

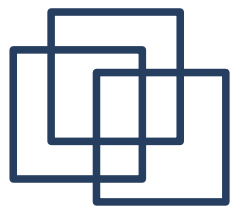
The stability of typological features

Dan Dediu

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

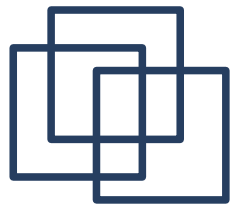
18 January, 2011
Edinburgh

Nijmegen, The Netherlands



Overview

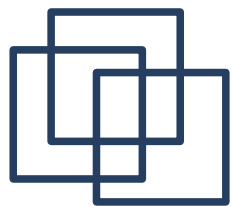
- Stability in biology
- Stability in language: basic vocabulary
- Stability in language: typology
- Four approaches
- Dediu (2011): Bayesian phylogenetics
- Comparing the approaches
- Conclusions



Stability in biology

(Nearly) Neutral Loci

- **Rate of evolution** (k) → *complex outcome*
 - **Lineage**-specific & -independent components
- *Neutral loci* → $k = u$ (mutation rate) → **molecular clock**
- *Nearly neutral loci* → $k \propto u/N$ (mutation ↔ population size)
 - **balance mutation - drift**
 - *Mutation rate* → **varies** across genome, species, time
 - DNA error correction, metabolism, life history, age, gender, environmental stress...

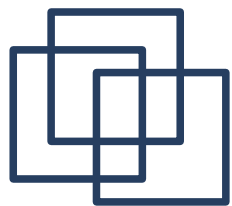


Stability in biology

Loci under selection

- **Selection** complicates picture even more:
 - **Purifying** → resist change
 - **Positive** → increased rates of evolution
- **Highly conserved** genes:
 - ribosomal RNA (*rRNA*) → all cellular life
 - *Pax6* → eye development
- **Very fast evolving** genes:
 - immune system
 - male reproductive biology
 - HARs, microcephaly genes, *FOXP2*





Stability in language

The basic vocabulary/cognate classes


- **Swadesh list** (100/200 meanings) → cognacy judgments in a set of languages → **cognate classes**

- **Rate** of cognate class change

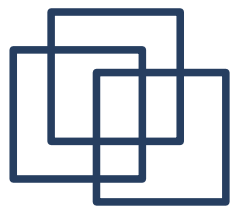
- **Bayesian phylogenetics**

- **Pagel *et al.* 2007: IE:**

- wide diffs in stability:
- most stable (“two”, “who”...): **< 1 replacement** per 10,000 years (half-life **> 10,000** years)
- most unstable (“dirty”, “guts”...): **< 9 (~750yr)**
- present-day **frequency of use** (Spanish, English, Russian, Greek): **$r = -0.37, p < 0.0001$**



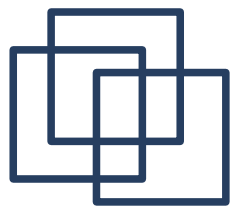
Language	"Bone"	Cognacy	Binary
Proto-Austronesian	*CuqelaN	1	10000
Paiwan	tsuqela	1	10000
Itbayaten	tuqgan	1	10000
Tagalog	butó	5	00001
Bare'e	wuku	2	01000
Mangarrai	toko	2	01000
Numfor	kor	3	00100
Motu	turia	4	00010
Fijian (Bau)	sui-na	4	00010
Tongan	hui	4	00010
Maori	iwi	4	00010



Stability in language

The basic vocabulary/cognate classes

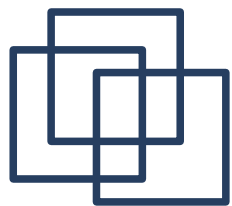
- **Pagel & Meade (2006)**: compare IE & Bantu
 - Bantu changes faster than IE
 - rates correlate across families: $r = 0.28, p < 0.03$
- **Greenhill *et al.* (2010)**: IE & Austronesian:
 - Austronesian faster than IE
 - rates correlate: $\rho = 0.37, p < 0.0001$
- **Pagel (2009)**: IE vs Starostin (14 lg fams) $\rightarrow r = 0.65$
 - \rightarrow meanings seen to have an **intrinsic** (i.e., cross-family & area) stability
 - \rightarrow accessible through very **different methods**
 - \rightarrow partially explained by **frequency of use**



Stability in language

What about typological features?

