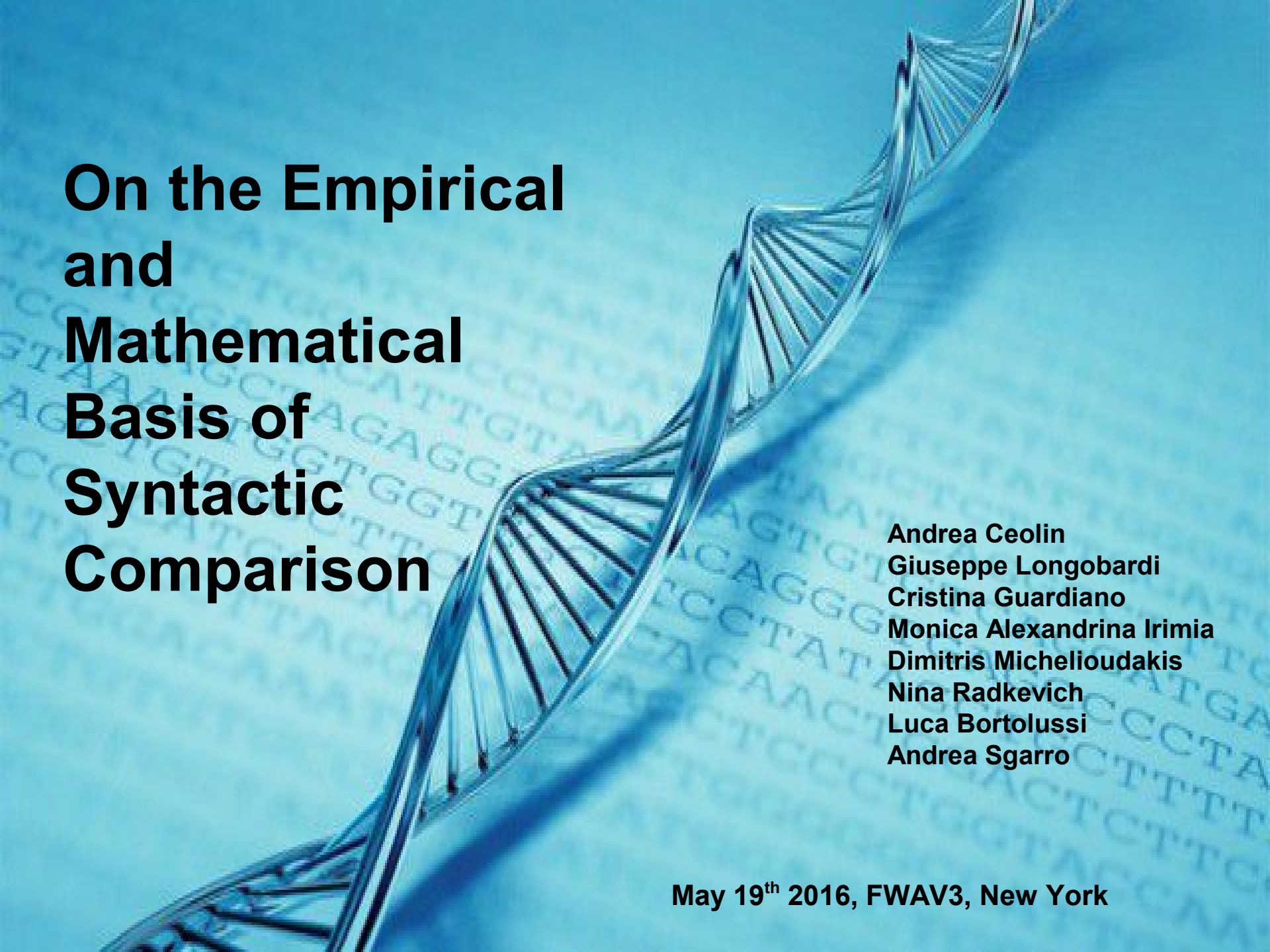
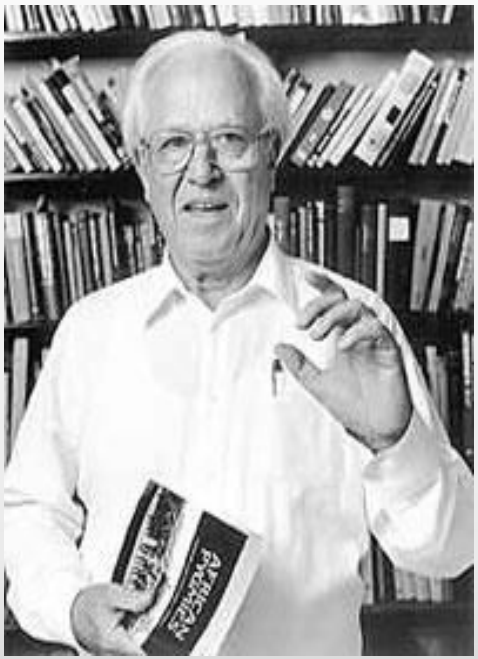


