

**Workshop on Language Variation
and Change and Cultural Evolution:
February 12-14, 2015, York, UK**

Algorithmic generation of random languages argues for syntax as a source of phylogenetic information

**Giuseppe Longobardi*[^], Andrea Ceolin*, Aaron Ecaj*, Cristina Guardiano[°], Monica
Alexandrina Irimia*, Nina Radkevich*, Dimitris Michelioudakis*, Luca Bortolussi[^], Andrea
Sgarro[^]**

***University of York**

[°]Università di Modena e Reggio Emilia

[^]Università di Trieste

Language comparison and phylogenetic reconstruction:

- 1) Correspondence problem
- 2) Metric problem
- 3) Probability problem

but in particular...

Language comparison and phylogenetic reconstruction:

4) **Globality problem**



The classical comparative method and modern lexical/morpho-phonological methods have addressed problems 1-3 (with some difficulties):

- 1) Correspondence problem: Swadesh-list “meaning” comparison (e.g. Dyen et al. 1992), “root” comparison (e.g. Gray and Atkinson 2003)
- 2) Metric problem: “distances” (e.g. Dyen et al. 1992), “characters” (e.g. Ringe et. al 2002)
- 3) Probability problem (e.g. Ringe 1992, Nichols 1996)

No solution has been found for the **Globality problem**.

Reason: Lexical and morphological characters are not comparable across established families (e.g. all etymological unrelated characters, all maximum distances).

Solution: Exploring different domains, e.g. Syntax (Guardiano and Longobardi 2005, Longobardi and Guardiano 2009)



Available online at www.sciencedirect.com



Lingua xxx (2009) xxx–xxx

Lingua

www.elsevier.com/locate/lingua

Evidence for syntax as a signal of historical relatedness

Giuseppe Longobardi^{a,*}, Cristina Guardiano^b

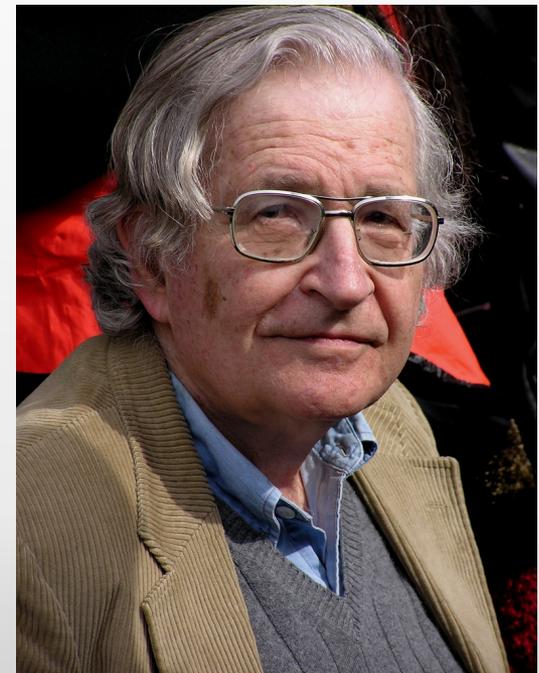
^a*Laboratorio di Linguistica e antropologia cognitiva, DSA, Università di Trieste, Italy*

^b*Dipartimento di Scienze del Linguaggio e della Cultura, Università di Modena e Reggio Emilia, Italy*

Received 15 January 2007; received in revised form 9 September 2008; accepted 9 September 2008

Problems 1,2 and 4 are solved through the notion of “syntactic parameter” (Chomsky 1981, Baker 2001).

Parameters are species-invariant, ultimately biological, options: the same open choices are presupposed by every language (i.e. by every infant) and environmentally set during the acquisition



● Parametric Minimalism: the Principles&Schemata model

- The description of parameters as an invariant set of innate questions might seem to imply that large amounts of irrelevant information must be stored in the minds of language learners. This is problematic (Berwick & Chomsky 2011, Boeckx 2011)
- However, it is possible to reformulate this description in terms of a Principles&Schemata approach (Longobardi 2014), where there **are** only a small number of parameter schemata preset at S_0 , which learners instantiate as needed, eventually reaching a set of fixed parameters at S_s

- 1) Correspondence problem: parameters are **unambiguous**
- 2) Metric problem: parameters are **discrete** (binary values)
- 4) Globality problem: parameters are **universal**

3) Probability problem



Parametric Comparison Method (PCM) (Guardiano and Longobardi 2005, Longobardi and Guardiano 2009)

Languages are encoded as lists of binary parameters (+,-).

Parametric Comparison Method (PCM) (Guardiano and Longobardi 2005, Longobardi and Guardiano 2009)

Problem: parameters are not **independent**, but there are **implications** which make some parameter values predictable.

Grammaticalized definiteness and enclitic articles

| | English | Norwegian | Russian |
|----------------------------------|---------|-----------|---------|
| p10: gramm. Def. (articles) | + | + | - |
| p14: enclitic articles (+p10) | - | + | ? |

Grammaticalized definiteness and enclitic articles

| | English | Norwegian | Russian |
|----------------------------------|---------|-----------|---------|
| p10: gramm. Def. (articles) | + | + | - |
| p14: enclitic articles (+p10) | - | + | 0 |

Sample parameters:

65 syntactic parameters from the Nominal Domain (DPs)

Sample languages:

21 IE languages (5 Romance, 5 Germanic, 5 Slavic, 3 Indo-Iranian, 2 Celtic, 1 Greek)

3 Finno-Ugric languages (Finnish, Estonian, Hungarian)

2 Altaic languages (Turkish, Buryat)

2 Semitic languages (Hebrew, Arabic)

2 Basque varieties

2 Chinese (Mandarin, Cantonese)

1 Wolof

Coding the data: **Characters** or **Distances**?

