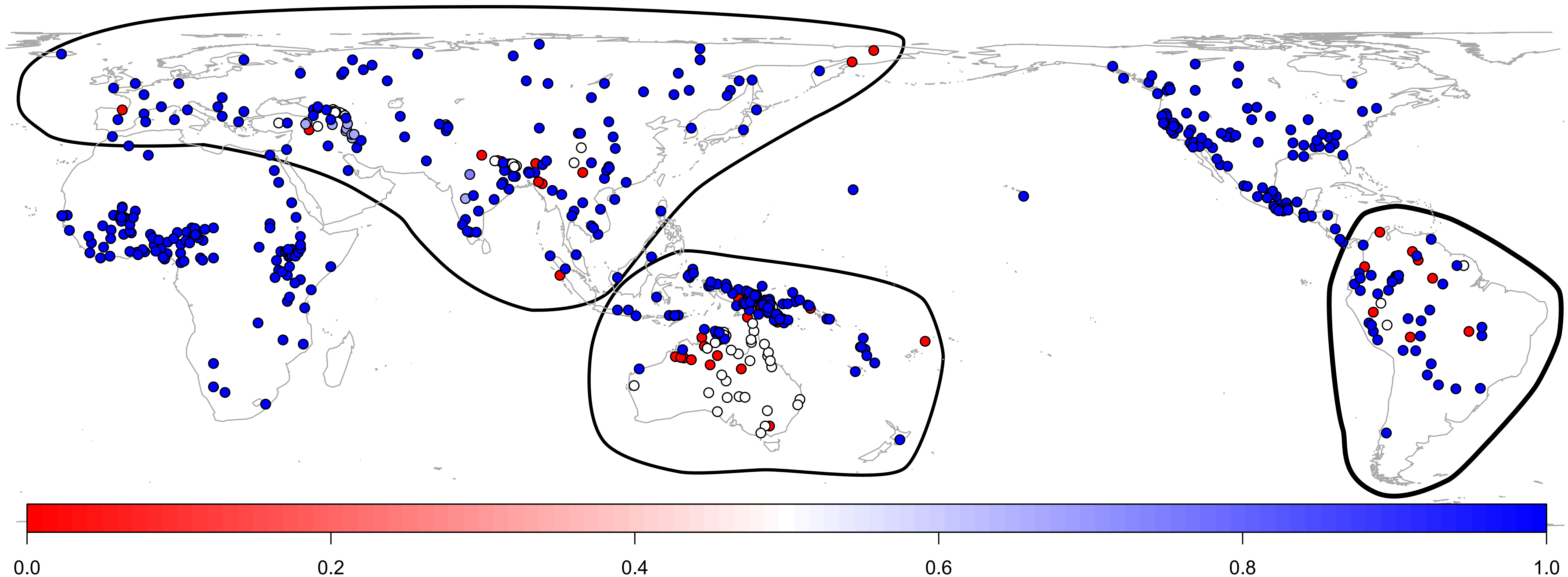
Although Hindi NOM structurally includes and often prefers a P-reading, the processing system first interprets it as S or A!

Hypothesis:

- If the Anti-Ergative Effect indeed applies universally to every unmarked initial NP, and if systems adapt to their processing environment, expect them
  - to attempt to reanalyze initial NPs as covering {S,A}
  - to avoid reanalyzing initial NPs as covering {S,P}

