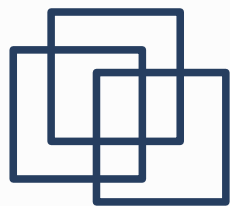


Are some **structural features** of language **more stable** than others and if so, why?



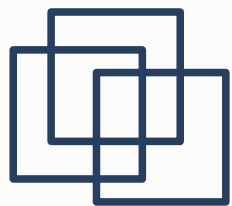
BEDLAN
22 September, 2011

Dan Dediu
Language and Genetics
The Max Planck Institute for Psycholinguistics
and
The Donders Institute for Brain, Cognition and Behaviour,
Nijmegen
The Netherlands



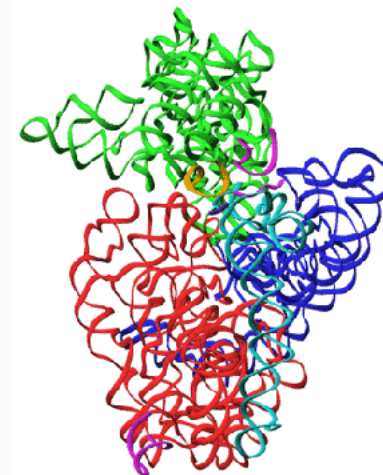
Overview

- Stability in biology
- Stability in language: basic vocabulary
- Stability in language: typology
- Dediu (2011): Bayesian phylogenetics
- Comparing the approaches
- Variation between language families
- Conclusions

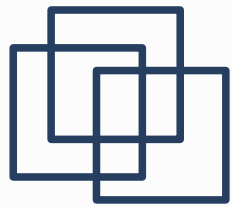


Stability in biology

- **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)

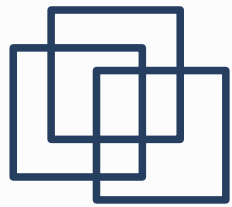


Stability of basic vocabulary

- **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) → $r = 0.65$

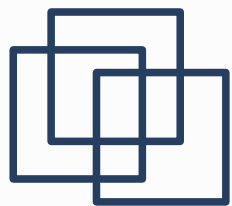


Language	"Bone"	Cognacy	Binary
Proto-Austronesian	*CuqelaN	1	10000
Paiwan	tsuqela	1	10000
Itbayaten	tuqgan	1	10000
Tagalog	butó	5	00001
Bare'e	wuku	2	01000
Mangarrai	toko	2	01000
Numfor	kor	3	00100
Motu	turia	4	00010
Fijian (Bau)	sui-na	4	00010
Tongan	hui	4	00010
Maori	iwi	4	00010



Stability of basic vocabulary

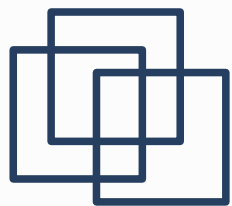
- Meanings seem to have an **intrinsic** (i.e., cross-family & area) **stability**
- Accessible through very **different methods**
- Partially explained by **frequency of use**



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

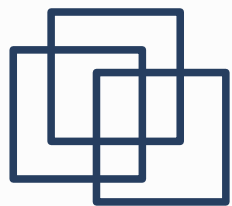
Too few language families to allow generalization!



Stability in language

2. Non-Phylogenetic/ad-hoc concepts

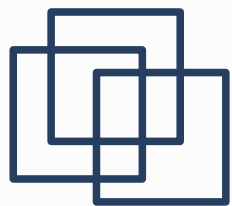
- Devised mostly by typologists
- *Literature survey* (**Dediu & Cysouw, in preparation**)
- **Criteria:**
 - *published*
 - *concept of stability* ~ “easiness” with which features change values across time, under the influence of various processes
 - *quantifiable, objective and repeatable*
 - *many features* (preferably WALS-compatible)
 - *many language families*
 - produce at least a *ranking* of features



Stability in language

Cysouw, Albu & Dress (2008)

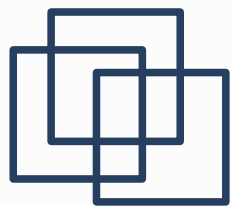
- **Consistency** of individual features vs overall patterns
 - **Typological distance** $D_F(L_1, L_2) = \begin{cases} 0, F_{L_1} = F_{L_2} \\ 2, F_{L_1} \neq F_{L_2} \\ 1, F_{L_1} \vee F_{L_2} \text{ undef} \end{cases}$
 - For N features F_i and M languages $L_j \rightarrow$
 - dist matrices b/w languages $D_i = (d_i)_{j,k} = D_{F_i}(L_j, L_k)$
 - overall dist matrix $D = (d)_{j,k} = \underset{F_i \text{ def } L_j, L_k}{\text{mean}} D_{F_i}(L_j, L_k)$
 - Quantify fit b/w D_i and D :
 - Mantel (**CM**), coherence (**CC**) and rank (**CR**)
 - Not good intercorrelation, consistent features *might* also be genealogically stable
-



