HOW TIME-BOUND IS YOUR GRAMMAR? PUSHING THE LIMITS OF COMPARATIVE METHODS THROUGH SYNTAX

LUCA BORTOLUSSI, ANDREA CEOLIN, GUIDO CORDONI, DIMITAR KAZAKOV,

CRISTINA GUARDIANO, MONICA IRIMIA, GIUSEPPE LONGOBARDI,

NINA RADKEVICH AND ANDREA SGARRO

LONG-STANDING PROBLEM OF HISTORICAL LINGUISTICS

- **How much similarity is required to demonstrate true language relatedness** when regular sound correspondences are not retrievable?

THE PARAMETRIC COMPARISON METHOD

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Lingua 119 (2009) 1679–1706

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

Available online 7 January 2009

A MODEL OF UG AND VARIATION

- Conceptual simplification on the basis of a **Principles and Schemata Model**

  (Longobardi 2005, 2017)

  - ‘Initial stage of UG (S_0) only contains parameter schemata, and not an extensional list of parameters’

PARAMETER SCHEMATA (LONGOBARDI 2017)

- Is F, F a feature, **grammaticalized**?
- Does F, F a grammaticalized feature, **Agree** with a category X (i.e. does F probe X)?
- Is F, F a grammaticalized feature, "strong" (i.e. does F **overtly attract** X, probe X with an EPP feature)?
- Is F, F a grammaticalized feature, **spread** on a category X?
- Does a functional category (a set of lexically co-occurring grammaticalized features) X have a **phonological matrix** Φ?
- Does F, F a grammaticalized feature, **probe** the minimal accessible category of type X (or is **pied-piping** possible)?
- Are f1 and f2, the respective values of two grammaticalized features, associated on a category X?
- Are f1 and f2, two feature values associated on X, optionally associated?
- Does a functional feature (set) exist in the vocabulary as a bound/free morpheme?

Nothing else is a parameter

91 PARAMETERS DEFINING THE SYNTAX OF THE NOMINAL DOMAIN

- The status of **features associated with D**, e.g. person, number, gender and definiteness
- Syntactic properties of **adjectives**, **relative clauses**, **genitival arguments** and **possessives**, **demonstratives**
- Type and scope of **N-movement**

PARAMETER IMPLICATIONS

- Parameter values: **[+]** or **[-]**
- But parameter values **imply each other** as well: **0** is the state of a parameter which is completely irrelevant owing to the settings of other parameters
- Large number of **parametric interdependencies** (implicational universal principles)
  - Out of 50x91=4550 cells (parameter states), 2045 are null (contain 0), i.e. **44.9%** of the information is redundant

THE EFFECT OF PARAMETRIC IMPLICATIONS

- To appreciate the typological restrictiveness of a realistic parameter system, we must calculate the **number of possible languages generated**.
- The **first 30 parameters** from TableA (less implicationally constrained than the successive ones) **generate less than 2^{19} admissible grammars** (Bortolussi et al. 2011, Ceolin et al. submitted), at least **eleven orders of magnitude less** than the 2^{30} expected under total independence (i.e. less than 500k as opposed to more than 1 billion)

TWO FURTHER STEPS TOWARDS PARSIMONY

i) implied parameters **do not play a role in acquisition** (are never present in the child's mind)

ii) only **[+] valued** parameters are present in the child's mind

iii) to take into account (i) and (ii), we calculated syntactic pairwise distances according to the **Jaccard** formula:

$$\Delta \text{Jaccard}(A,B) = \frac{[\,N_{-+} + N_{+-}\,]}{[\,N_{-+} + N_{+-} + N_{++}\,]}$$