Reasons why character-based approaches are problematic for syntactic comparison:

- 1) Homoplasy and borrowing are possible
- 2) Lack of information about directionality of change
- 3) **Shared innovations are not always detectable (e.g. implicated values, low number of parameters)**

Coding the data: **Characters** or **Distances**?

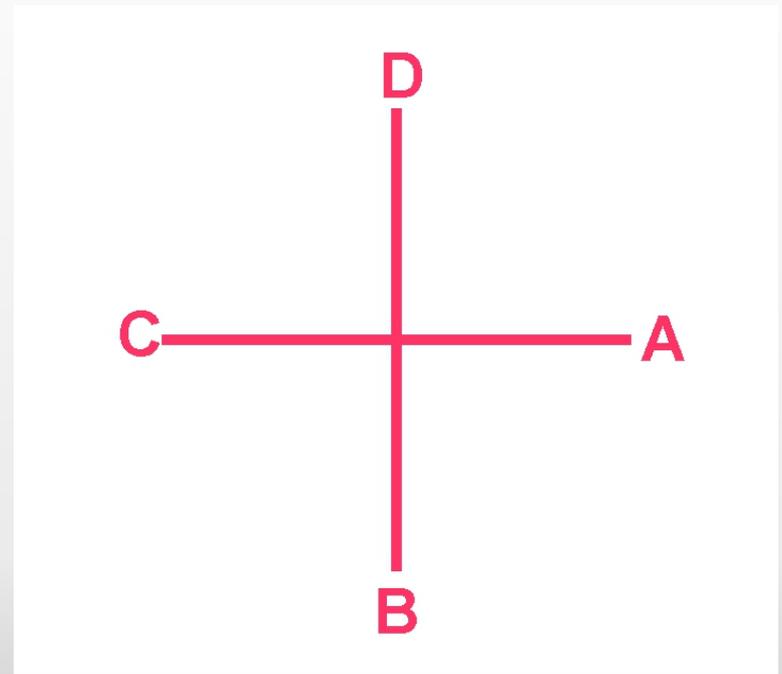
Consider the following example:

| | |
|---|--------|
| A | ++++-- |
| B | ----++ |
| C | ----- |
| D | ----- |

Coding the data: **Characters** or **Distances**?

| | |
|---|----------|
| A | ++++-- |
| B | ----++++ |
| C | ----- |
| D | ----- |

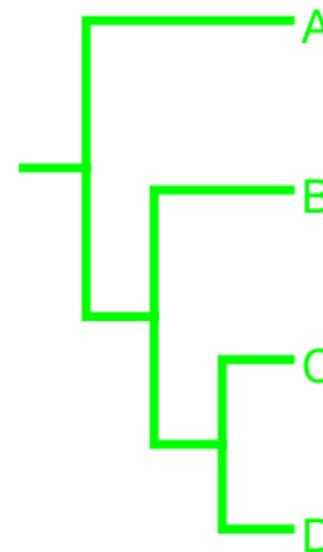
PAUP*



Coding the data: **Characters** or **Distances**?

KITSCH or UPGMA

| | |
|---|---------|
| A | ++++-- |
| B | ----+++ |
| C | ----- |
| D | ----- |



Coding the data: **Characters** or **Distances**?

Now consider the following example:

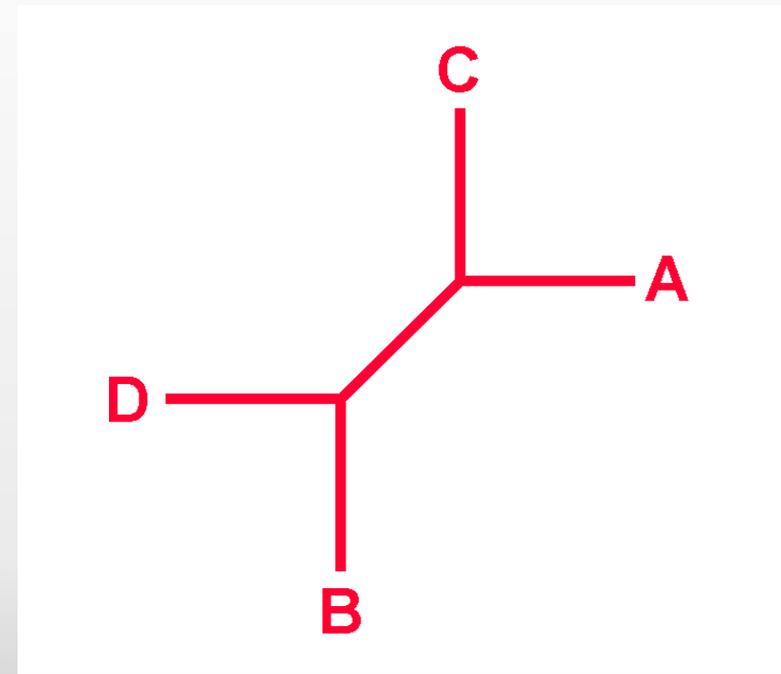
| | |
|---|------------------|
| A | ++++----- |
| B | -----++++ |
| C | ----- |
| D | -----+ (circled) |

(borrowing, parallel
development, resetting,
etc. etc.)