On the Empirical and Mathematical Basis of Syntactic Comparison

Andrea Ceolin
Giuseppe Longobardi
Cristina Guardiano
Monica Alexandrina Irimia
Dimitris Michelioudakis
Nina Radkevich
Luca Bortolussi
Andrea Sgarro

May 19th 2016, FWAV3, New York

Cavalli Sforza et al. (1988)



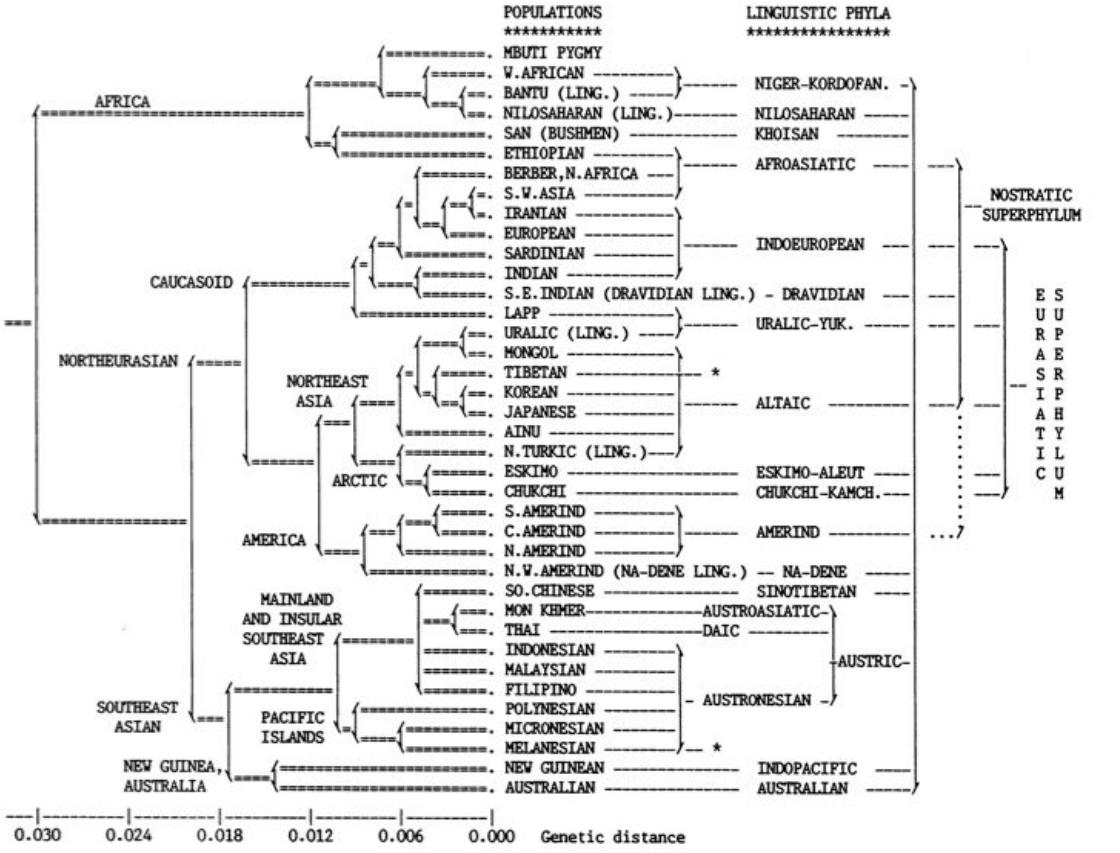


FIG. 1. Comparison of genetic tree and linguistic phyla. See text for details. (Ling.) indicates populations pooled on the basis of linguistic classification. The tree was constructed by average linkage analysis of Nei's genetic distances. Distances were calculated based on 120 allele frequencies from the following systems: *A1A2BO, MNS, RH, P, LU, K, FY, JK, DI, HP, TF, GC, LE, LP, PEPA, PEPB, PEPC, AG, HLAA* (12 alleles), *HLAB* (17 alleles), *PI, CP, ACP, PGD, PGM1, MDH, ADA, PTC, E1, SODA, GPT, PGK, C3, SE, ESD, GLO, KM, BF, LAD, E2, GM*, and *PG*.

Science AAAS

Home News Journals Topics Careers

Science Science Advances Science Immunology Science Robotics Science Signaling Science Translational Medicine

SHARE

REPORT

Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa

Quentin D. Atkinson[1,2,*]

+ Author Affiliations

*E-mail: q.atkinson@auckland.ac.nz

Science 15 Apr 2011:
Vol. 332, Issue 6027, pp. 346-349
DOI: 10.1126/science.1199295

Atkinson (2011)

Creanza et al. (2015)

PNAS PNAS

A comparison of worldwide phonemic and genetic variation in human populations

Nicole Creanza[a], Merritt Ruhlen[b], Trevor J. Pemberton[c], Noah A. Rosenberg[a], Marcus W. Feldman[a,1], and Sohini Ramachandran[d,e,1]