- Very interesting for:
 - typology
 - linguistic theory
- **Genetic biasing hypothesis** (Dediu & Ladd, 2007):
 - genetically “anchored” cultural features → **tend** to change slower
- **Previous work:**
 - (1) explicitly phylogenetic framework
 - (2) “ad-hoc”/non-phylogenetic concepts

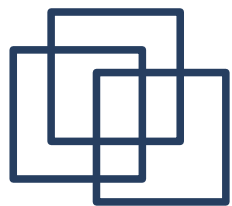


Stability in language

1. Phylogenetic framework

- **Dunn et al. (2005, 2008)**: Oceanic & Papuan → typology preserves info beyond comparative method
- **Hunley et al. (2008)**: typology resists borrowing better than genetics!
- **Greenhill et al. (2010)**: IE & Austronesian →
 - vocabulary more tree-like
 - similar rates for typology & vocabulary
 - very weak corr b/w fams: $\rho = 0.17$, $p = 0.1$→ no “intrinsic” stability? more stable than vocabulary?
→ could be the **pattern** which is more stable (genetics)

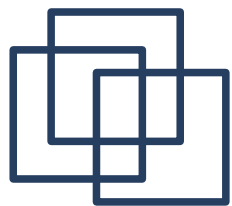
Too few language families to allow generalization!



Stability in language

2. Non-Phylogenetic/ad-hoc concepts

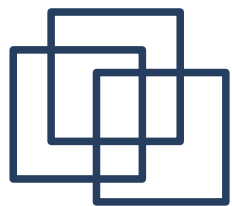
- Devised mostly by typologists
- *Literature survey* (**Dediu & Cysouw, *in preparation***)
- **Criteria:**
 - *published*
 - *concept of stability* ~ “easiness” with which features change values across time, under the influence of various processes
 - *quantifiable, objective and repeatable*
 - *many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



Stability in language

Cysouw, Albu & Dress (2008)

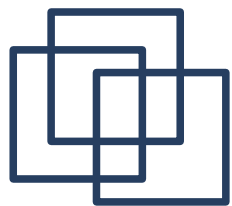
- **Consistency** of individual features vs overall patterns
 - **Typological distance** $D_F(L_1, L_2) = \begin{cases} 0, & F_{L_1} = F_{L_2} \\ 2, & F_{L_1} \neq F_{L_2} \\ 1, & F_{L_1} \vee F_{L_2} \text{ undef} \end{cases}$
 - For N features F_i and M languages $L_j \rightarrow$
 - dist matrices 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

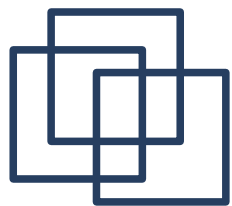
Wichmann & Holman (2009)

- 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

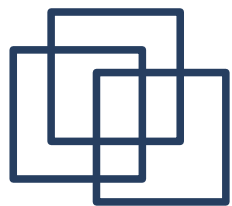
$$S_F = \frac{R_F - U_F}{1 - U_F}$$



Stability in language

Dediu (2011): The primary data

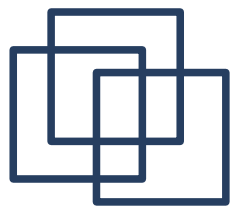
- **Phylogenetic** method → **Bayesian (model-based)**
 - As many language families and features as possible
 - **Primary data**
 - *Features & values* from **WALS** (www.wals.info)
 - **Selection and recoding:**
 - **UNORDERED** ($v_i \leftrightarrow v_j$)
 - **RANKED** ($v_1 \leftrightarrow v_2 \dots \leftrightarrow v_n$)
 - **CUSTOM** (user-defined transition matrix)
 - **Binary aspects:** e.g. *Tone* → binary *Tone1* (no tone vs all) & *Tone2* (complex tone vs all)
 - **Polymorphic** and **Binary** features
-