AN EMPIRICAL IMPLEMENTATION OF THE PCM

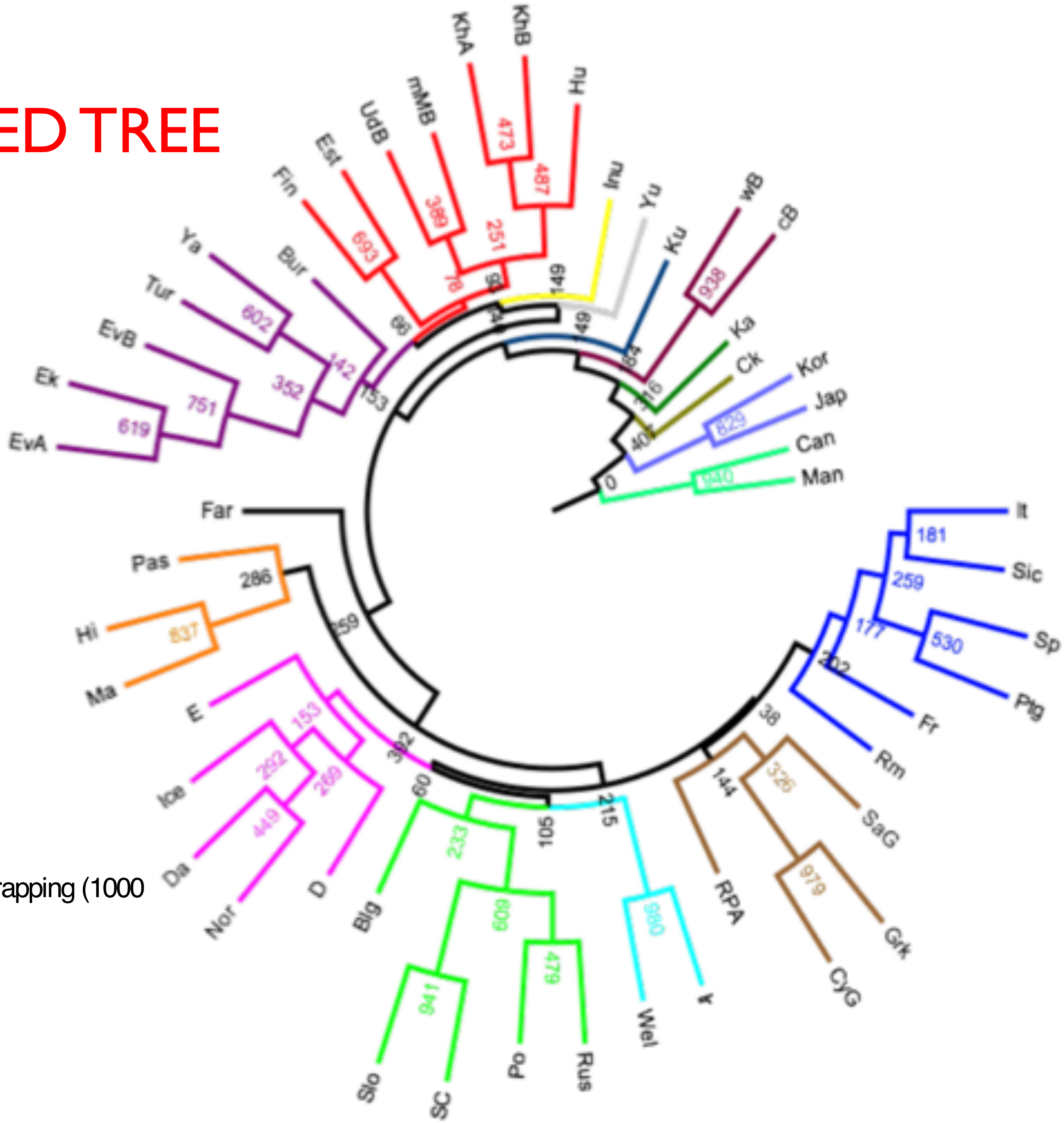
- **50** languages
- **14** traditionally recognized and irreducible families (scattered across Europe, Asia and the Americas)

- **Sino-Tibetan**: Mandarin (Man) and Cantonese (Can).
- **Korean**: Korean (Kor).
- **Japonic**: Japanese (Jap).
- **Muskogean**: Chickasaw (Ck).
- **Guaicuruan**: Kadiweu (Ka).
- **Basque**: Western and Central Basque (wB and cB).
- **Carib**: Kuikuro (Ku).

- **Indo-European**:
  - Indo-Iranian: Farsi (Far), Pashto (Pas), Marathi (Ma), Hindi (Hi);
  - Greek: Romeyka Pontic (RPA), Salento Greek (SaG), Standard Modern Greek (Grk), Cypriot Greek (CyG);
  - Romance: Romanian (Rm), French (Fr), Italian (It), Sicilian (Sic), Spanish (Sp), Portuguese (Ptg);
  - Germanic: English (E), German (D), Icelandic (Ice), Danish (Da), Norwegian (Nor);
  - Celtic: Irish (Ir), Welsh (Wel);
  - Slavic: Bulgarian (Blg), Russian (Rus), Polish (Po), Serbo-Croat (SC), Slovenian (Slo).