[a]Department of Biology and [b]Department of Anthropology, Stanford University, Stanford, CA 94305; [c]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, Canada R3E 0J9; and [d]Department of Ecology and Evolutionary Biology and [e]Center for Computational Molecular Biology, Brown University, Providence, RI 02912

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2013.

Contributed by Marcus W. Feldman, December 17, 2014 (sent for review July 16, 2014; reviewed by Quentin D. Atkinson and Keith Hunley)

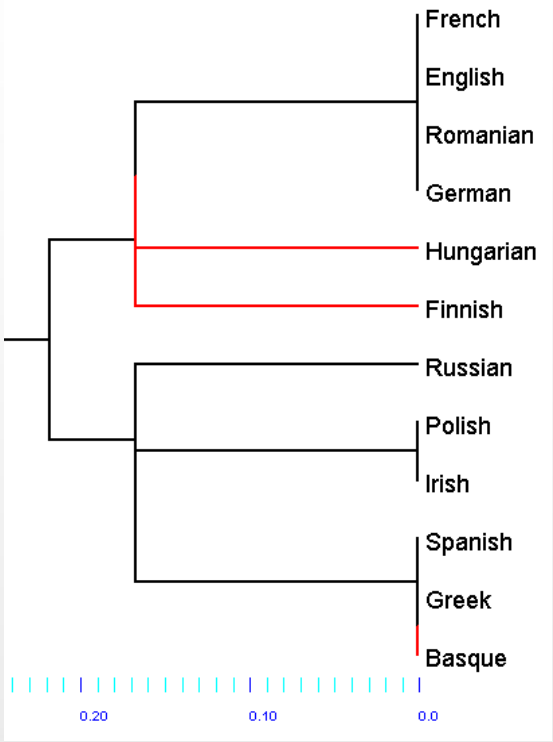
Worldwide patterns of genetic variation are driven by human demographic history. Here, we test whether this demographic history has left similar signatures on phonemes—sound units that distinguish meaning between words in languages—to those it has left on genes. We analyze, jointly and in parallel, phoneme inventories from 2,082 worldwide languages and microsatellite polymorphisms from 246 worldwide populations. On a global scale, both

compares the signatures of human demographic history in microsatellite polymorphisms from 246 worldwide populations (20) and complete sets of phonemes (phoneme inventories) for 2,082 languages; these are the largest available datasets of both genotyped populations and phonemes, the smallest units of sound that can distinguish meaning between words. Languages do not hold information about deep ancestry as genes do, and

Which kind of information do phonological and phonetic databases provide about Language? Are they compatible with our knowledge about historical language families?