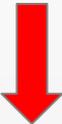
Coding the data: **Characters** or **Distances**?

| | |
|---|-----------|
| A | ++++----- |
| B | -----++++ |
| C | ----- |
| D | -----+ |

PAUP*

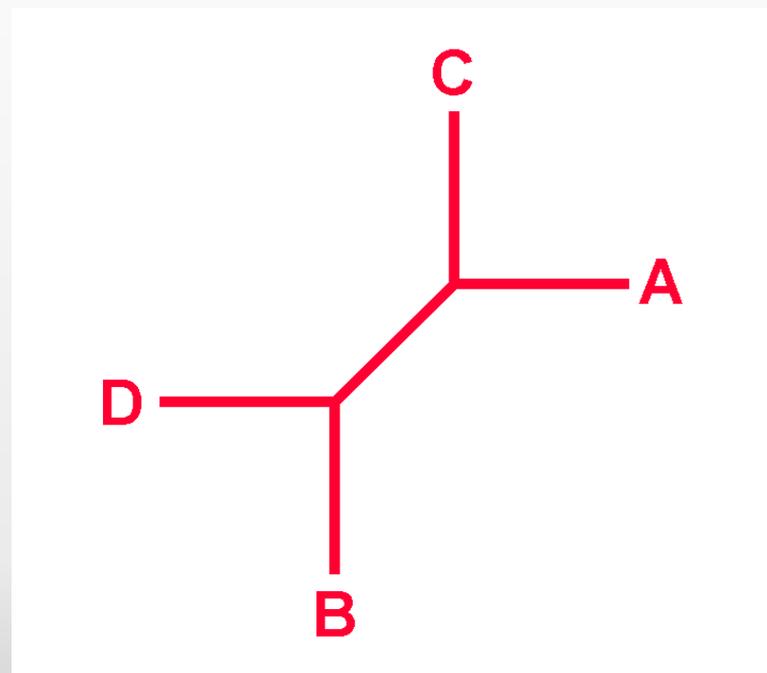


Coding the data: **Characters** or **Distances**?



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A | + | + | + | + | - | - | - |
| B | - | - | - | - | + | + | + |
| C | - | - | - | - | - | - | - |
| D | - | - | - | - | - | - | + |

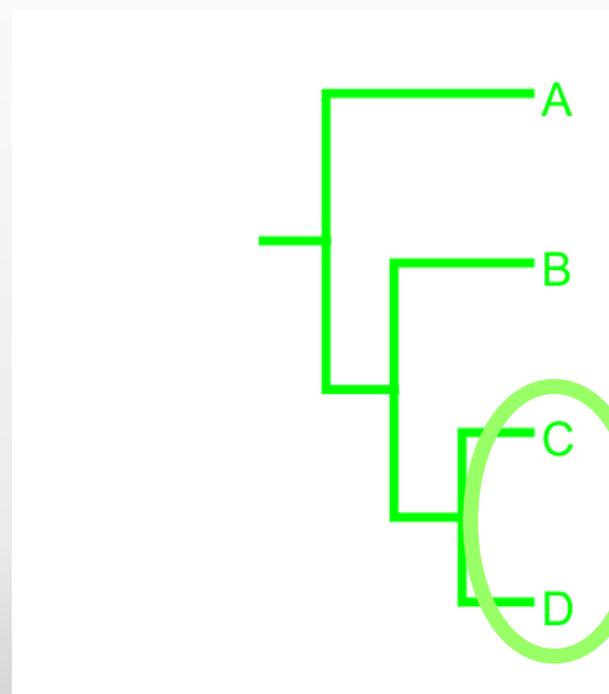
PAUP*



Coding the data: **Characters** or **Distances**?

KITSCH or UPGMA

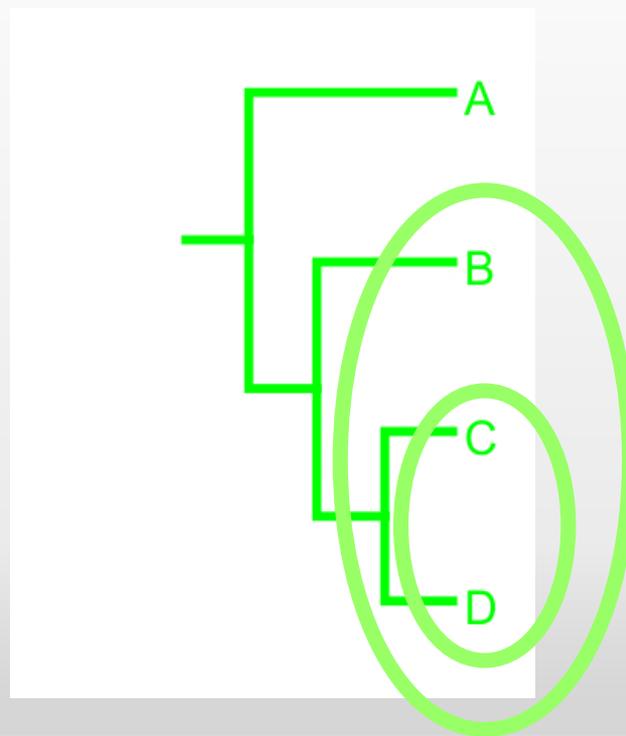
| | |
|---|-----------|
| A | ++++----- |
| B | -----++++ |
| C | ----- |
| D | -----+ |



Coding the data: **Characters** or **Distances**?