- **Yukaghir**: Yukaghir (Yu).
- **Inuit**: Inuktitut (Inu).
- **Mongolian**: Buryat (Bur).
- **Turkic**: Turkish (Tur), Yakut (Ya).
- **Tungusic**: two varieties of Even (EvA, EvB), Evenki (Ek).
- **Uralic**: Balto-Finnic: Finnish (Fin), Estonian (Est); Mari: Meadow Mari (mM); Permic: Udmurt (Ud); Ugric: Hungarian (Hu), two varieties of Khanty (KhA, KhB).

DISTANCE-BASED TREE



KITSCH Tree from the 50 languages.

Consensus tree calculated after a bootstrapping (1000 samples).

MACRO-FAMILIES

- Tree topologies cannot prove genetic relationship. One needs to rely on **statistical testing** (Kessler and Lehtonen 2006)

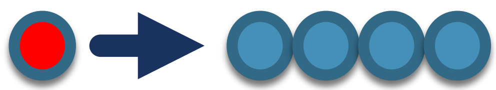
- First step: determining a **null distribution**

NULL DISTRIBUTION

- With only **14 families** it's **difficult to determine a null distribution by internal sampling**. Also, with **most** of the languages being from **Eurasia**, it's **difficult to control for historical relations**
- Solutions: we generate artificial languages by recombining the parametric values of the sample. Values are chosen probabilistically, using evidence weighted for families **Implicational constraints are also applied**
- We generate a sample of **7000 languages**

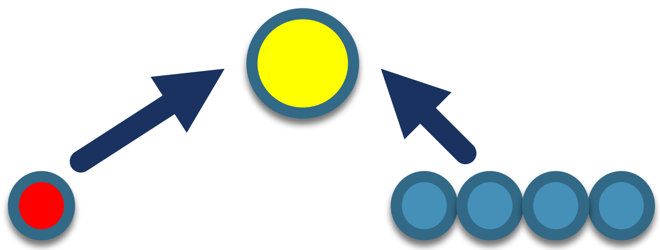
MEDIAN TESTS

- We **compare distributions** of real and artificial distances by means of Mann-Whitney U tests.
- **First test**:
  - Compare median of distances calculated from a **group of real languages** (e.g. Indo-European) versus median of distances calculated from a **group of artificial languages of the same size**.
  - Repeat 1000 times.
  - Check p-values range.



MEDIAN TESTS

- **Second test**:
  - Compare median of distances calculated from a **group of real languages of size N** versus median of distances calculated from a **bigger sample** (2000 languages).
  - Repeat the same using each of the 1000 artificial samples of the same size N instead.
  - Check p-values range.