Empirical Test: Europe (different language families, well-studied area)

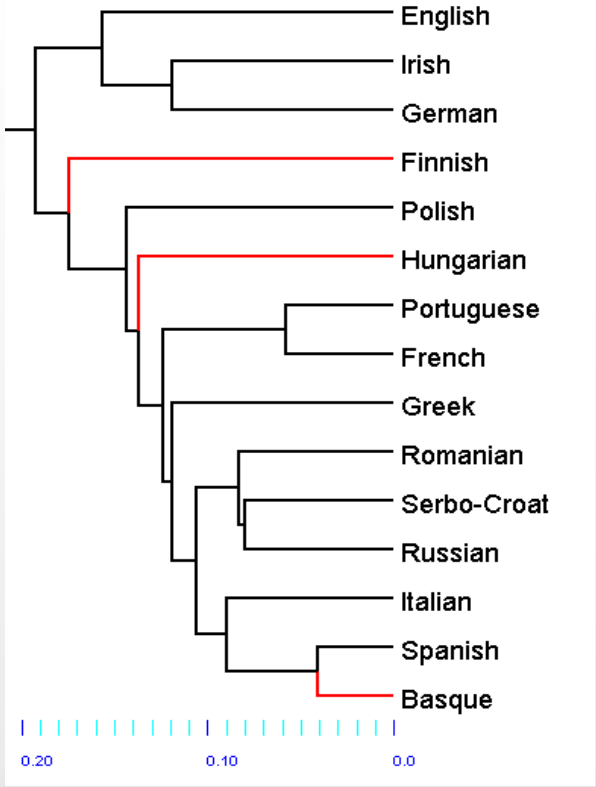
We can compute phylogenetic trees to study the vertical signal of phonological and phonetic data.



Distance-based trees

KITSCH (Phylip package)

Felsenstein (2004)



Atkinson 2011

WALS – Vowel and Consonant Inventories

Creanza et al. 2015

Ruhlen phonemic database

Can linguists do better?

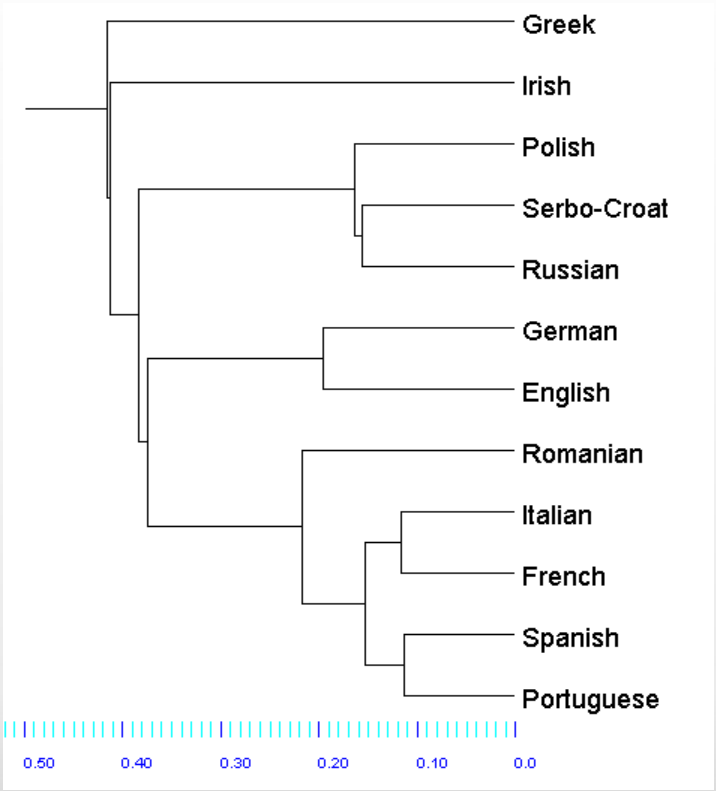
The **Classical Comparative Method** provided excellent results.
Additionally, **Swadesh word-lists** have been proposed to study relationships between well-established families .

Dyen, Kruskal and Black 1992 and Ringe, Warnow and Taylor 2002

but.....