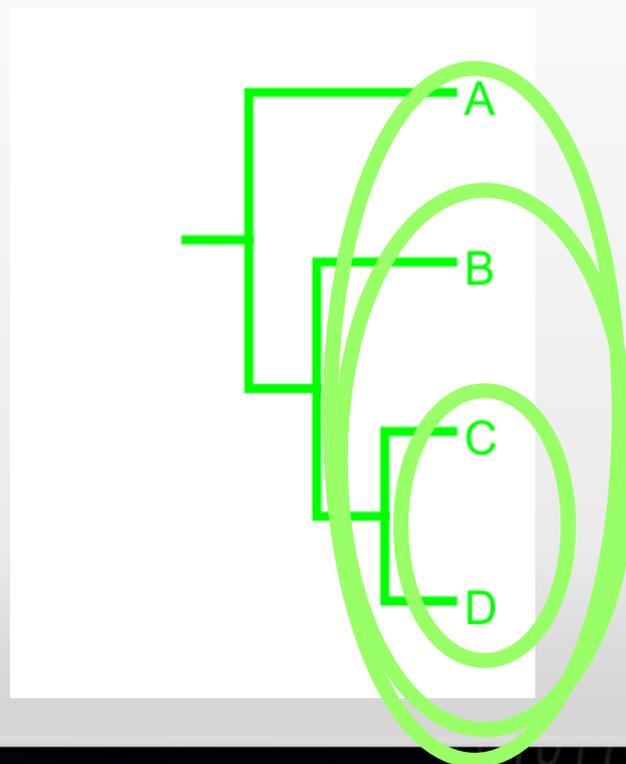
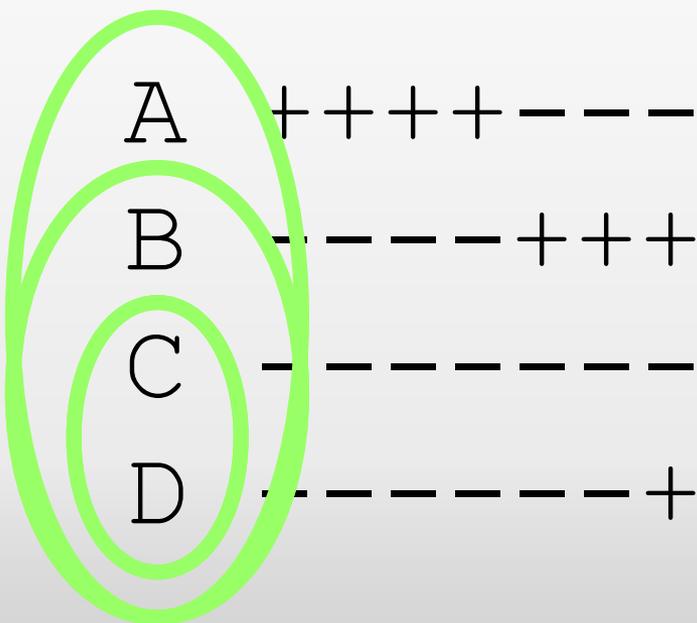
| | |
|---|-----------|
| A | ++++----- |
| B | -----++++ |
| C | ----- |
| D | -----+ |

KITSCH or UPGMA



Coding the data: **Characters** or **Distances**?

KITSCH or **UPGMA**



Coding the data: **Characters** or **Distances**?

Reasons why distances-based programs are better for syntactic comparison:

- 1) More resistant to single borrowing/parallel development effects
- 2) No need of notions like directionality of change/ancestral values/outgroups (when using the molecular clock hypothesis)
- 3) They take into account also parsimony uninformative values (and thus **shared retentions**, like in lexicostatistics)

How to choose a distance measure?

Since we have a lot of '0' values, we cannot rely on a simple Hamming distance.

We can use a **Jaccard-Tanimoto distance** between “comparable” values:

$$\delta(A,B) = d(A,B) / [d(A,B) + i(A,B)]$$

= differences / identities + differences

E.g.: Italian-English: (35 id., 6 diff.) $\delta = 6 / 41 = 0.146$

Are these distances **significant** from a statistical
viewpoint?

Comparison of “real” distances versus “randomly
generated” distances (Bortolussi et al. 2011).

Bortolussi et al. (2011): Sampling over a uniform distribution of languages

| Bortolussi et al. (2011) | L1 | L2 | L3 | L4 | Probability of P_ = + |
|--------------------------|-------------|-------------|-------------|-------------|-----------------------|
| P1 | + | + | + | - | 0.75 |
| P2 (only if +P1) | + | + | - | 0 | 0.67 |
| P3 (only if +P2) | + | - | 0 | 0 | 0.5 |
| Probability of L_ = + | 0.25 | 0.25 | 0.25 | 0.25 | |

Problem: since the sample is made of mostly IE languages, this kind of sampling generates IE-like languages.

