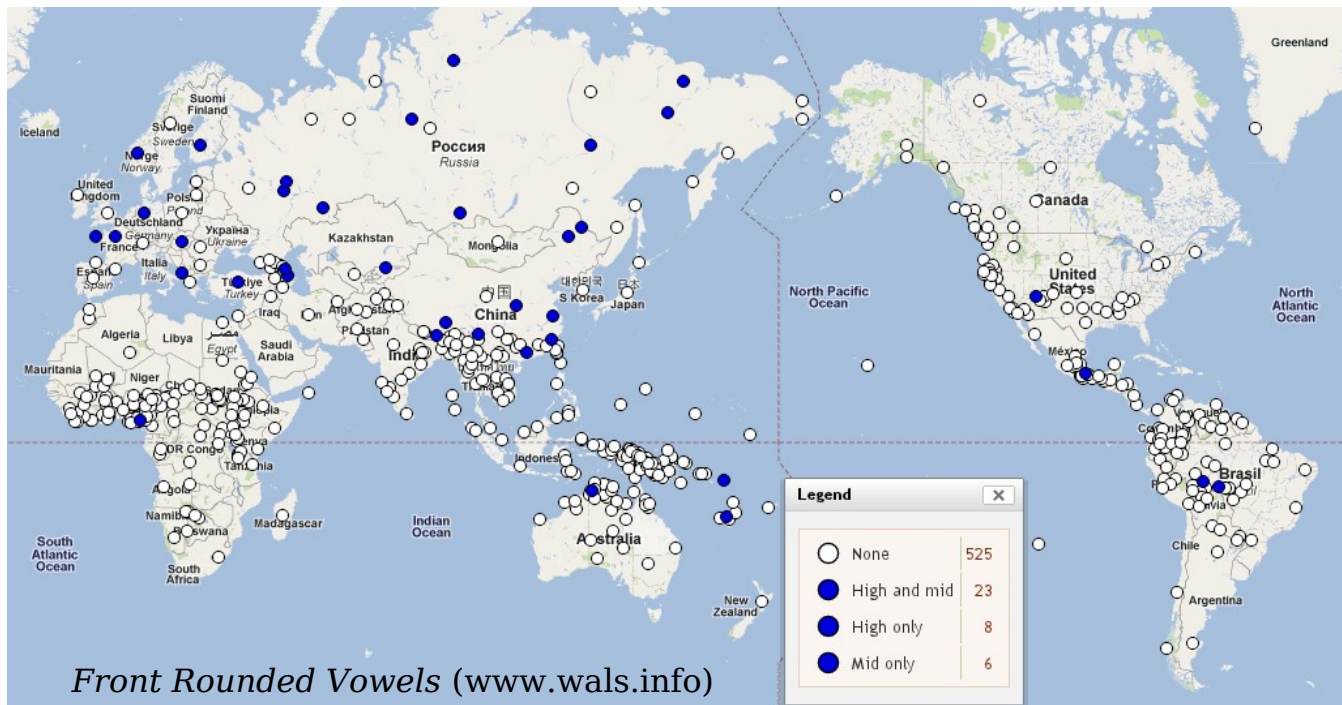


# The stability of typological features



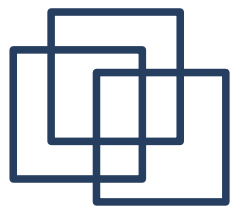
*Coelacanths* ("living fossil")

Dan Dediu

The Max Planck Institute for Psycholinguistics  
The Donders Institute for Brain, Cognition and Behavior

28 July, 2011  
ICHLXX, Osaka, Japan

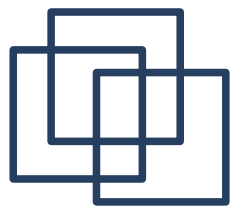
Nijmegen, The Netherlands



# Overview

---

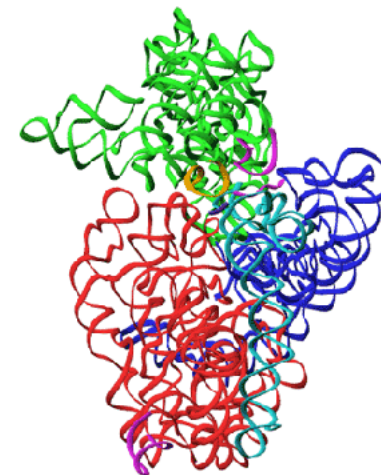
- Stability in biology
- Stability in language: basic vocabulary
- Stability in language: typology
- Quantifying stability
- Comparing methods
- Conclusions