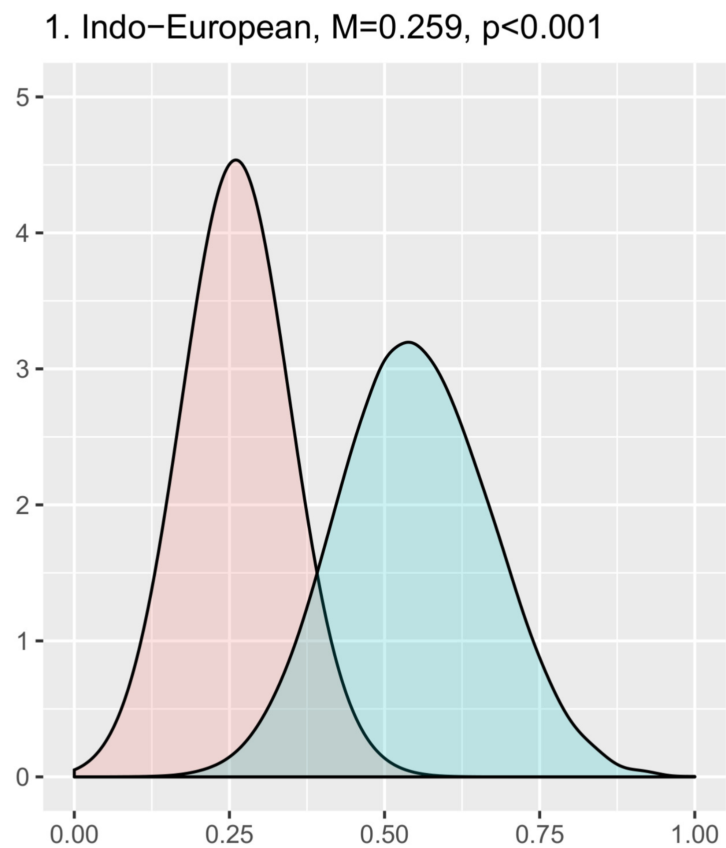


WHICH MACRO-FAMILIES?

- Potentially, many ($2^N-(N+1)$).
  - We start from **two well-established families**, IE and Uralic.
  - We move to **previously hypothesized macrofamilies**, Altaic and Indo-Uralic.
  - We test two groups suggested by the tree, Uralo-Altaic and Uralo-Altaic-Inuktitut.
- **Important**: when testing across families, **remove family-internal pairs** (and adjust artificial sample accordingly through sampling)

A. Well-established families

1. Indo-European
2. Uralic



n_pairs = 277

median = 0.259, sd = 0.058

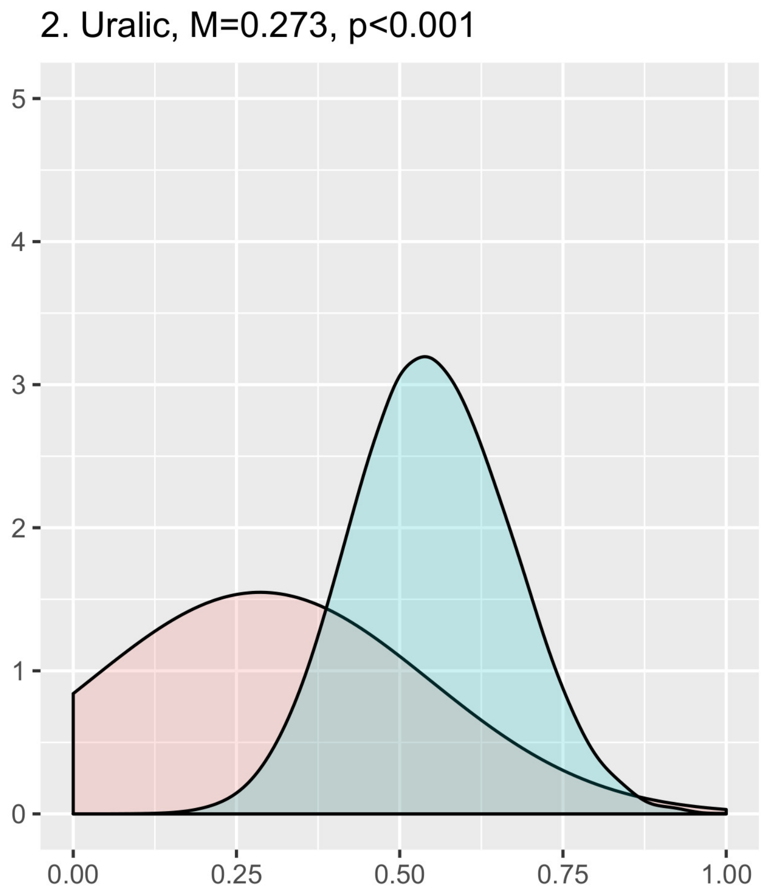
range_mannwhitney = $[1.47*10^{-6}, 0.0002]$