A new sampling algorithm: assumption of a uniform distribution of **parameters**

| Our algorithm | L1 | L2 | L3 | L4 | Probability of P_ = + |
|-----------------------|-------|------|------|-----|-----------------------|
| P1 | + | + | + | - | 0.5 |
| P2 (only if +P1) | + | + | - | 0 | 0.5 |
| P3 (only if +P2) | + | - | 0 | 0 | 0.5 |
| Probability of L_ = + | 0.125 | 0.25 | 0.25 | 0.5 | |

Independent parameters are first assigned a value by chance.
Following parameters are checked for implications.

Pseudo code:

```
given [P1, P2(if and only if +P1), P3....]
set P1 (either + or -, 0.5 chance)
check the implications of P2
if P2 can be set,
    set P2 (either + or -, 0.5 chance)
else,
    assign 0 to P2 and move to P3
check the implications of P3
if P3 can be set....
...
return the string generated
```

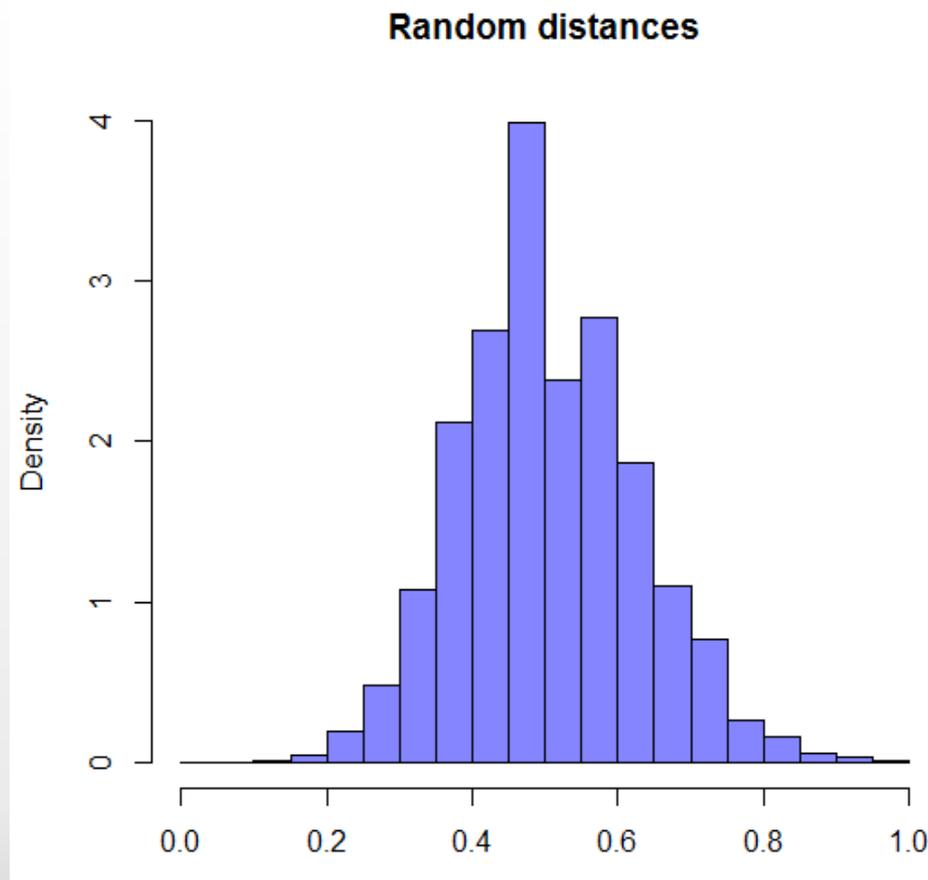
Random sample