Stability in language

Parkvall (2008)

- **Borrowability** vs genealogical stability
- Genealogically stable vs borrowable/transferable through contact
- Genealogical (families & subfamilies) vs areal units
- For unit U and feature $F \rightarrow$ **Herfindahl-Hirschman index** (Gini coefficient) $D_U = 1 - \sum_{i=1}^n P_i^2$, with P_i = prop of lgs in U with $\text{val}(F)=i$
- **Homogeneity** of U , $H_U = 1/D_U \rightarrow$ averaged over all families, H_F^{fam} and all areas, H_F^{are}
 \rightarrow stability of F is the ratio $\frac{H_F^{fam}}{H_F^{are}}$



Stability in language

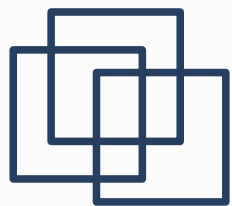
Wichmann & Holman (2009)

- Stable features tend to “**stay in the family**”
- The prob that F does not change within a language during a given time period, whatever the reason
- “**Metric C**”: genealogical group G of n_G lgs $\rightarrow \pi_G^F =$ proportion of pairs of lgs in G with same value of F :

$$R_F = \sum_G \frac{\pi_G^F}{\sqrt{n_G}}$$

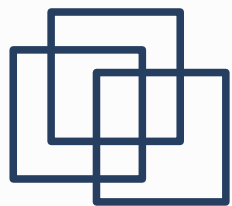
- Similar \rightarrow prop of pairs of *unrelated* lgs sharing F , U_F
- **Stability of F** is

$$S_F = \frac{R_F - U_F}{1 - U_F}$$



Stability of typological features - Dediu, 2011

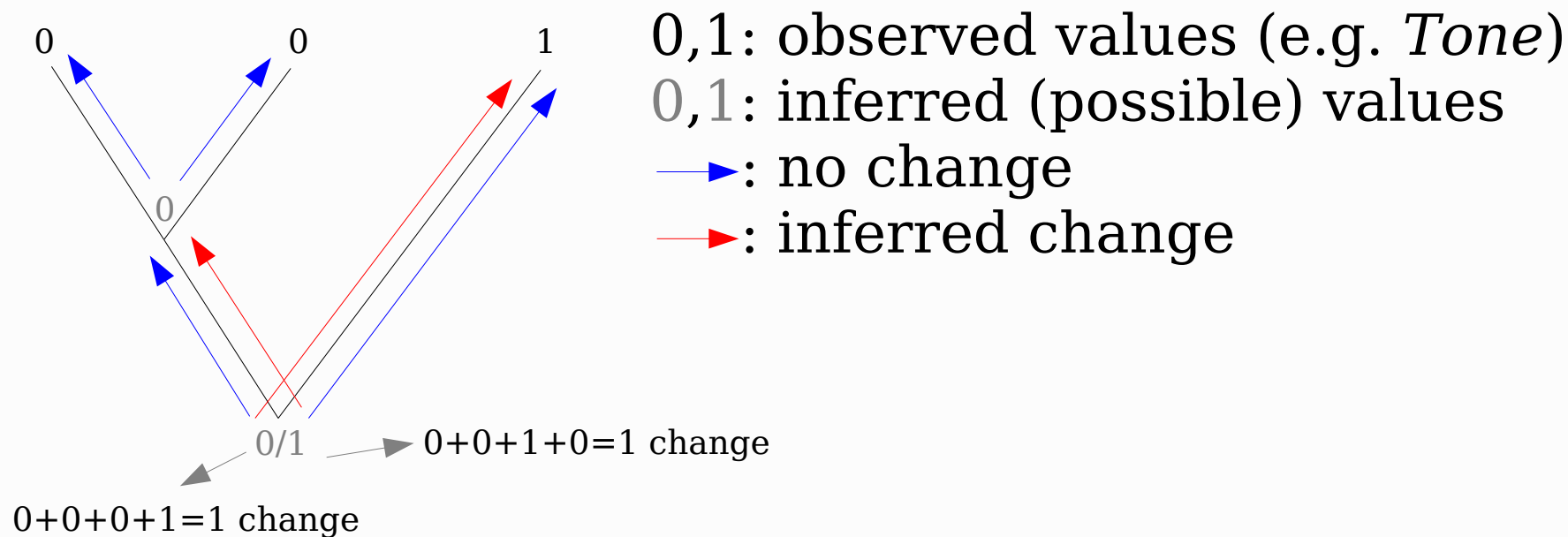
- **Phylogenetic** method → **Bayesian (model-based)**
- As many language families and features as possible
- **Primary data: WALS** (www.wals.info)
- **Selection** and **recoding**: unordered, ranked, custom
- **Binary “aspects”**: e.g. *Tone* → binary *Tone1* (no tone vs all) & *Tone2* (complex tone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