Test1 = <0.001

mannwhitney_bigsample = $1.34*10^{-172}$

min_artificial = $2.61*10^{-16}$

Test2 = <0.001



n_pairs = 17

median = 0.273, sd = 0.113

range_mannwhitney = $[9.98*10^{-7}, 0.029]$

Test1 = <0.001

mannwhitney_bigsample = $1.57*10^{-10}$

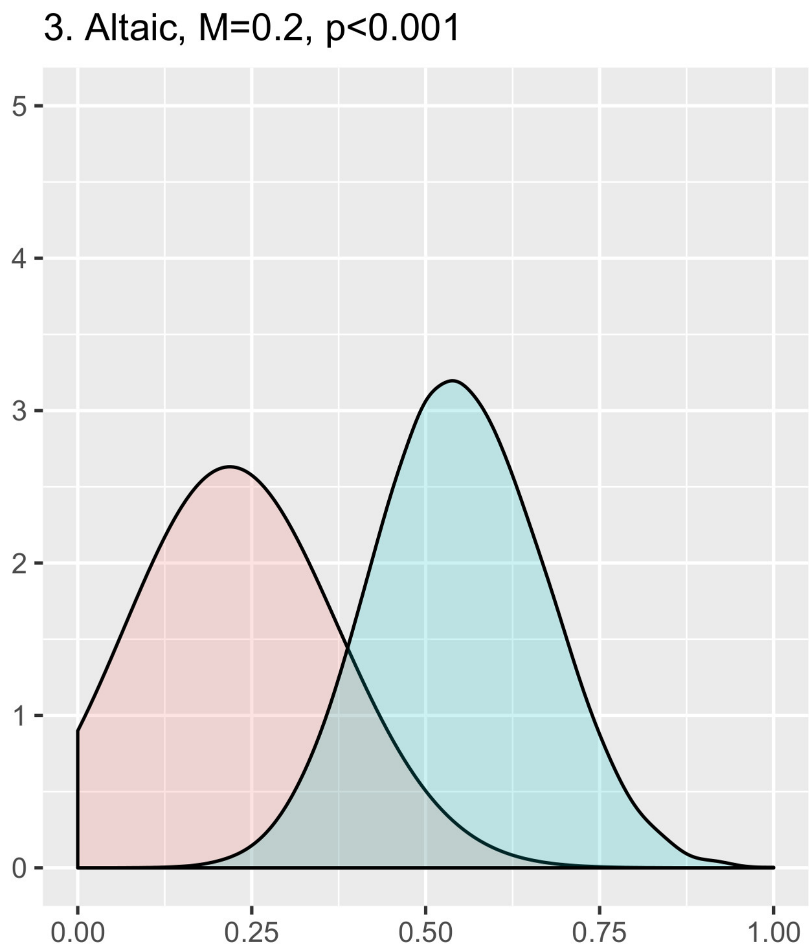
min_artificial = $3.15*10^{-9}$

Test2 = <0.001

B. **Hypothesized macro-families**

3. Altaic

4. Indo-Uralic



n_pairs = 11

median = 0.2, sd = 0.068

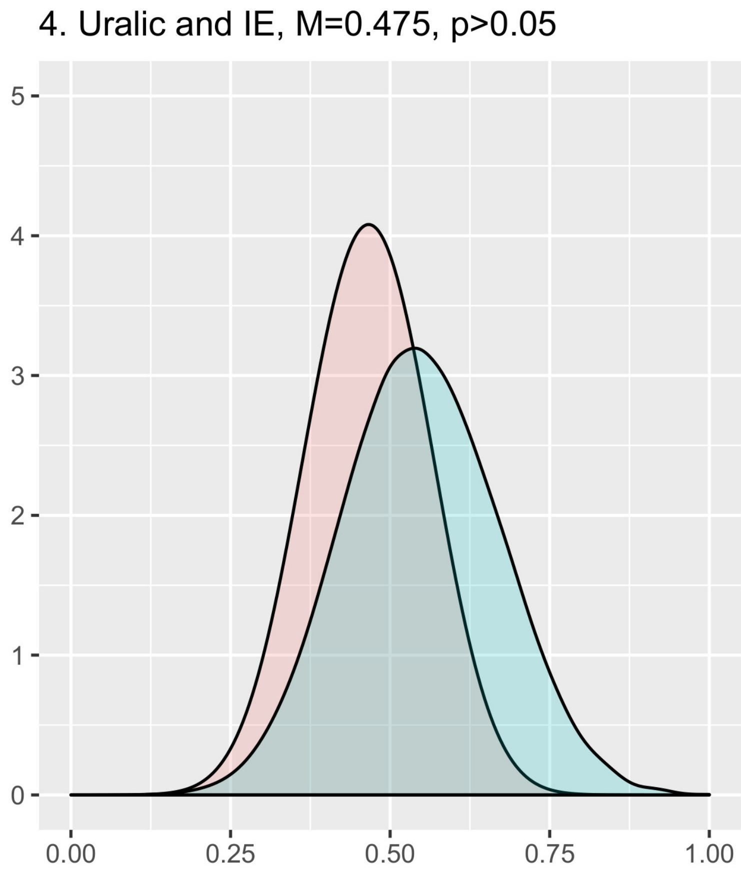
range_mannwhitney = [0.001, 0.007]