Both the Classical Comparative Method and Swadesh-lists are limited. **They cannot be used to establish relations which go beyond a certain amount of time**.

IELex (Bouckaert et al. 2012, 2013)



ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

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]

Syntactic Analysis

Analysis of Morphosyntactic Features

Syntactic Parameters (Chomsky 1981)

Features on Functional Items ('Borer & Chomsky conjecture', Baker 2008).

Research Article

Across language families: Genome diversity mirrors linguistic variation within Europe

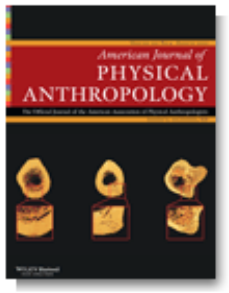
OPEN

Giuseppe Longobardi[1,2], Silvia Ghirotto[3], Cristina Guardiano[4], Francesca Tassi[3], Andrea Benazzo[3], Andrea Ceolin[1] and Guido Barbujani[3,*]

Article first published online: 8 JUN 2015

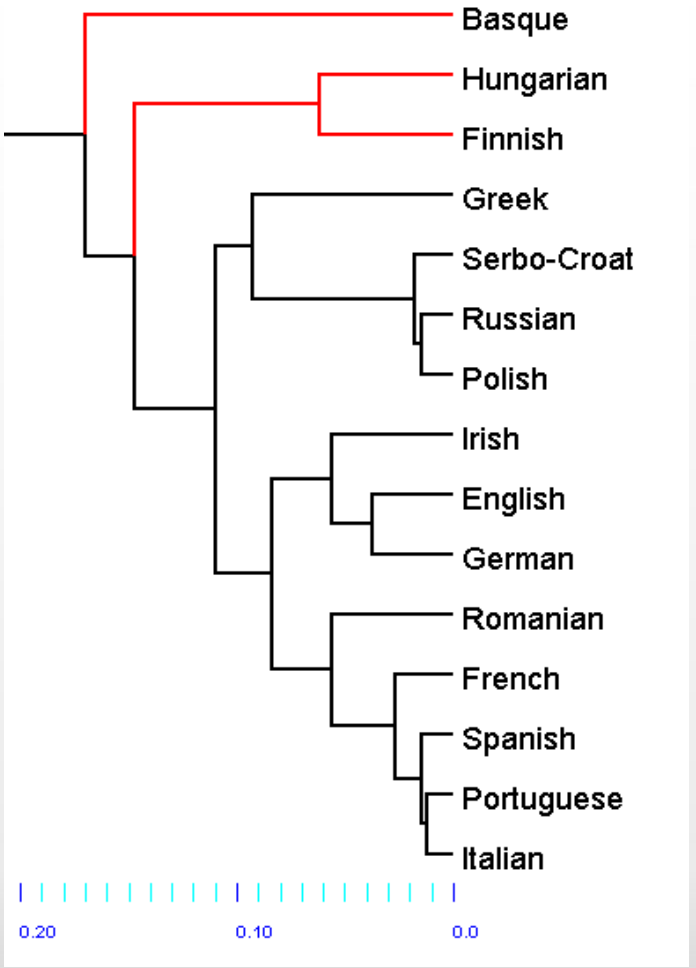
DOI: 10.1002/ajpa.22758

Issue



American Journal of Physical Anthropology

Volume 157, Issue 4, pages 630–640, August 2015



TableA (**Longobardi et al. 2015**)

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

Grammaticalized definiteness and enclitic articles

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

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

Grammaticalized definiteness and enclitic articles

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

Sample parameters (82):

82 syntactic parameters from the Nominal Domain (DP)

Sample languages (40):

23 IE languages (Portuguese, Spanish, French, Romanian, Italian, English, German, Danish, Nowegian, Icelandic, Welsh, Irish, Slovenian, Serbo-Croat, Bulgarian, Polish, Russian, Standard Greek, Pontic Greek, Cypriot Greek, Pashto, Hindi, Marathi)