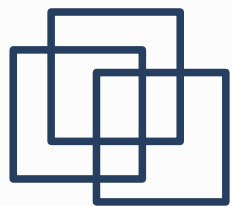
Estimating rates of change

Principles

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

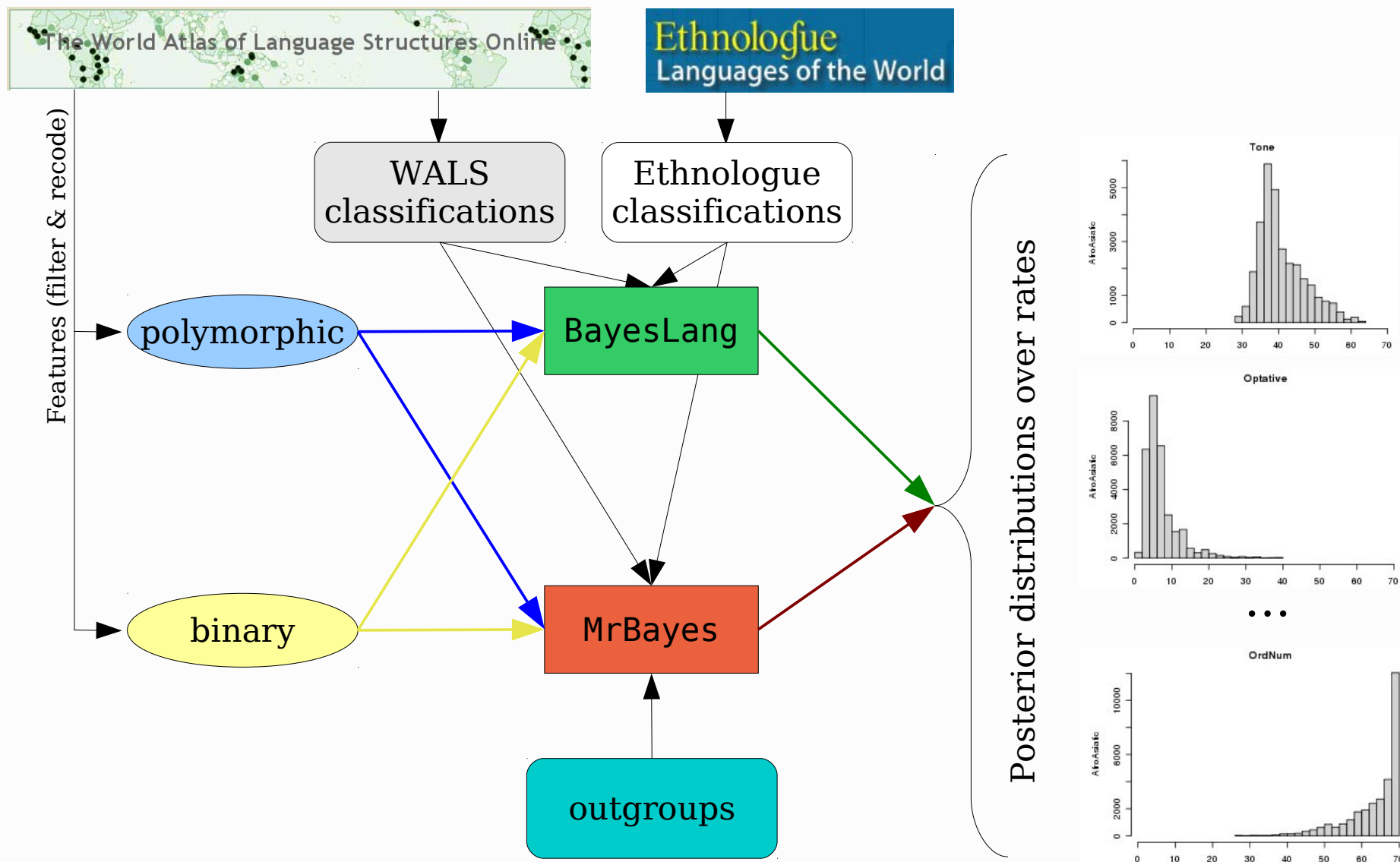


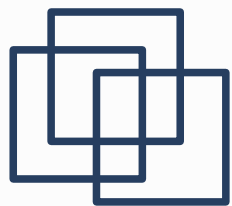
- **Counting changes** ~ maximum parsimony
- Model of evolution (branch length, probability of change, etc) → **rate of change**