1000 random languages

~500.000 random pairs

Mean: 0.510

Median: 0.5



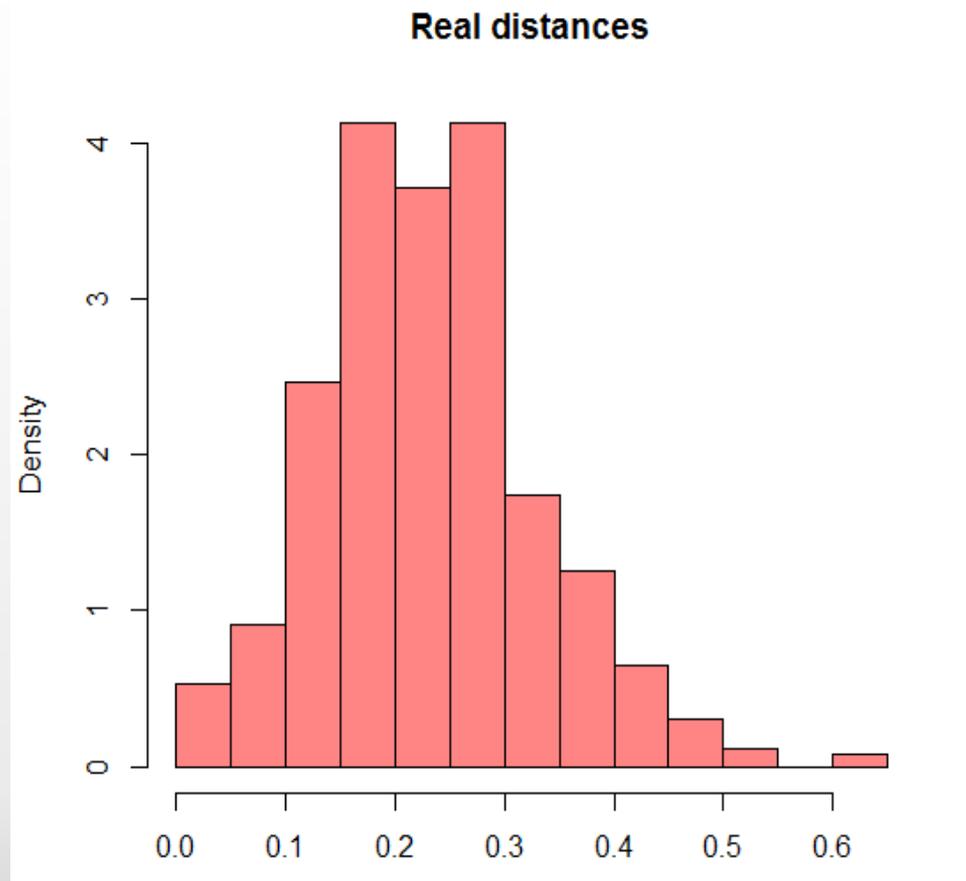
Real sample

33 real languages

528 real pairs

Mean: 0.234

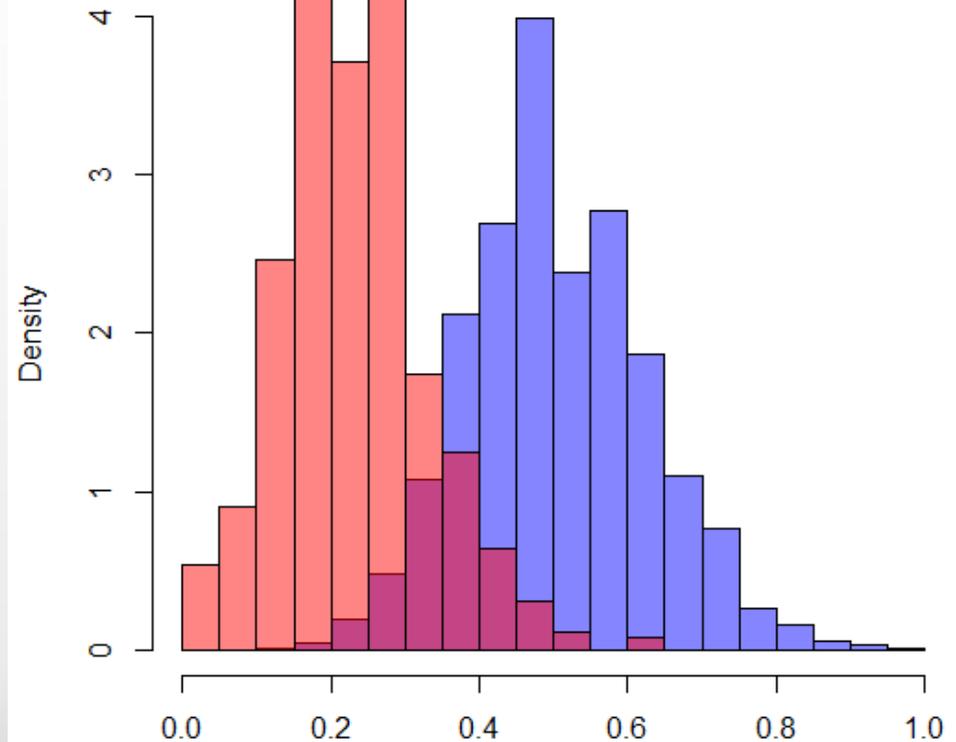
Median: 0.226



Mann-Whitney U test = $p < 2.2 * e^{-16}$

The two datasets are unlikely to
be drawn from the same distribution

Real vs Random distances

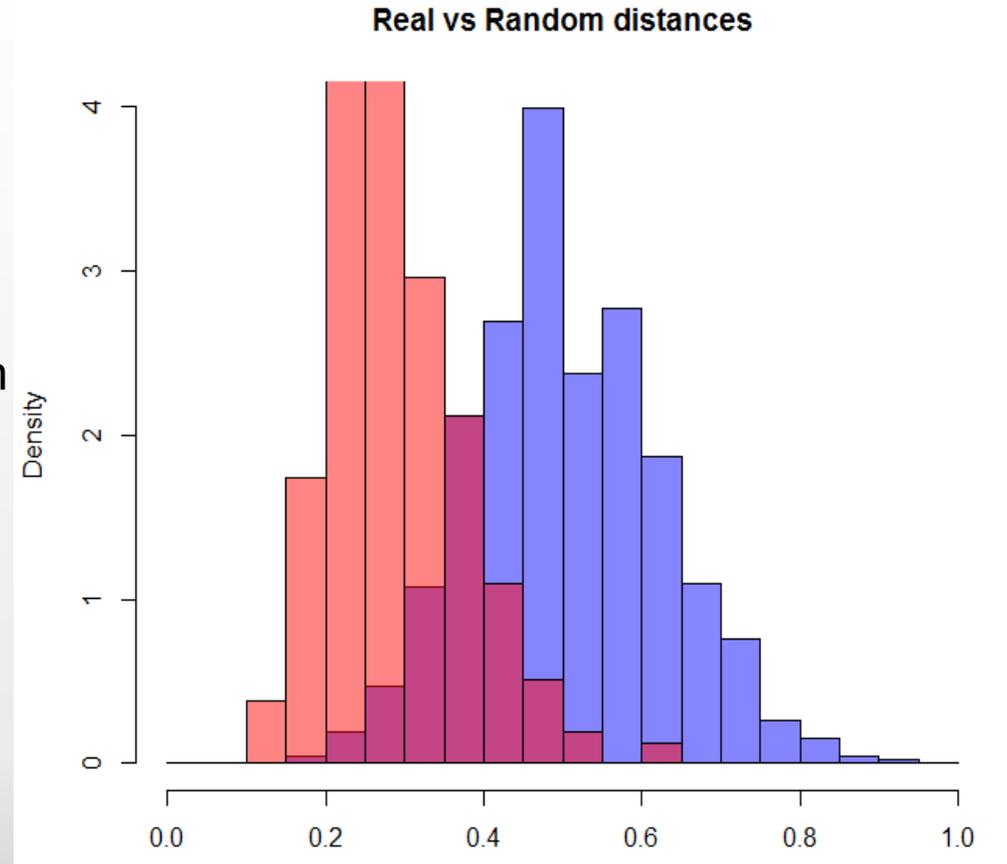


Family-external pairs only

Mann-Whitney U test = $p < 2.2 * e^{-16}$

The two datasets are unlikely
to be drawn from the same distribution

(**Anti-Babelic principle**, Guardiano
and Longobardi 2005)



Possible explanation?

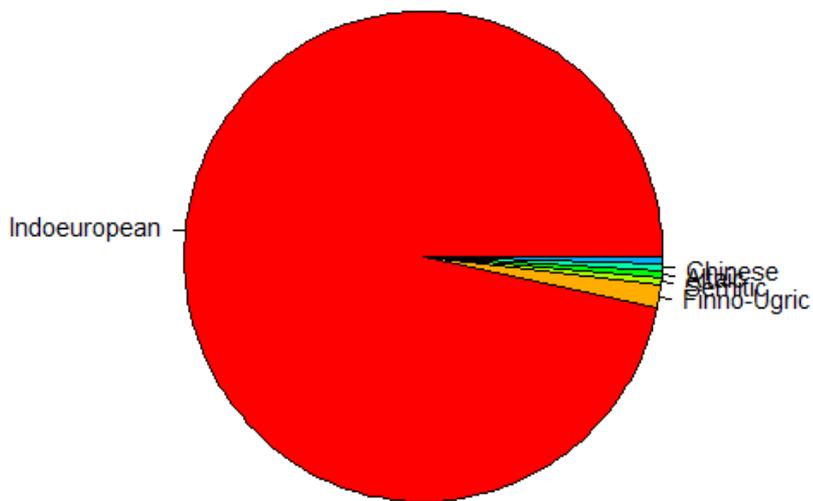
| Chance threshold | Distance value | Pairs expected by chance | Pairs under the threshold |
|------------------|----------------|--------------------------|---------------------------|