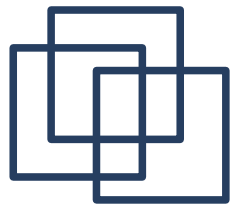
# Stability in biology

---

- **Rate of evolution** → *complex outcome*
- **Lineage-specific** & -independent comp
- *Neutral loci* → molecular clock
- *Nearly neutral loci* → mutation - drift
- *Mutation rate* → varies across genome, species, time...
- *Selection* → purifying vs positive
- **Highly conserved genes:** *rRNA*, *Pax6*
- **Very fast evolving genes:** immune system, male reproductive biology, HARs, microcephaly genes, *FOXP2*...




*rRNA* (30S)

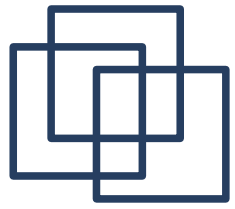


# Stability of basic vocabulary

- **Swadesh list** → cognate classes → **rate** of change
- **Pagel *et al.* 2007** (IE):
  - most *stable* (“two”, “who”...): half-life > 10,000 years
  - most *unstable* (“dirty”, “guts” ...): half-life ~750 years
  - present-day *frequency of use*
- **Pagel & Meade 2006** (IE vs Bantu):  $r = 0.28$ ,  $p < 0.03$
- **Greenhill *et al.* 2010** (IE vs Austronesian):  $\rho = 0.37$ ,  $p < 0.0001$
- **Pagel (2009)**: IE vs Starostin (14 lg fams) →  $r = 0.65$



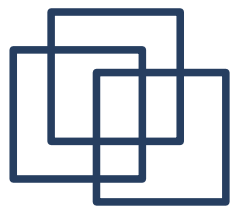
Language	"Bone"	Cognacy	Binary
Proto-Austronesian	*CuqelaN	1	10000
Paiwan	tsuqela	1	10000
Itbayaten	tuqgan	1	10000
Tagalog	butó	5	00001
Bare'e	wuku	2	01000
Mangarrai	toko	2	01000
Numfor	kor	3	00100
Motu	turia	4	00010
Fijian (Bau)	sui-na	4	00010
Tongan	hui	4	00010
Maori	iwi	4	00010



# Stability of basic vocabulary

---

- Meanings seem to have an **intrinsic** (i.e., cross-family & area) **stability**
- Accessible through very **different methods**
- Partially explained by **frequency of use**

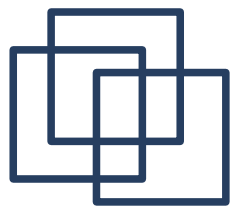


# Stability of typological features

---

- **Genetic biasing hypothesis** (Dediu & Ladd, 2007): genetically “anchored” cultural features → **tend** to change slower
- **Dunn et al. 2005, 2008**: Oceanic & Papuan
- **Hunley et al. 2008**: typology resists borrowing
- **Greenhill et al. 2010**: IE & Austronesian →
  - similar rates for typology & vocabulary
  - very weak corr b/w lg fams:  $\rho = 0.17$ ,  $p = 0.1$
- **Dunn et al. 2011**: lg fam-specific processes

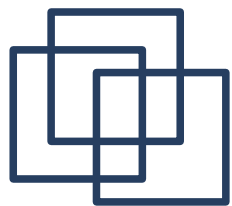
**Too few language families to allow generalization!**