5 Finno-Ugric languages (Finnish, Estonian, Hungarian, Udmurt, Khanty)

5 Altaic languages (Turkish, Buryat, Even, Evenki, Yakut)

2 Basque varieties (Western Basque, Central Basque)

2 Sinitic languages (Mandarin, Cantonese)

Yukaghir

Japanese

Korean

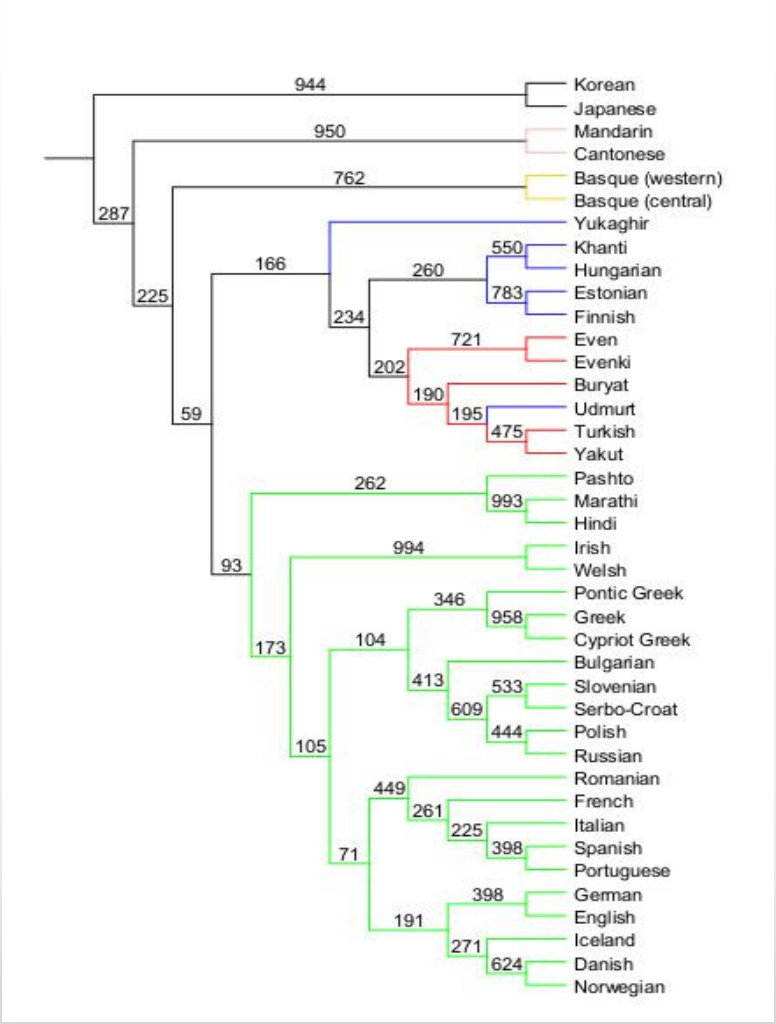
Distance-based trees

KITSCH (Phylip package)

Felsenstein (2004)

Bootstrapping procedure to remove homoplastic/horizontal effects

The only assumption is equidistance from the root to the leaves (the Molecular Clock Hypothesis)



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

Character-based methods work in a controlled environment (no horizontal transmission, no homoplasy, shared innovations, independence assumption)

Syntactic variation does not meet these requirements. Distance-based methods are more neutral (i.e. less sensitive to local phenomena)

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.) $\delta = 6 / 41 = \mathbf{0.146}$