Case Study #2: The Anti-Ergative Hypothesis

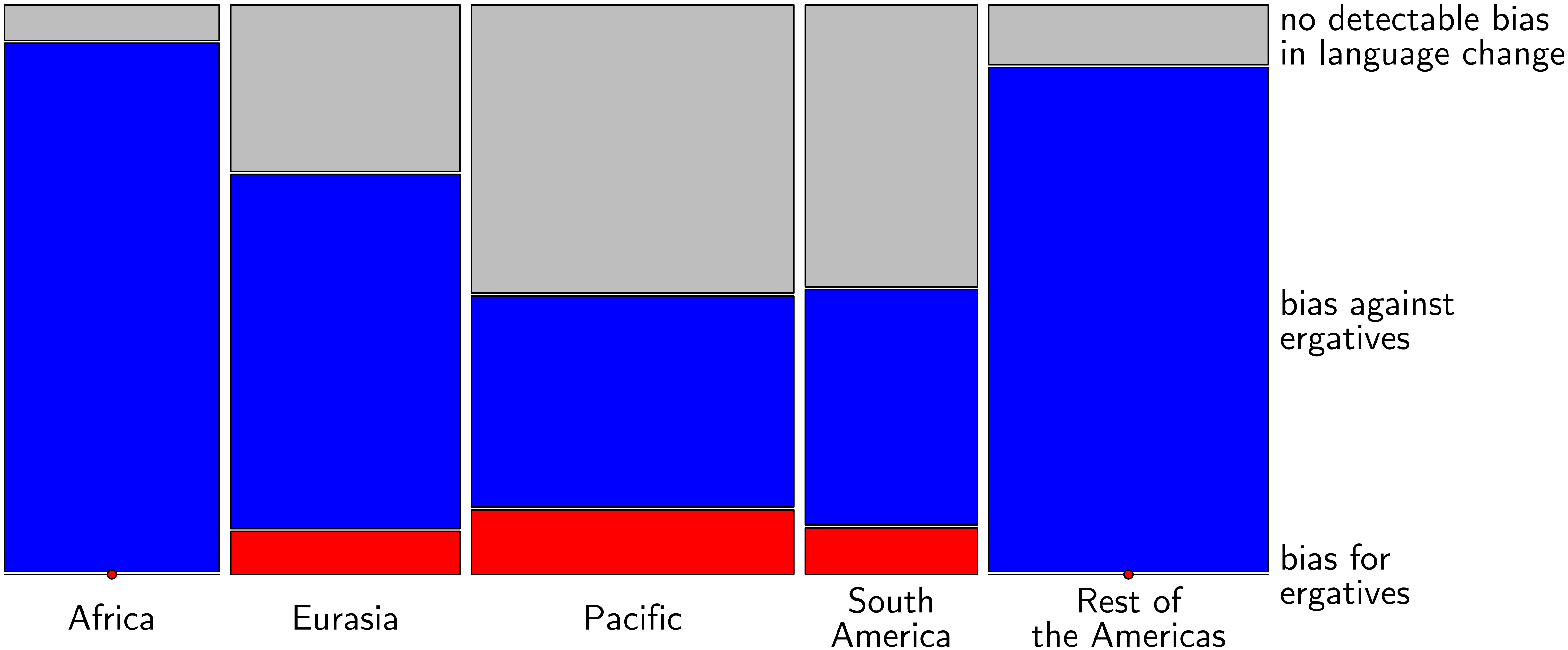
- Tested on 617 languages, 712 subsystems (e.g. past vs. nonpast); excluding V-initial structures
- Controlling for possible event-based areal diffusion effects



E (S≠A) A (S=A)

(means per language, across all NP types, clause types, and valency classes)

Case Study #2: The Anti-Ergative Hypothesis



Bias for ergatives vs. against ergatives is determined both by:

- contact histories (AREA × BIAS DIRECTION, LR $p<.01$)
- Anti-Ergative Effect: more ergative biases than anti-ergative biases across all areas (binomial $ps<.05$)

Results are the same across methods and genealogical data (set-based vs tree-based estimates, AUTOTYP vs. GLOTTOLOG trees etc.)

Conclusions

- Causal theories are tricky in traditional, Pāṇinian linguistics
- Alternative: theories of historical contact events and functional constraints
  $\rightarrow$ causes for biases in language change
- Now testable (though we obviously still need better methods, e.g. sensitive to partial tree or network structures in families)
- Describe language so we can test theories: descriptions need to become even more typologically informed than in the past