Estimating rates of change

Workflow





Estimating rates of change

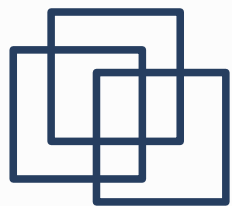
Datasets & outgroups

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

- **Outgroups:**

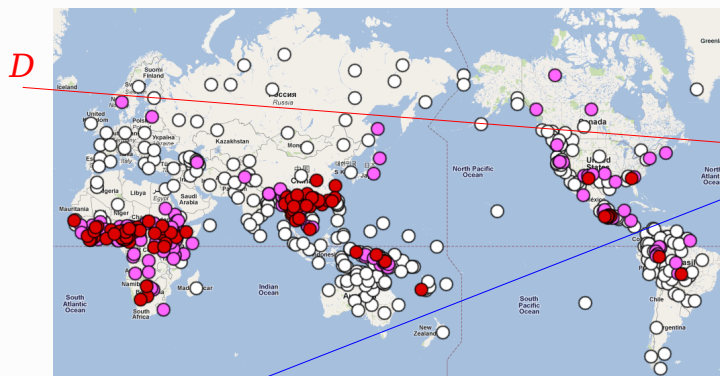
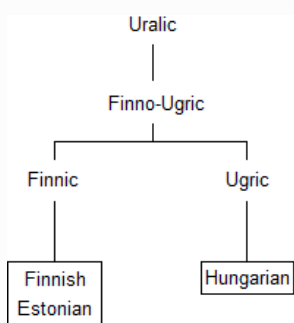
- Required by **MrBayes** for rooting
- Isolates: problematic → using many for inter-correlations & correlations with **BayesLang**

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

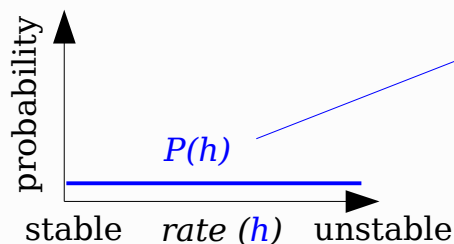


Bayesian principles

- **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

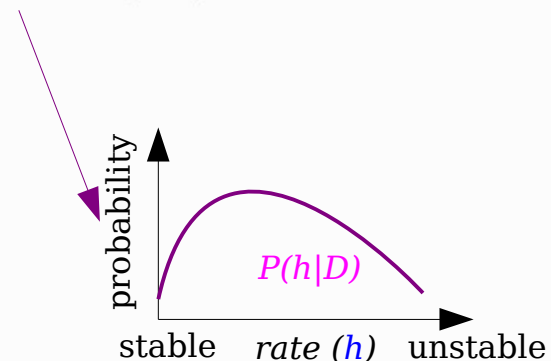


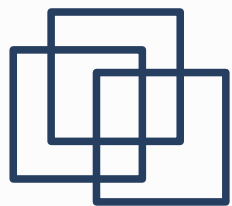
$$P(h | D) = \frac{P(D | h)P(h)}{P(D)}$$



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

Model $P(D|h)$



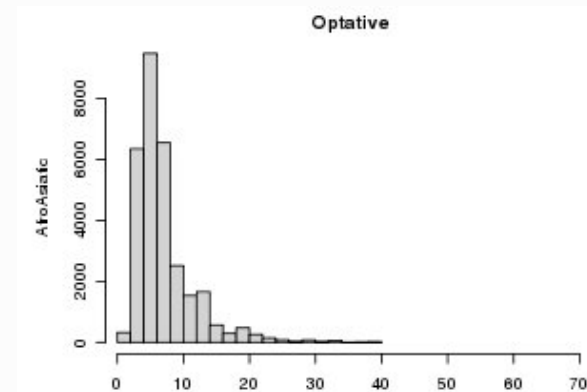


Estimating rates of change

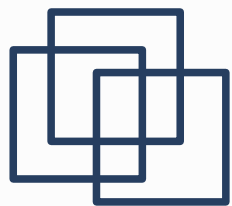
Output

- **Bayesian approach:** for each phylogeny → **posterior distribution** of 100,000+ **samples** of relevant params
- **Here:** for each *language family* ϕ 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



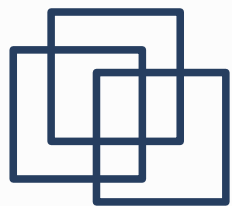
- Absolute rates cannot be compared across trees, methods & datasets → conversion to **ranks**
- **Summarize** rank estimate distributions: **mean & median**