# Stability of typological features

---

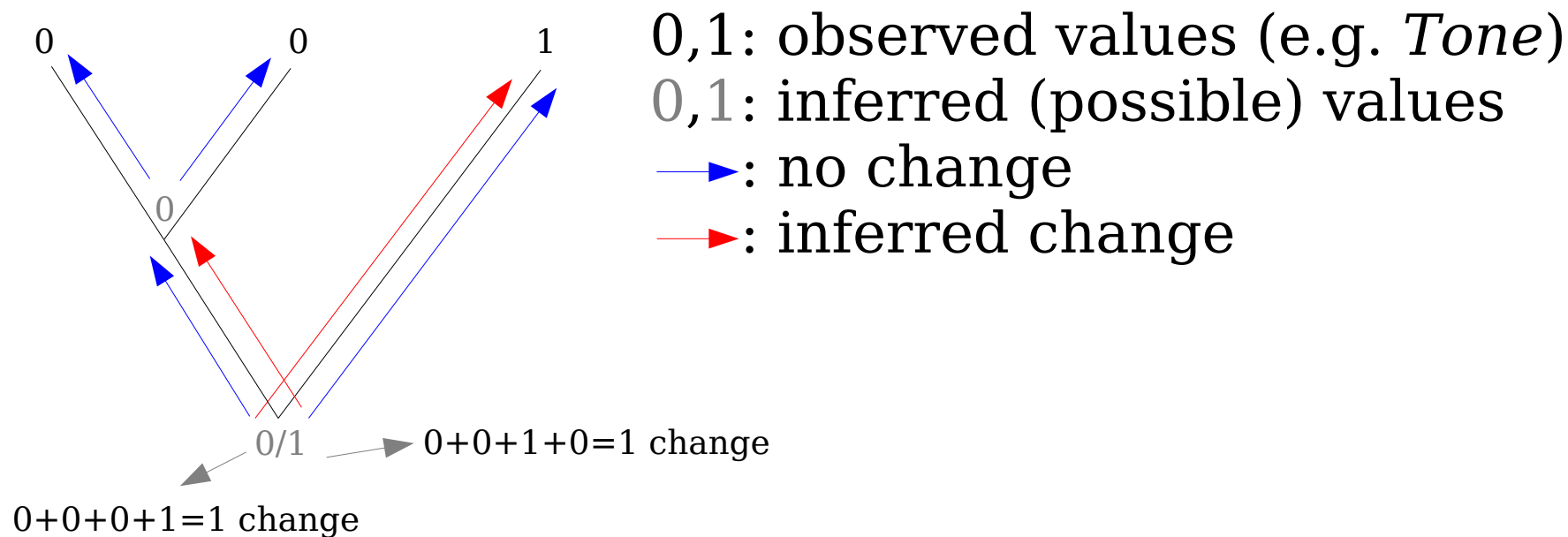
- **Phylogenetic** method → **Bayesian (model-based)**
- As many language families and features as possible
- **Primary data: WALS** ([www.wals.info](http://www.wals.info))
- **Selection** and **recoding**: unordered, ranked, custom
- **Binary “aspects”**: e.g. *Tone* → binary *Tone1* (no tone vs all) & *Tone2* (complex tone vs all)
- **Polymorphic** and **Binary** features
- **Historical classifications**: *WALS* & *Ethnologue* → not fully independent
- **Software**: *MrBayes 3* & *BayesLang*



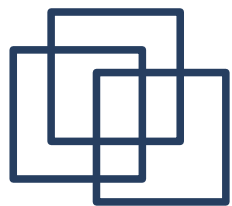
# Estimating rates of change

## Principles

- **Given** linguistic classification → **infer** feature history

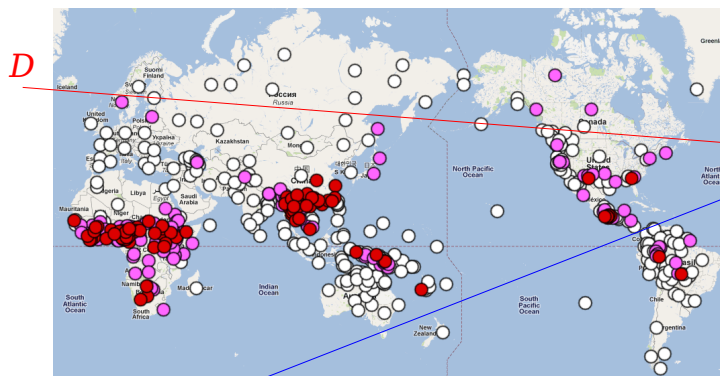
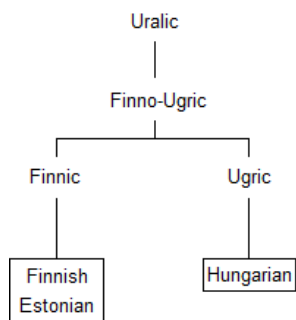