Stability in language

Dediu (2011): The historical classifications

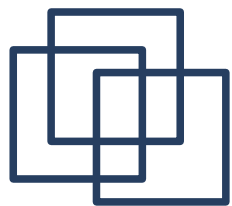
- Control for the effects of historical classification
- Classification in **WALS**:
 - 3 levels (“Family”, “Genus”, “Language”): e.g.,
[Indo-European](#) > [Romance](#) > [Romanian](#)
 - “Subfamily” - sporadic(e.g. Niger-Congo) → ignored
- The **Ethnologue** (www.ethnologue.com):
 - generally, much more resolved
 - number of levels varies from 1 (isolates) to 14 (Austronesian)
 - [Indo-European](#) > [Italic](#) > [Romance](#) > [Eastern](#) > [Romanian](#)
- **Not independent**: WALS inspired by 14 ed ethnologue



Stability in language

Dediu (2011): Stability estimates

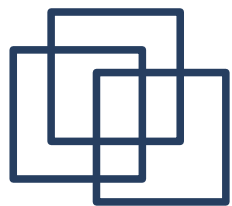
- Control for assumptions behind estimates & software
- **Two software packages:**
- **MrBayes 3.1.2** (<http:// mrbayes.csit.fsu.edu/index.php>)
 - widely used & very flexible Bayesian phylogeny
 - rates → **gamma** distribution of rates across sites
 - language families → **constraints** on topology
 - **outgroups**: language isolates
 - **MC³**: 3+1 chains, 5,000,000 generations → 1,000 sampling freq + first 1,000 samples burn-in
 - **convergence**: log-likelihood plots & Potential Scale Reduction factor