Estimating rates of change

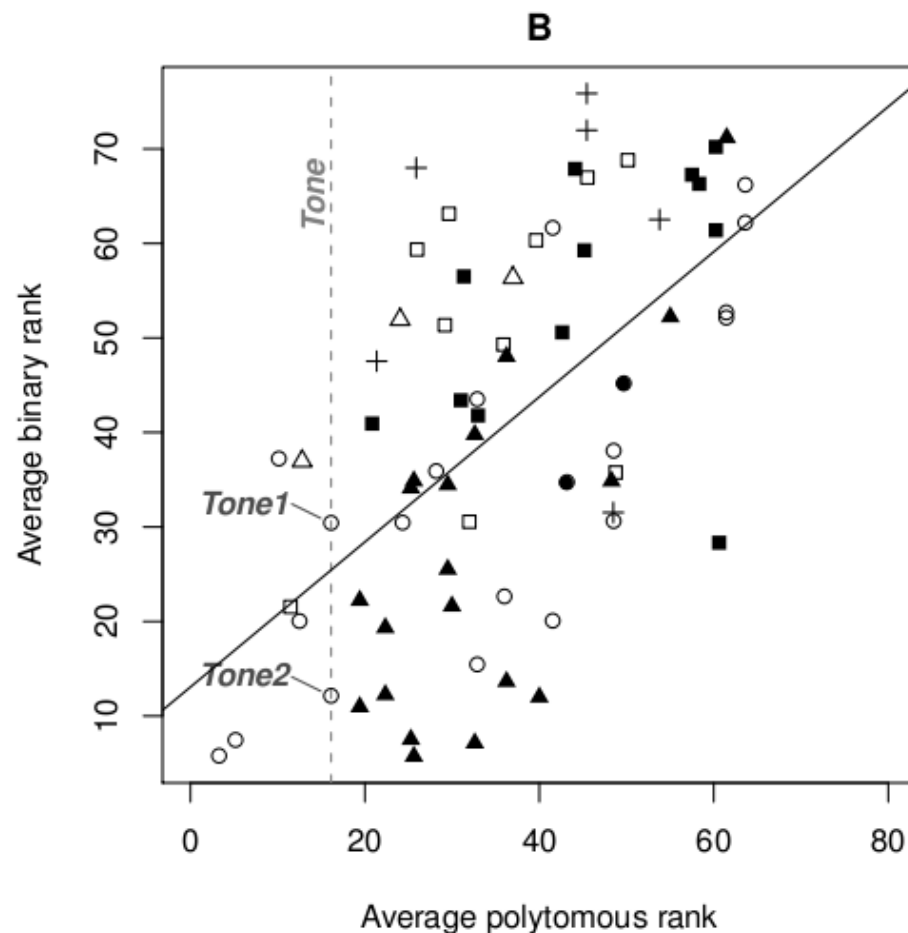
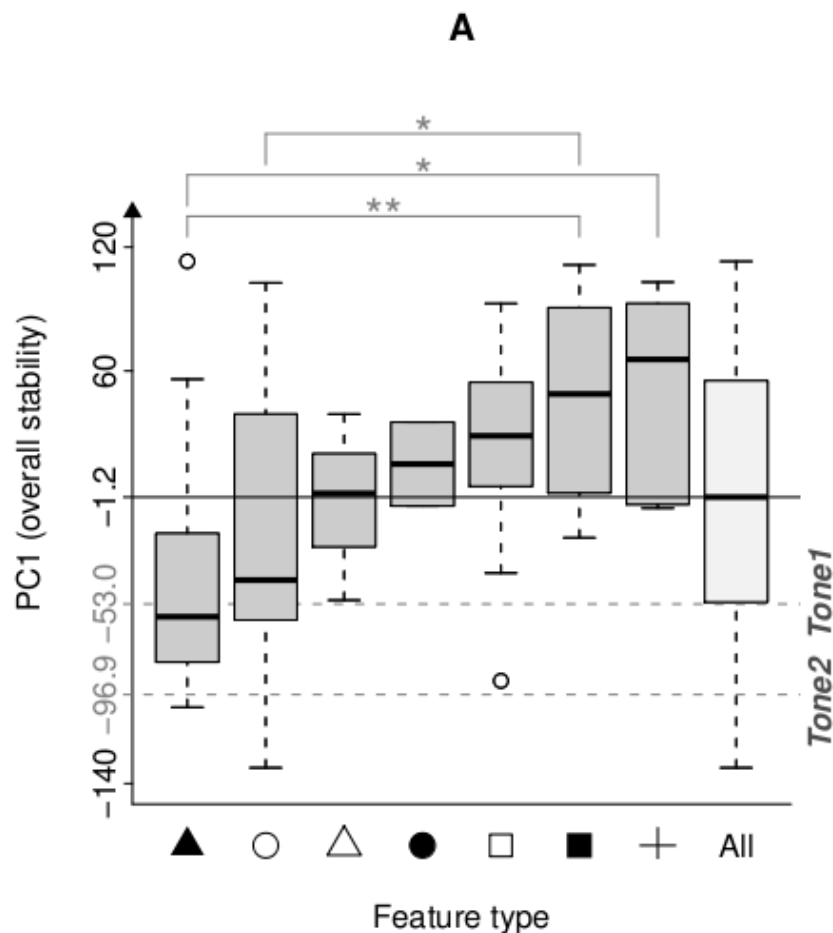
Results