- **Counting changes** ~ maximum parsimony
- Model of evolution (branch length, probability of change, etc) → **rate of change**

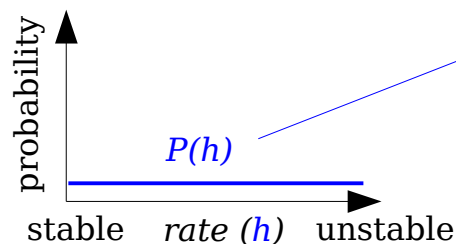


# Bayesian principles

- **Set of hypotheses** (rates of change, ...)
- **Model** (likelihood function → how features change)
- Observed **data** (feature values and languages tree)
- **A priori** probability of these rates
- ⇒ **a posteriori** probability of these rates = the results

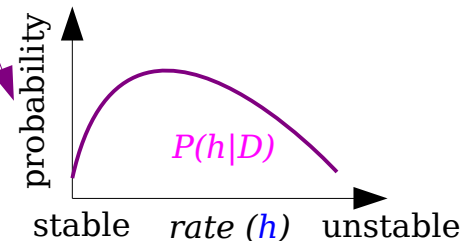


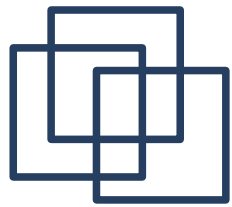
$$P(h | D) = \frac{P(D | h)P(h)}{P(D)}$$



$$Q = \begin{pmatrix} - & \theta_{01}\pi_1 & \theta_{02}\pi_2 \\ \theta_{01}\pi_0 & - & \theta_{12}\pi_2 \\ \theta_{02}\pi_0 & \theta_{12}\pi_1 & - \end{pmatrix} \mu$$