Stability in language

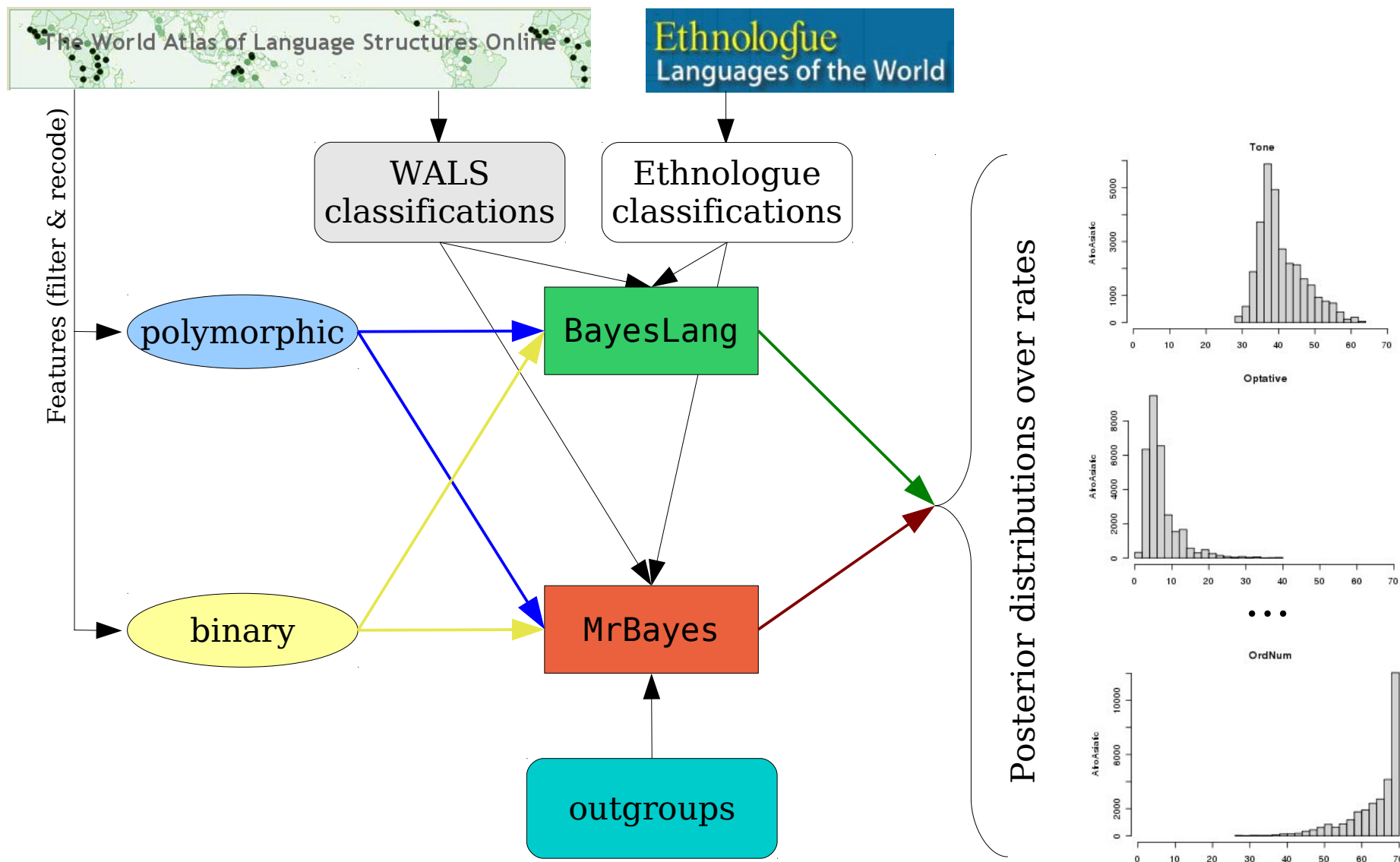
Dediu (2011): Stability estimates

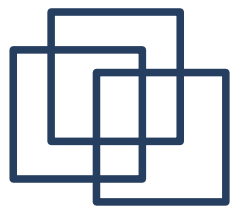
- **BayesLang** (<http://www.mpi.nl/people/dediu-dan/tool>)
 - written from scratch in cross-platform C++; GPL
 - standard Bayesian phylogenetics
 - multi-threaded MC³, customized Nexus format
 - takes rooted trees and custom transition matrix
 - rates → **min number of changes** (parsimony)
 - language families → as **rooted trees**
 - **MC³**: 6+1 chains, 5,000,000 generations, first 1,000,000 generations burn-in
 - **convergence**: log-likelihood plots



Stability in language

Dediu (2011): Workflow





Stability in language

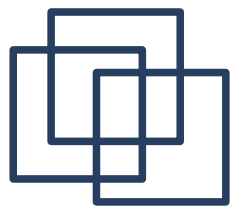
Dediu (2011): Datasets & outgroups

<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**

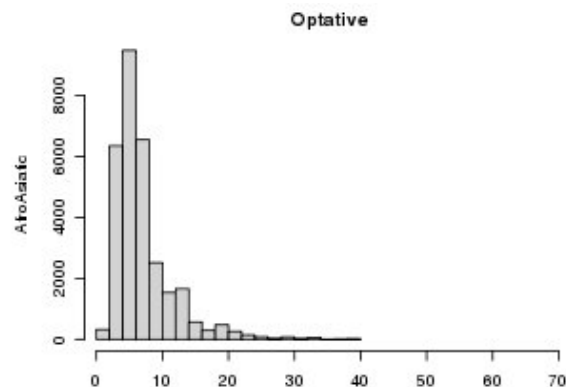


Stability in language

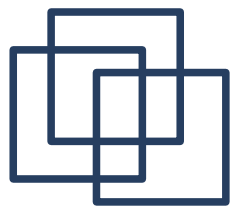
Dediu (2011): Output

- **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



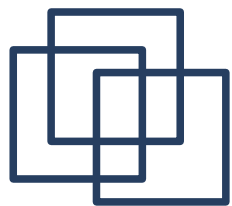
- Absolute rates cannot be compared across trees, methods & datasets → conversion to **ranks**
- **Summarize** rank estimate distributions: **mean** & **median**



Stability in language

Dediu (2011): Results

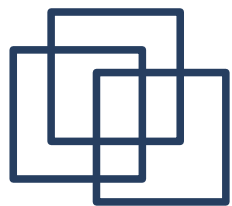
- Family-level distibs → uni-, bi- and (rarely) multi-modal
- **Summaries** by means & medians: $0.94 \leq r \leq 0.98, p < 10^{-6}$
- **WALS & Ethnologue**: $0.96 \leq r \leq 0.99, p < 10^{-6}$
- **Outgroup** → negligible: $0.49 \leq r \leq 0.92, p < 10^{-6}; \bar{r} = 0.78$
PC₁: 79% variance
- **Binary**: $0.59 \leq r \leq 0.98, p < 10^{-8}; \bar{r} = 0.78; PC_1: 81.4\%$
 - *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**