- **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: $0.49 \leq r \leq 0.92, p < 10^{-6}; \bar{r} = 0.78$
PC₁: 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: PC₁ (agreement): 67.4%, PC₂ (bin vs poly): 16.1%
-

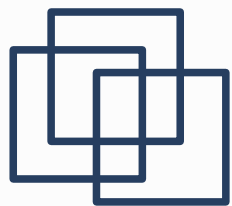


Estimating rates of change

Results (2)



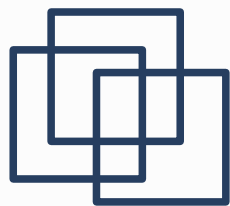
▲ = Word Order, ○ = Phonology, △ = Nominal Syntax, ● = Morphology, □ = Verbal Categories, ■ = Nominal Categories, + = Simple Clauses, and All = all types of features combined.



Stability in language

Comparing all four methods

- Cysouw, Albu & Dress (2008): **CM, CC, CR**
- Dediu (2011): **D**
- Parkvall (2008): **P₁, P₂**
- Wichamnn & Holman (2009): **W**
- **62** WALS features shared by all



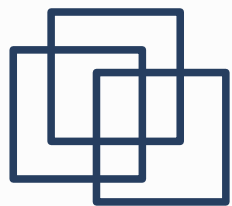
Stability in language

Comparing all four methods

- PCA:

Loadings	PC_1	PC_2	PC_3	PC_4
% variance explained	44.3%	19.6%	15.4%	9.8%
CM	0.18	0.60	0.45	0.49
CC	0.45	-0.35	0.17	0.19
CR	0.50	-0.21	0.25	0.19
D	0.46	-0.29	0.04	-0.40
P ₁	0.42	0.29	-0.30	0.11
P ₂	0.24	0.12	-0.78	0.25
W	0.27	0.54	0.08	-0.68

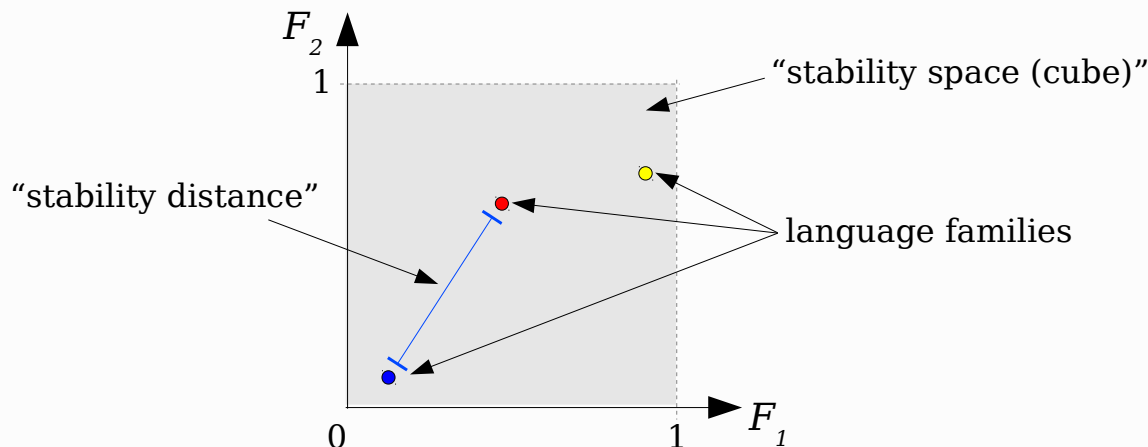
- *Tend to agree more than disagree* → **stability** has an **intrinsic component**
- *Unexpected results*: **D** (strongly phylogenetic) was expected to agree with **W** (conceptually phylogenetic), but **D** agrees with **CC** and **CR** (non-genealogical)
- The agreement between all methods (PC1) classifies *Tone* as **16** of **62**

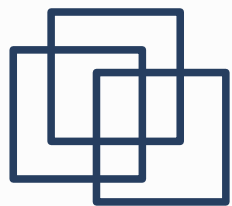


Variation among Families

“Stability space” & “stability distances”

- Multi-dimensional space of “**typological stability**” →
 - each feature = one dimension
 - possible values: **0..1** (rescaled ranks) →
“**stability hyper-cube**”
- Each language family → “**stability profile**”
- “**Stability distances**” between language families in this stability space → **max possible distance** \sqrt{N}