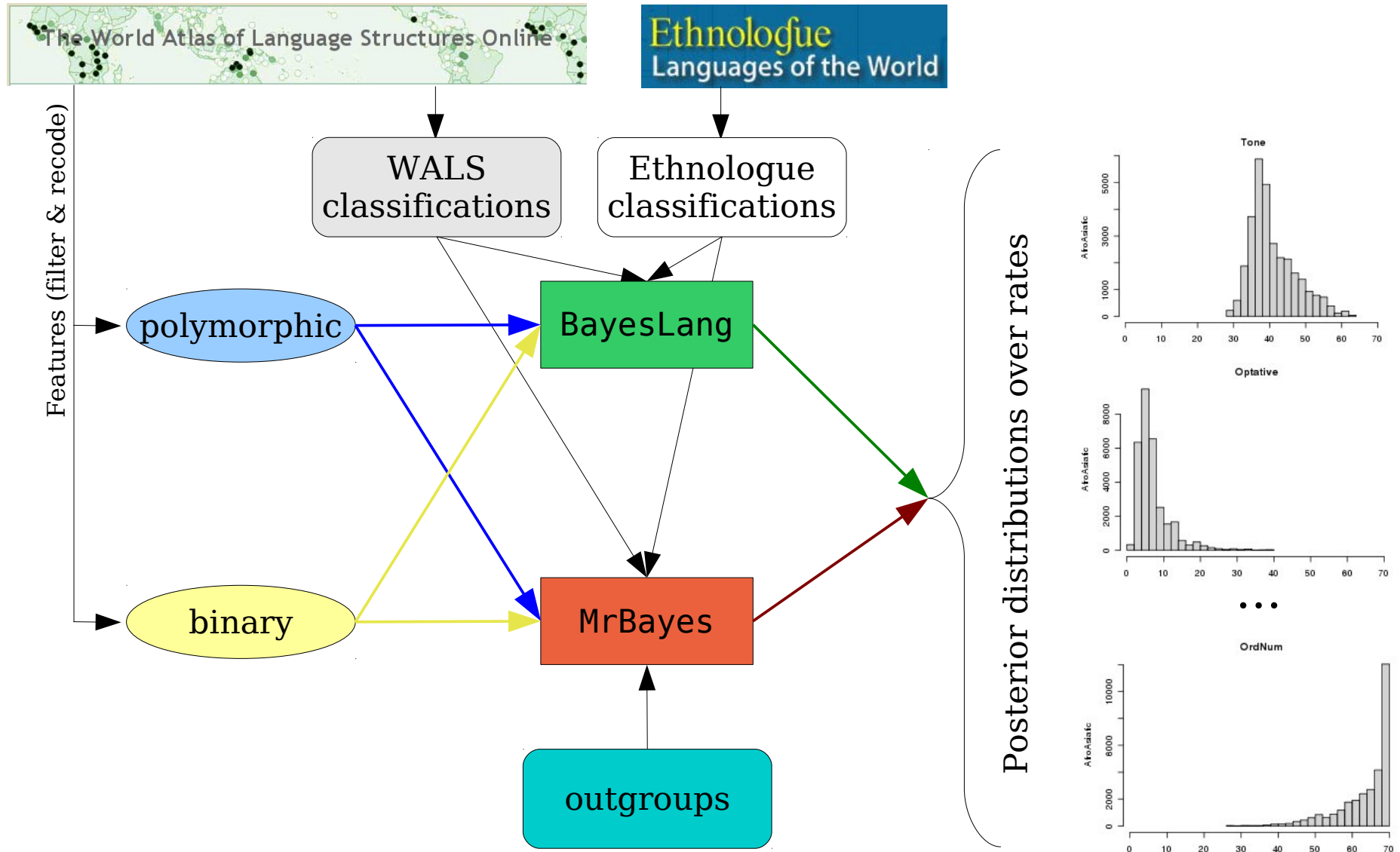
*Model P(D|h)*

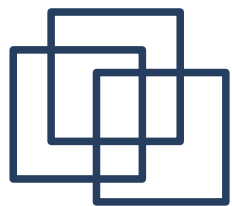




# Estimating rates of change

## Workflow





# Estimating rates of change

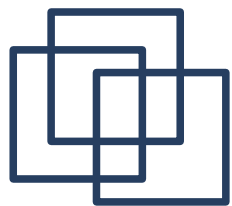
## Datasets & outgroups

<i>Dataset</i>	<i>Software</i>	<i>Coding</i>	<i>Classif</i>	<i>Feats</i>	<i>LgFams</i>	<i>Lgs</i>	<i>Outgroups</i>
<b>BM</b>	MrBayes	Binary	WALS	86	25	255	Basque & Ainu
			Ethnologue	86	33	320	23 isolates
<b>BB</b>	BayesLang	Binary	WALS	86	26	186	
			Ethnologue	86	39	303	
<b>PM</b>	MrBayes	Poly	WALS	68	18	162	Basque & Ainu
			Ethnologue	70	28	278	23 isolates
<b>PB</b>	BayesLang	Poly	WALS	70	25	249	
			Ethnologue	70	28	195	

- **Outgroups:**

- Required by **MrBayes** for rooting
- Isolates: problematic → using many for inter-correlations & correlations with **BayesLang**

- **54 datasets → 113,246 phylogenies**

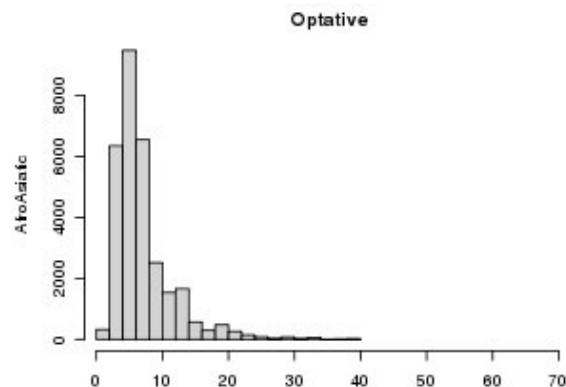


# Estimating rates of change

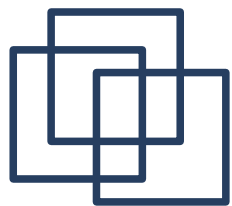
## Output

- **Bayesian approach:** for each phylogeny → **posterior distribution** of 100,000+ **samples** of relevant params
- **Here:** for each *language family*  $\phi$  and *feature*  $F$  → **rate estimate**  $r_{\phi, F}$  → sample of such rate estimates