| $P=10^{-4}$ | 0.111 | 0.05 | 42 |
| $P=10^{-3}$ | 0.174 | 0.528 | 146 |
| $P=10^{-2}$ | 0.250 | 5.28 | 310 |

Possible explanation?

| Chance threshold | Distance value | Pairs expected by chance | Pairs under the threshold |
|------------------|----------------|--------------------------|---------------------------|
| $P=10^{-4}$ | 0.111 | 0.05 | 42 |
| $P=10^{-3}$ | 0.174 | 0.528 | 146 |
| $P=10^{-2}$ | 0.250 | 5.28 | 310 |

History as the driving force?

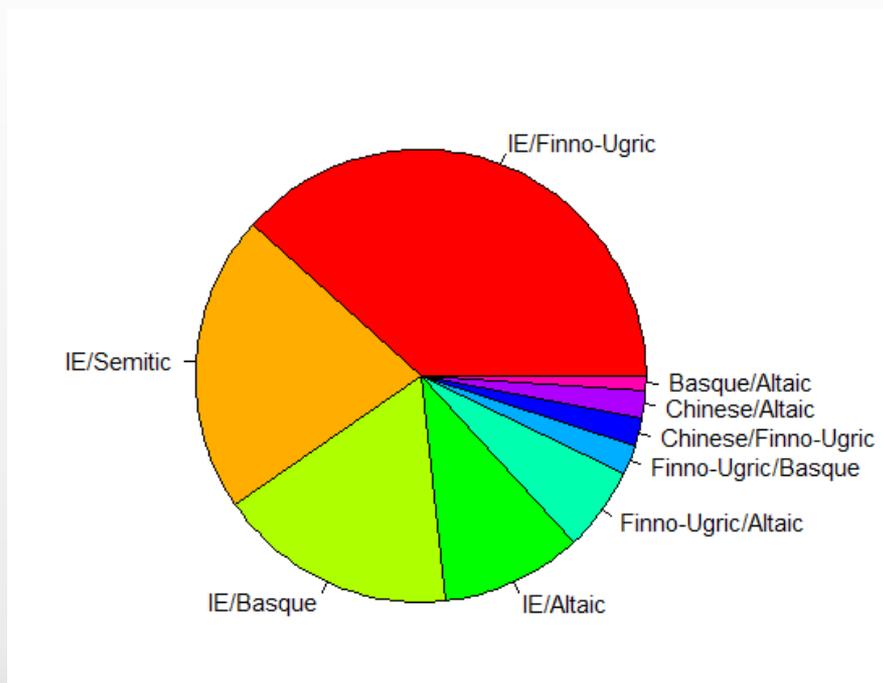
Family-internal pairs (211)



| | Under threshold | Total | Percentage |
|---------------|-----------------|-------|------------|
| Indo-European | 204 | 210 | 97.14% |
| Finno-Ugric | 3 | 3 | 100.00% |
| Semitic | 1 | 1 | 100.00% |
| Basque | 1 | 1 | 100.00% |
| Altaic | 1 | 1 | 100.00% |
| Chinese | 1 | 1 | 100.00% |

History as the driving force?