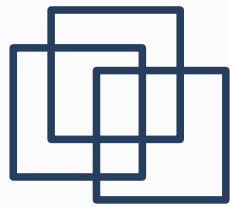


Variation among Families

Distribution of families in stability space

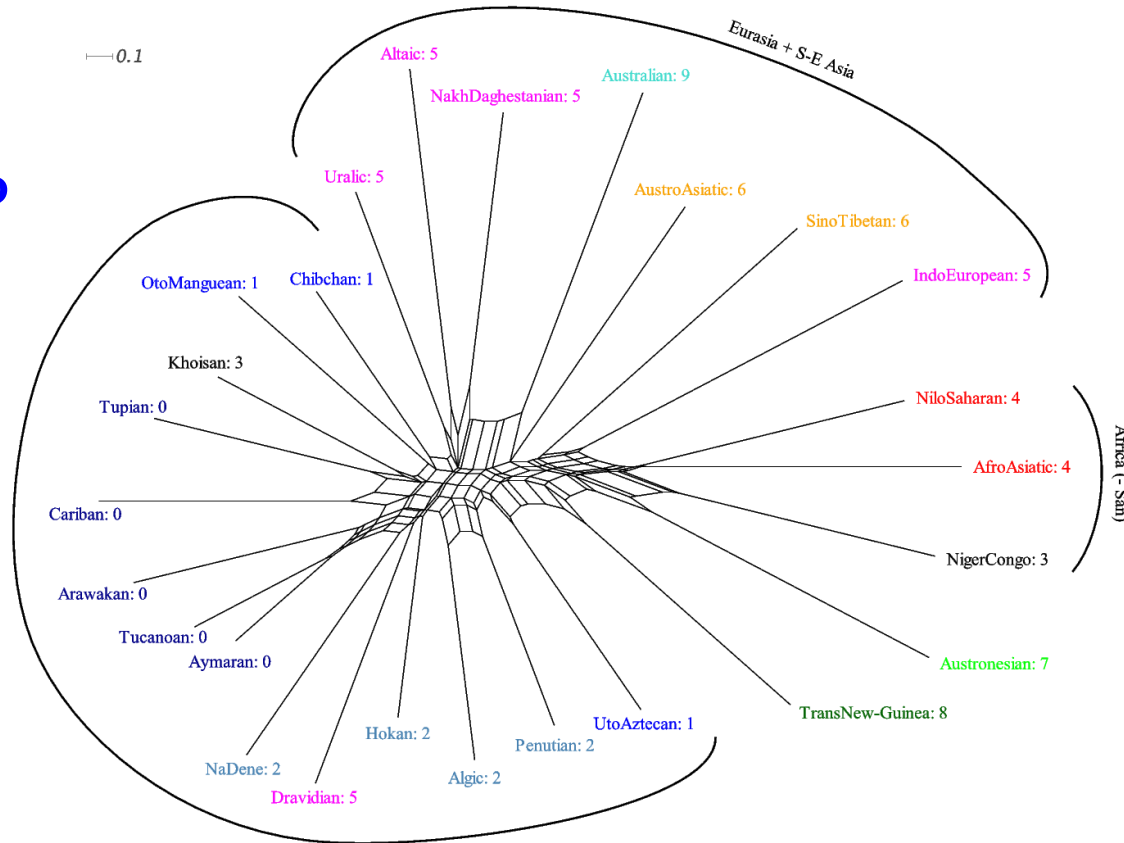
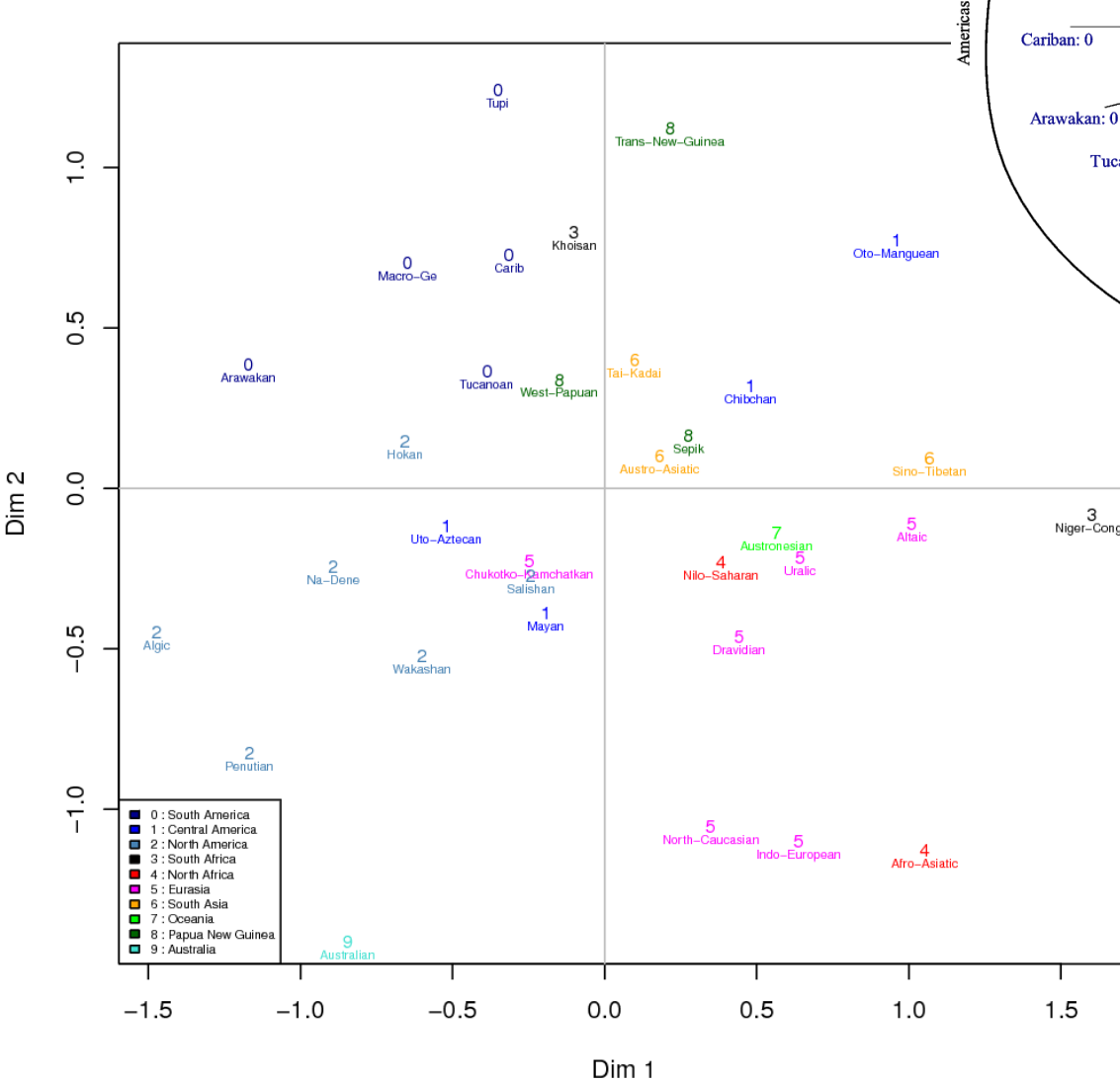
- How are lg.fams. **distributed** in this stability space:
 - Random? Clustered? Dispersed?
- Compare with randomly simulated 10,000 lg.fams. and max possible distance → **more compact** (smaller nearest-neighbor & mean distances, $p < 10^{-4}$)
- **Generalized Ripley's K function** → 10,000 Poisson process → empirical points outside 99% CI (whole min-max range) of simulated K values → **very strong clustering**