Test1 = <0.001

mannwhitney_bigsample = $1.58*10^{-8}$

min_artificial = $3.35*10^{-6}$

Test2 = <0.001



n_pairs =184

median = 0.475, sd = 0.059

range_mannwhitney = $[5.4*10^{-60}, 0.98]$

Test1 = 0.116

mannwhitney_bigsample = $3.53*10^{-26}$

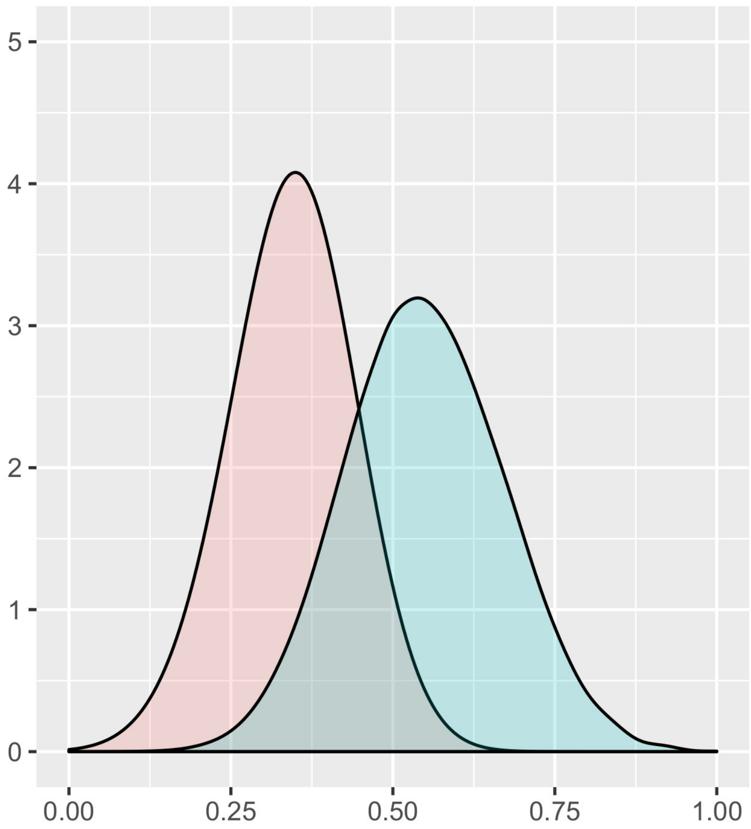
min_artificial = $1.74*10^{-80}$

Test2 = 0.073

C. Further groups suggested by the tree

5. Uralo-Altaic
6. Uralo-Altaic-Inuktitut

5. Uralic and Altaic, M=0.35, p<0.001



n_pairs = 42

median= 0.35, sd = 0.057

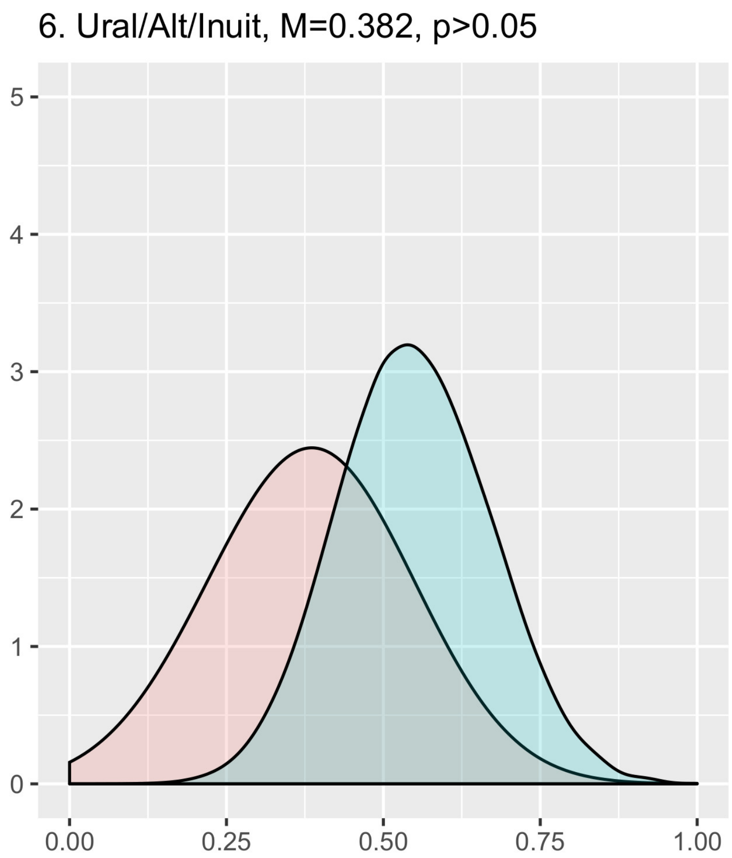
range_mannwhitney = $[4.12*10^{-15}, 0.007]$