Feature Sample	$F_1$	$F_2$	...	$F_N$
1	0.1	0.01		0.9
2	0.15	0.04		0.87
...				
1,000,000	0.14	0.02		0.83



- Absolute rates cannot be compared across trees, methods & datasets → conversion to **ranks**
- **Summarize** rank estimate distributions: **mean & median**

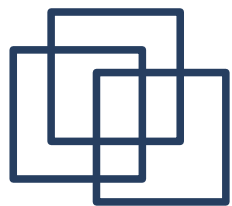


# Estimating rates of change

## Results

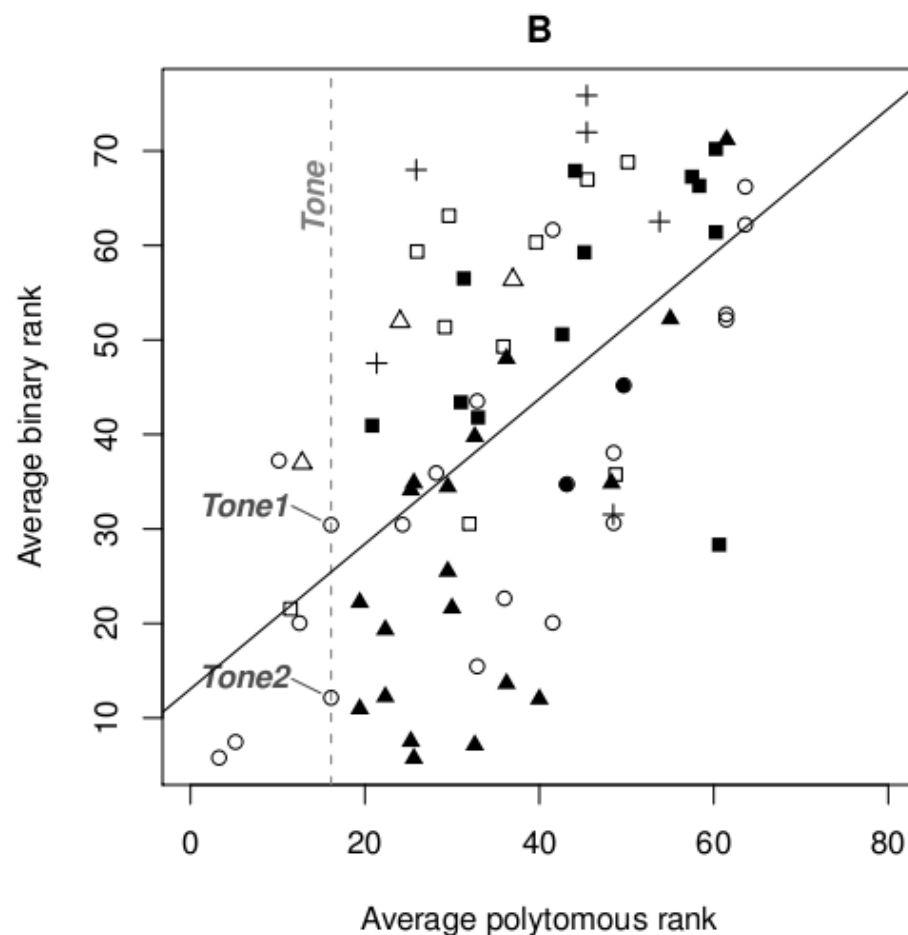
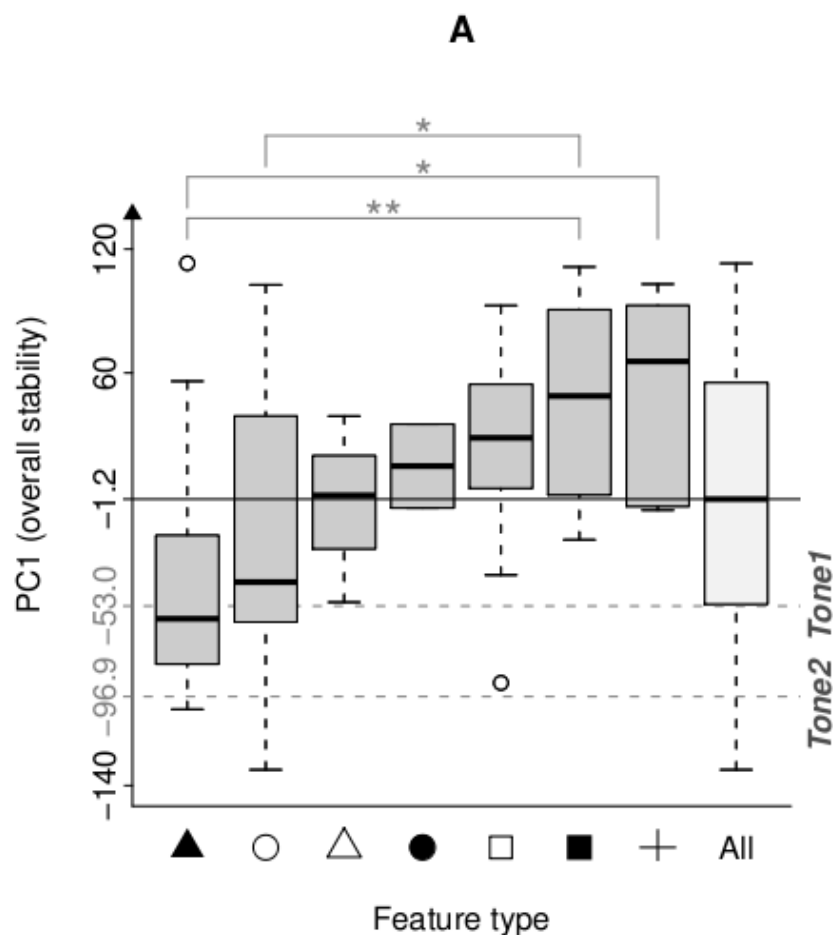
---