→ **stability of typology is much more similar across language families than expected**

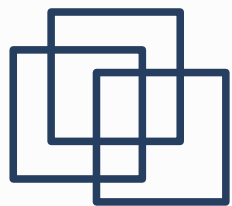


How does it look?

MDS for MrBayes : Binary : Ethnologue : All



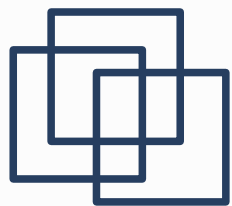
- **Americas** → cluster NS
- **NE Eurasia** → America
- Eurasia (?)
- Africa w/o Khoisan (?)
- **Very weak patterns!**



Variation among Families

Geography vs stability distances

- **Geographical distances** → waypoints between continents → min/max/mean/median/sd of distances between pairs of languages in different lg.fams.
- Weak → **moderate** mantel correlations $\min(r)=0.10$, $\max(r)=0.46$, **mean(r)=0.25**, most significant 0.01
- Families closer together in space tend to have similar stabilities:
 - **contact/areal** and/or
 - **deep genealogical** signal

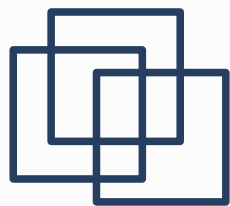


Variation among Families

Internal consistency of macro-areas

- **Randomization test:** for N language families, compare their stability distances with those of 10,000 permutations of N language families
- **Control for geography** (regression & pred 95% CI)
- **Americas** are clearly coherent (N & S but not C)
- **NE Eurasia** (Chukotko-Kamchatkan and Yukagir) + **Americas**
- ~ **“Core Eurasian”** (Altaic, Dravidian, Indo-European, Uralic and Caucasian)
- ~ **Austro-Tai & Papuan**
- This test **very conservative** and **signal very weak!**

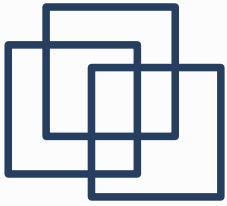
→ **signal of deep relationships and/or contact?**



Conclusions

and future work

- **Basic vocabulary** → mechanism (*frequency of use*) → intrinsic stability
- **Typology** → no such simple mechanism → *complex mixture* of intrinsic and context-dependent factors
- Looking at abstract, **higher-level properties** → maybe go beyond ~10,000 years horizon?
- Looking at **differences** between language families & areas → Neandertals, Denisovans (Dediu & Levinson, *submitted*)
- **Deep genealogical signal** in typological stability???



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.



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