Family-external pairs (99)

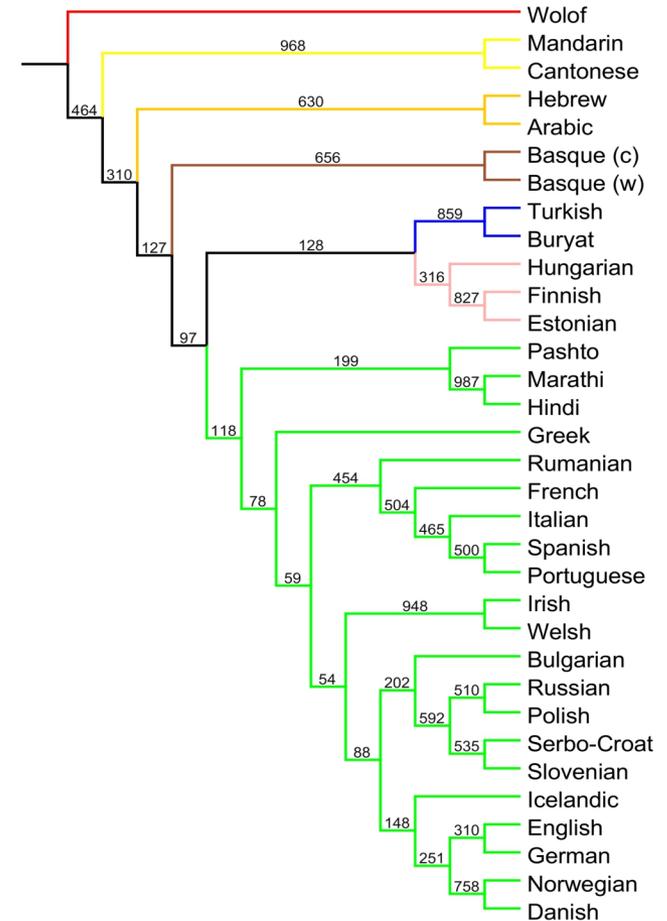


| | Under threshold | Total | Percentage |
|---------------------|-----------------|-------|------------|
| Finno-Ugric/Altaic | 6 | 6 | 100.00% |
| IE/Finno-Ugric | 38 | 63 | 60.32% |
| IE/Semitic | 21 | 42 | 50.00% |
| Chinese/Altaic | 2 | 4 | 50.00% |
| IE/Basque | 17 | 42 | 40.48% |
| IE/Altaic | 10 | 42 | 23.81% |
| Finno-Ugric/Basque | 2 | 6 | 33.33% |
| Finno-Ugric/Chinese | 2 | 6 | 33.33% |
| Altaic/Basque | 1 | 4 | 25.00% |

Phylogenetic tree (KITSCH, bootstrapped, 1000 resamples)



Supported by genetic evidence (cf. Pagani et al. 2015, in prep.)



Conclusions:

- We provided an algorithm for generating random languages and studying the space of variation modeled taking into account implications between parameters
- Investigating syntax within a generative framework focusing on the intricate system of implications between parameters proved that the parametric approach is able to retrieve a high level of correct historical information
- This is an indirect argument that parametric theories are a realistic approach to syntactic variation at a large cross-linguistic scale



Langelin has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement n° 295733. Langelin is developed with the participation of University of York, University of Ferrara and University of Bologna.

i'Ethiopian
L'Égyptien
Le Malais

THANKS!



Selected references:

THE UNIVERSITY *of York*

- Baker, M. 2001. *The Atoms of Language*. New York, Basic Books.
- Bortolussi, L., Longobardi, G., Guardiano, C., Sgarro, A. 2011. How many possible languages are there? In Bel-Enguix, G., Dahl, V., Jiménez-López, M. D. (eds). *Biology, Computation and Linguistics*, 168-179. Amsterdam, IOS Press.
- Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht, Foris.
- Dyen, I., Kruskal, J., Black, P. 1992. An Indo-European classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82 (5).
- Gray, R., Atkinson, Q. 2003. Language tree divergences support the Anatolian theory of Indo-European origin. *Nature*, 426: 435–439
- Guardiano, C., Longobardi, G. 2005. Parametric Comparison and Language Taxonomy. In Batllori, M., Hernanz, M. L., Picallo C., Roca F. (eds). *Grammaticalization and Parametric Variation*. 149-174. Oxford University Press.
- Longobardi, G., Guardiano, C. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119 (11): 1679-1706.
- Nichols, J. 1996. *The comparative method as heuristic* in Durie M., Ross. M. (Eds.), *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press
- Ringe, D. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society*, 82 (1): 1-110
- Ringe, D., Warnow, T., Taylor, A. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society*, 100 (1): 59-129.