Test1 = <0.001

mannwhitney_bigsample = $3.94*10^{-23}$

min_artificial = $4.12*10^{-22}$

Test1 = <0.001



n_pairs = 13

median= 0.382, sd = 0.069

range_mannwhitney = [$1.05*10^{-5}$, 0.797]

Test1 = 0.090

mannwhitney_bigsample = $1.48*10^{-6}$

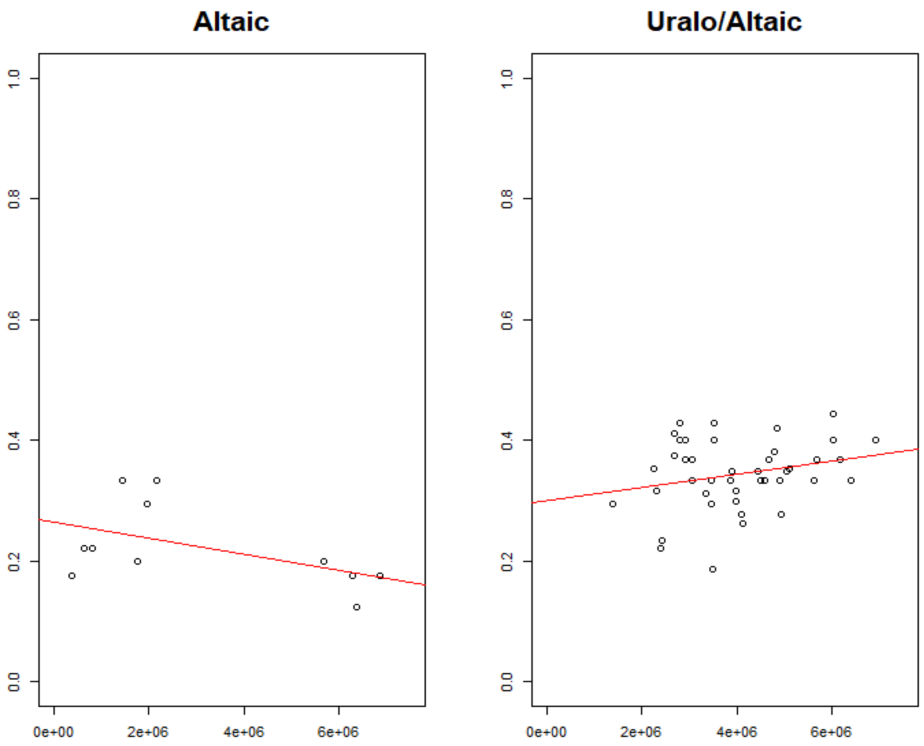
min_artificial = $4.93*10^{-7}$

Test1 = 0.005

SUMMARY

- **Indo-European** and **Uralic** are expectedly supported by the test.
- The **Altaic** hypothesis is corroborated.
- Further evidence for a **Uralo-Altaic unit**.
- Larger groups, like Indo-Uralic or Uralo-Altaic/Inuit, are weakly supported by the test.

GEOGRAPHY



	Altaic	Uralo/Altaic
Kendall tau	-0.29	0.124
p-value	0.223	0.2611
Pearson r	-0.51	0.247
p-value	0.107	0.114

CONCLUSIONS

1. Syntax contains a **detectable historical signal**
2. Unlike classical methods, syntax provides **long-range phylogenies encompassing distinct families**
3. Language relatedness can be **tested statistically** thanks to this syntax-based perspective