- **Summaries** by means & medians:  $0.94 \leq r \leq 0.98, p < 10^{-6}$
  - **WALS & Ethnologue**:  $0.96 \leq r \leq 0.99, p < 10^{-6}$
  - **Outgroup** → negligible:  $0.49 \leq r \leq 0.92, p < 10^{-6}; \bar{r} = 0.78$   
PC<sub>1</sub>: 79% variance
  - **Binary**:  $0.59 \leq r \leq 0.98, p < 10^{-8}; \bar{r} = 0.78; PC_1: 81.4%$ 
    - *Tone2*: **8** of **86**; *Tone1*: **23** of **86**
  - **Poly**:  $0.51 \leq r \leq 0.99, p < 10^{-5}; \bar{r} = 0.71; PC_1: 76.1%$ 
    - *Tone*: **8** of **68**
  - **Binary - Poly**: complex: PC<sub>1</sub> (agreement): 67.4%, PC<sub>2</sub> (bin vs poly): 16.1%
-



# Estimating rates of change

## Results (2)

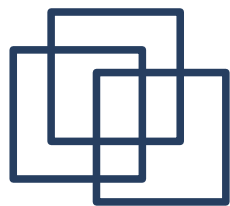


▲ = Word Order, ○ = Phonology, △ = Nominal Syntax, ● = Morphology, □ = Verbal Categories, ■ = Nominal Categories, + = Simple Clauses, and All = all types of features combined.