Stability in language

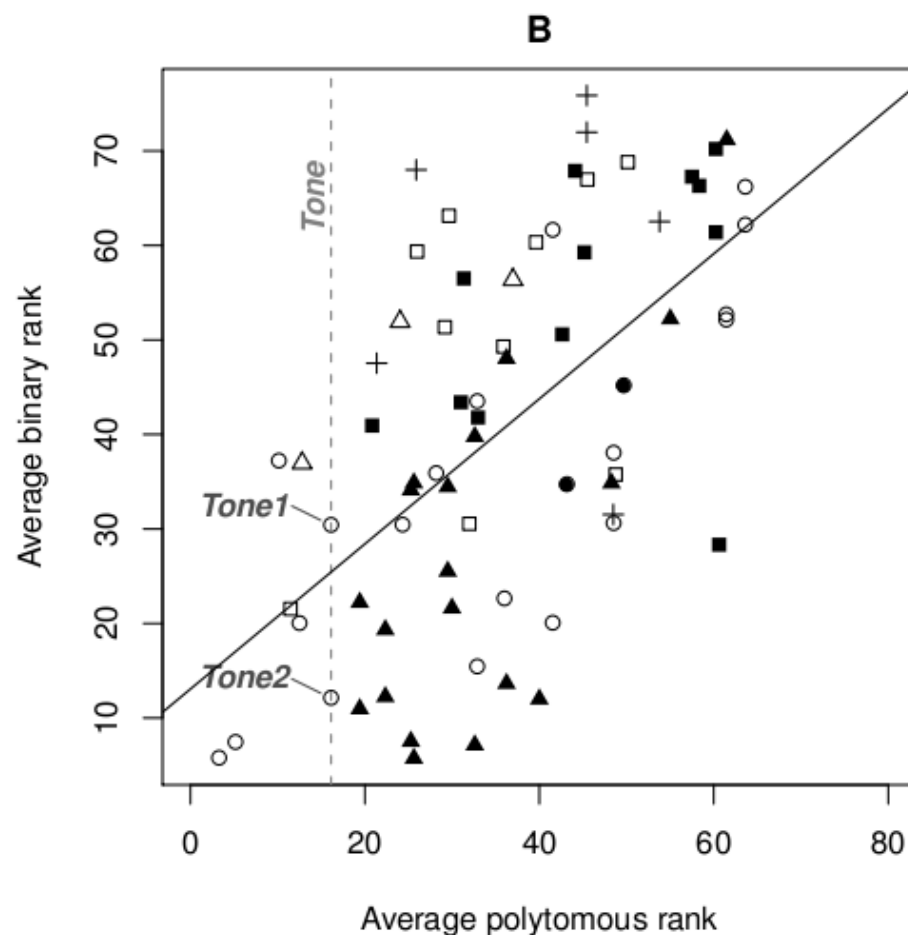
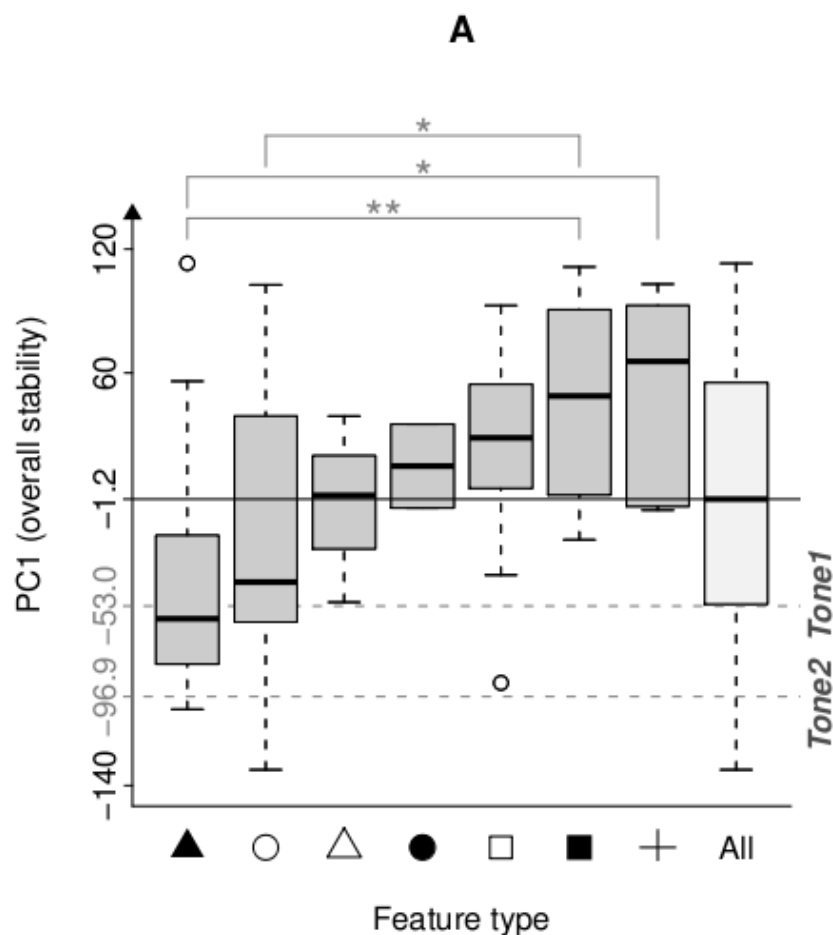
Dediu (2011): Results (2)