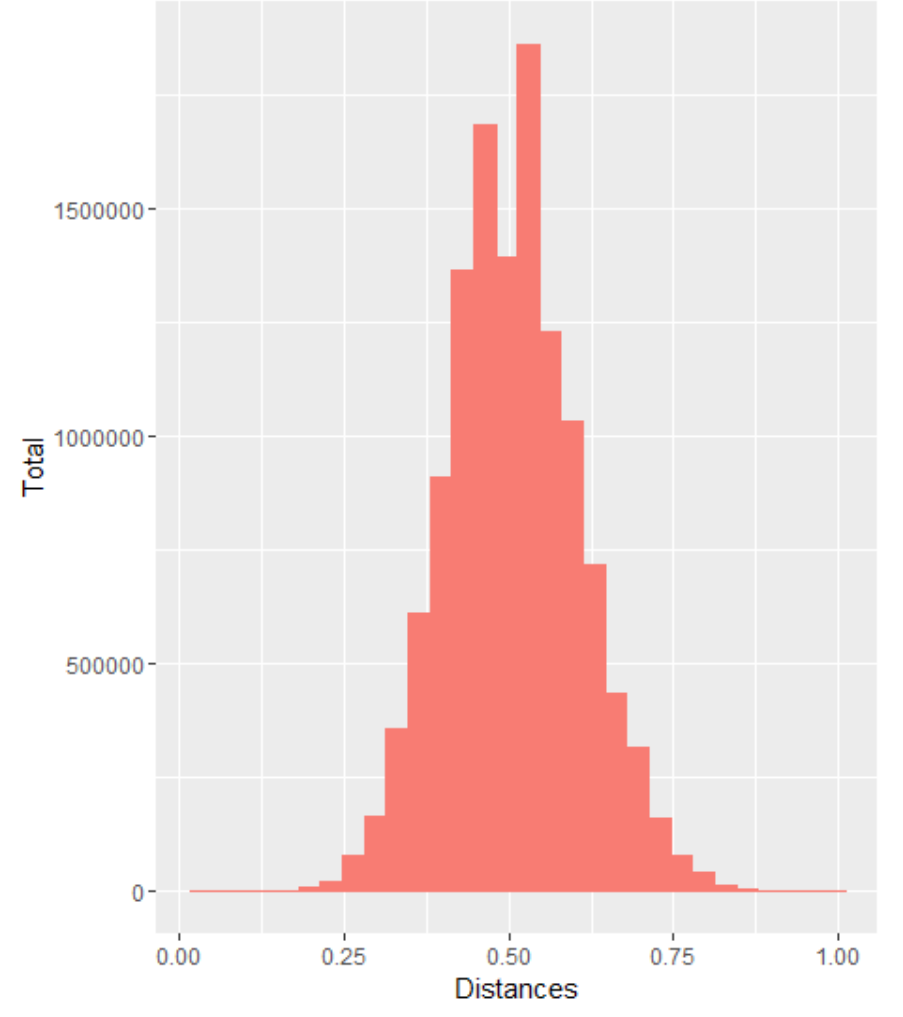
Random sample

5000 random languages

~12M random pairs

Mean: 0.5058

Median: 0.5



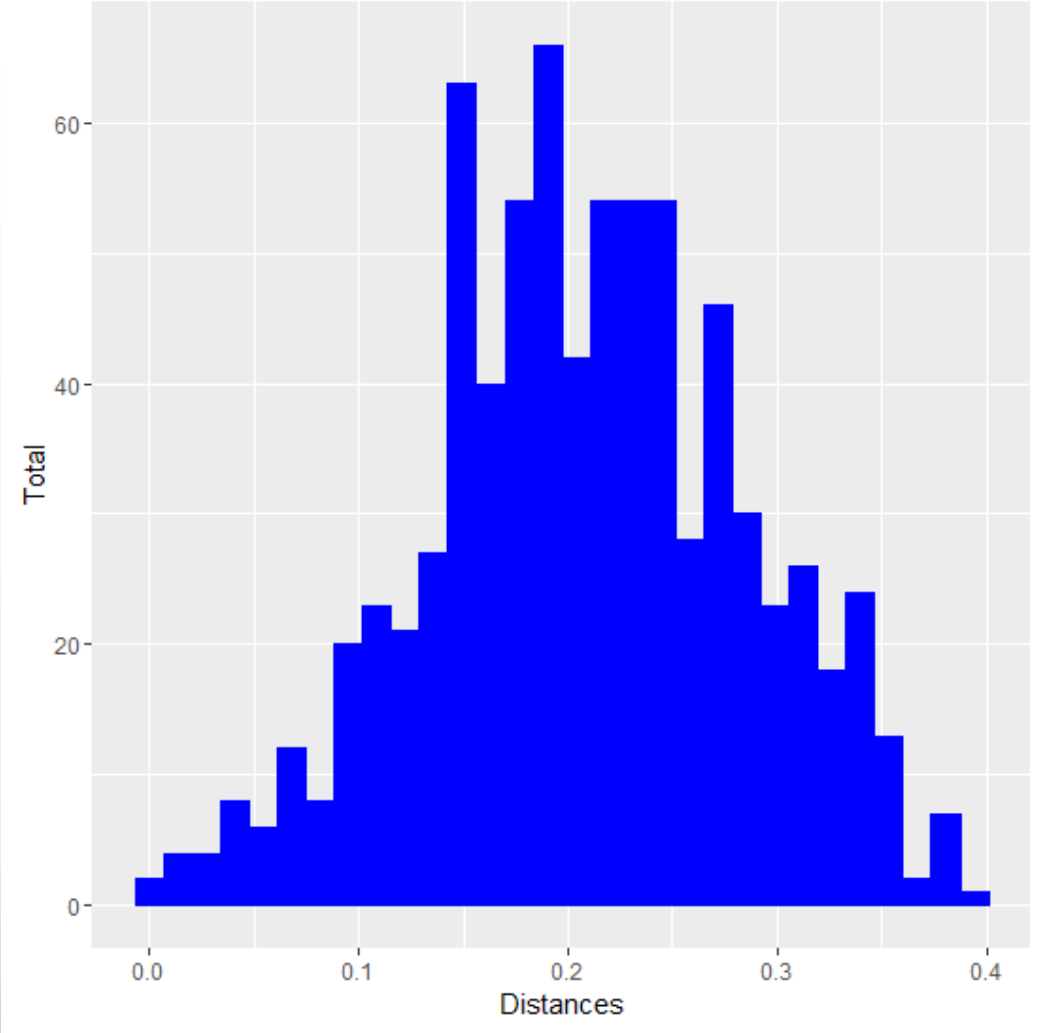
Real sample

40 real languages

728 real pairs

Mean: 0.2082

Median: 0.2085



We can check which kind of pairs we see at the left tail of the distribution.

A good threshold to start with can be the 10^{-3} quantile, which according to the random distribution should fall at d = **0.222**.

We have 442 / 780 **pairs** exhibiting distance < 0.222

Family-internal pairs (257)

	Under threshold	Total	Percentage
Indoeuropean	235	253	92.89%
Finno-Ugric	10	10	100.00%
Altaic	10	10	100.00%
Basque	1	1	100.00%
Chinese	1	1	100.00%

Family-external pairs (167)

	Under threshold	Total	Percentage
IE/Finno-Ugric	63	115	54.78%
IE/Altaic	39	115	33.91%
Finno-Ugric/Altaic	**25**	**25**	**100.00%**
IE/Basque	23	46	50.00%
IE/Chinese	8	46	17.39%
Altaic/Yukaghir	**4**	**5**	**80.00%**
Finno-Ugric/Chinese	3	12	25.00%
Finno-Ugric/Yukaghir	2	5	40.00%
Japanese/Korean	1	1	100.00%

The 10^{-4} quantile falls at d = **0.167**

We have 227 / 780 **pairs** exhibiting distance < 0.167