# Ranking of features

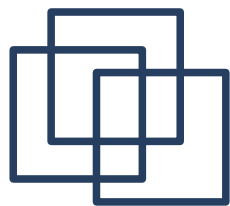
Rank	Feature	PC <sub>1</sub> score	Rank	Feature	Category	PC <sub>1</sub> score
1	AbsComC	-88.47	1	AbsComC	Phonology	-132.4
2	FrRoundV	-82.49	2	FrRoundV	Phonology	-124.9
3	VowelN	-75.09	3	NumN1	Word Order	-103.0
4	Optative	-74.75	4	OV1	Word Order	-97.6
5	UvulC	-67.85	<b>5</b>	<b>Tone2</b>	<b>Phonology</b>	<b>-96.9</b>
6	OlbPosInfl	-64.71	6	GenN1	Word Order	-96.7
7	NomLocPred	-59.97	7	SV1	Word Order	-91.1
<b>8</b>	<b>Tone</b>	<b>-53.80</b>	8	Optative	Verbal Categories	-90.2
9	NMPron	-52.22	9	AdjN1	Word Order	-88.0
10	GenN	-43.34	10	UvulC	Phonology	-86.4
11	ZeroCopPredNom	-41.88	11	SV2	Word Order	-74.2
12	NumClas	-41.00	12	GenN2	Word Order	-69.3
13	AntipassiveC	-39.93	13	IntPhCQ1	Word Order	-68.2
14	MTPron	-37.07	14	OVAdpNP	Word Order	-65.6
15	SV	-35.25	15	Vowel2	Phonology	-64.8
54	VoicPF	39.09	57	CVRatio1	Phonology	67.8
55	MorphImp	39.11	58	CVRatio2	Phonology	72.2
56	LmarkC	39.85	59	DistCDem	Nominal Categories	80.4
57	Perfect	48.11	60	PersMV	Simple Clauses	81.1
58	PersMV	55.68	61	OvSitEpi	Verbal Categories	83.1
59	OVAdjN	58.17	62	AsymCaseM	Nominal Categories	85.0
60	IndefArt	65.37	63	SymAsymStNeg1	Simple Clauses	92.7
61	P3PrDem	68.16	64	Perfect	Verbal Categories	92.7
62	DefArt	71.84	65	Cons2	Phonology	94.3
63	AsymCaseM	71.97	66	P3PrDem	Nominal Categories	96.2
64	OrdNum	73.48	67	IndefArt	Nominal Categories	100.1
65	PolQPart	75.03	68	Cons1	Phonology	102.7
66	CVRatio	76.96	69	SymAsymStNeg2	Simple Clauses	103.1
67	Ncases	78.86	70	DefArt	Nominal Categories	111.4
68	Cons	80.40	71	PolQPart	Word Order	113.1



# Interim conclusions

---

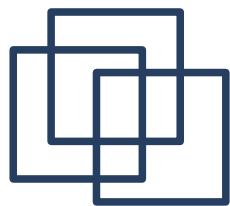
- *Tone* (*Tone1* & *Tone2*) seem to be quite stable
- Stability estimates seem to correlate across methods, datasets ... → maybe they reflect something **intrinsic**?
- **Universal** tendencies & lg fam-**idiosyncratic**
- **Caveats:**
  - *Trees*
  - Data *quality* & *quantity*
  - *Outgroups*



# Comparing methods

---

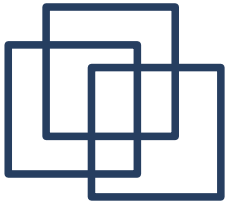
- With **Michael Cysouw**
- Compare published methods of stability estimation
- **Cysouw, Albu & Dress (2008)**: consistency with overall pattern
- **Parkvall (2008)**: borrowability vs genealogical stability
- **Wichmann & Holman (2009)**: Stable features tend to “stay in the family” (metric C)
- **Elena Maslova's** (as implemented by Cysouw & Dediu)
- **Dediu (2011)**: phylogenetic stability
- 62 shared features
- Correlations: some variation but most positive
- **PC<sub>1</sub>: 48.3%** (agreement)
- All these different methods seem to agree on stability



# Conclusions

---

- Stability estimates seem to **agree** across very different methods and concepts
- **Universal** tendencies → why?
  - universals of usage?
  - cognitive, articulatory, etc biases?
  - “**hubs**” in a “**linguistic network**”
- **Idiosyncratic** (family-specific) component
- **Method specificity**
- **Future work:**
  - *patterning of variation* (in progress; S. Levinson)



# THANK YOU!

---

Dediu, D. (2011). A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society of London, B-Biological Sciences* **278**(1704), 474-479. doi:10.1098/rspb.2010.1595.



Echidna

**Special thanks to:** Bob Ladd, Kenny Smith, Fiona Jordan, Gwen Hyslop, Steve Levinson, Michael Cysouw, Michael Dunn, Russel Gray, Simon Greenhill and Alexandra Dima.

---