- **Binary - Poly**: complex → the two aspects might have different stabilities!
- $0.18 \leq r \leq 0.99$; $\bar{r} = 0.61$, median $p = 6.5 \times 10^{-9}$
- PC_1 (agreement): 67.4%, PC_2 (bin vs poly): 16.1%
- *Tone*: **very** stable as **poly**: $t_{56} = 9.07$, $p = 1.35 \cdot 10^{-13}$
- *Tone2*: **very** stable as **bin** ($t_{70} = 12.04$, $p < 2.2 \cdot 10^{-16}$) & **overall** ($t_{70} = 12.27$, $p < 2.2 \cdot 10^{-16}$)
- *Tone1*: **relatively** stable as **bin** ($t_{70} = 4.35$, $p = 6.7 \cdot 10^{-9}$) & **overall** ($t_{70} = 4.35$, $p = 4.5 \cdot 10^{-5}$)

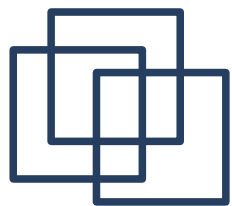


Stability in language

Dediu (2011): Results (3)



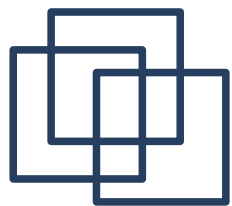
▲ = Word Order, ○ = Phonology, △ = Nominal Syntax, ● = Morphology, □ = Verbal Categories, ■ = Nominal Categories, + = Simple Clauses, and All = all types of features combined.



Stability in language

Dediu (2011): Conclusions & caveats

- *Tone* (*Tone1* & *Tone2*) seem to be quite stable
- Stability estimates seem to correlate across methods, datasets ... → maybe they reflect something **intrinsic**?
- **Caveats:**
 - *Trees!!!*
 - Data *quality* & *quantity*
 - *Outgroups*
 - “*Averaging out*” → looking at variation as well
- However, it's **first step** → supports the genetic biasing hypothesis of tone
- Suggests **new candidates**: *FrRoundV*, *NMPron...*

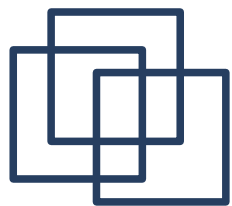


Stability in language

Comparing all four methods

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

ρ \ r	CM	CC	CR	D	P ₁	P ₂	W
CM	-	0.08 0.534	0.25* 0.047	-0.08 0.523	0.29* 0.021	-0.03 0.838	0.36** 0.003
CC	0.10 0.428	-	0.82** < $2.2 \cdot 10^{-16}$	0.65** $8.34 \cdot 10^{-9}$	0.33** 0.008	0.21 0.108	0.12 0.353
CR	0.12 0.346	0.83** < $2.2 \cdot 10^{-16}$	-	0.73** $7.78 \cdot 10^{-12}$	0.49** $5.34 \cdot 10^{-5}$	0.17 0.189	0.21 0.102
D	-0.17 0.177	0.60** $1.35 \cdot 10^{-7}$	0.68** < $2.2 \cdot 10^{-16}$	-	0.45** $2.20 \cdot 10^{-4}$	0.20 0.127	0.31* 0.014
P ₁	0.15 0.254	0.42** $7.02 \cdot 10^{-4}$	0.54** $4.20 \cdot 10^{-6}$	0.52** $1.32 \cdot 10^{-5}$	-	0.49** $4.89 \cdot 10^{-5}$	0.41** $8.63 \cdot 10^{-4}$
P ₂	0.23 0.075	0.33** 0.008	0.18 0.154	0.29* 0.020	0.48** $7.28 \cdot 10^{-5}$	-	0.16 0.215
W	0.28* 0.028	0.23 0.072	0.25* 0.048	0.39** 0.001	0.51** $1.83 \cdot 10^{-5}$	0.46** $1.87 \cdot 10^{-4}$	-



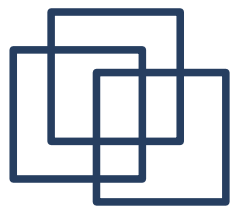
Stability in language

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

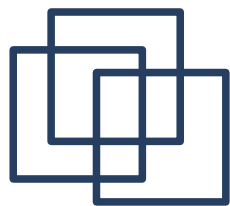
- *Tend to agree more than disagree* → **stability** has an **intrinsic component**
- *Unexpected results*: **D** (strongly phylogenetic) was expected to agree with **W** (conceptually phylogenetic), but **D** agrees with **CC** and **CR** (non-genealogical)
- The agreement between all methods (PC1)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



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 **multivariate patterns of typological values** → maybe go beyond ~10,000 years horizon?
- Looking at **differences** between language families & areas → Neandertals, Denisovans (Dediu & Levinson, *in preparation*)
- What about **selection**? Do we have a **null model** (“(nearly) neutral linguistic evolution”)? If not, how could we get one?



THANK YOU!

- Dediu, D., & Ladd, D. R. (2007). Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *PNAS*, **104**, 10944-10949.
- Dediu, D. (2011). A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proc. Royal Society B*. Advance online publication. doi:10.1098/rspb.2010.1595.
- Dediu, D. & Cysouw, M. (in preparation). Is the concept of “stability” meaningful in linguistic typology?
- Dediu, D. & Levinson, S. (in preparation). On the antiquity of language: the reinterpretation of Neandertal linguistic capacities and its consequences.
- Cysouw, M., Albu, M., and Dress, A. (2008). Analyzing feature consistency using dissimilarity matrices. *STUF*, **61**:263–279.
- Dunn, M., Levinson, S. C., Lindstrm, E., Reesink, G., and Terrill, A. (2008). Structural phylogeny in historical linguistics: Methodological explorations applied in island melanesia. *Language*, **84**:710–759.
- Dunn, M., Terrill, A., Reesink, G., Foley, R. A., and Levinson, S. C. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science*, **309**:2072–2075.
- Greenhill, S. J., Atkinson, Q. D., Meade, A., and Gray, R. D. (2010). The shape and tempo of language evolution. *Proc. R. Soc. B*.
- Hunley, K., Dunn, M., Lindstrm, E., Reesink, G., Terrill, A., Healy, M. E., Koki, G., Friedlaender, F. R., and Friedlaender, J. S. (2008). Genetic and linguistic coevolution in northern island melanesia. *PLoS Genet*, **4**(10):e1000239.
- Pagel, M. (2009). Human language as a culturally transmitted replicator. *Nat Rev Genet*, **10**:405–415.
- Pagel, M., Atkinson, Q. D., and Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, **449**:717–721.
- Pagel, M. and Meade, A. (2006). Estimating rates of lexical replacement on phylogenetic trees of languages. In Forster, P. and Renfrew, C., eds, *Phylogenetic methods and the prehistory of languages*, 173–182. McDonald Institute: UK.
- Parkvall, M. (2008). Which parts of language are the most stable? *STUF*, **61**:234–250.
- Wichmann, S. and Holman, E. W. (2009). *Assessing Temporal Stability for Linguistic Typological Features*. LINCOM Europa:M'nchen.

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