	Under threshold	Total	Percentage
Indoeuropean	149	253	**58.89%**
Finno-Ugric/Altaic	21	25	**84.00%**
IE/Finno-Ugric	15	115	13.04%
IE/Altaic	11	115	9.57%
Finno-Ugric	10	10	100.00%
Altaic	9	10	90.00%
Altaic/Yukaghir	2	5	40.00%
Altaic/Basque	2	10	20.00%
Finno-Ugric/Basque	2	10	20.00%
IE/Basque	2	46	4.35%
Basque	1	1	100.00%
Chinese	1	1	100.00%
Japanese/Korean	1	1	100.00%

Uralo-Altaic pairs are the only ones that seem to be consistently below the critical threshold.

How can we rule out horizontal transmission?

In Europe, the correlation with geography is weak (Longobardi et al. 2015):

Syntax/Geography(IE)	0.3879 (p= 0.0001)
Syntax/Geography (IE of Europe)	0.2774 (p= 0.002)

What about Uralo/Altaic?

Syntax/Geography(Uralo/Altaic) 0.2559 (p= 0.0331)

What about Uralo/Altaic?

The correlation with geography is less powerful and less significant in Uralo/Altaic than it is for the Indoeuropean family from a syntactic view point.

CONCLUSIONS

- Generative Syntax is a powerful tool to classify languages. Not only is it not entirely disrupted by horizontal transmission (like Phonetics and Phonology), but it can allow the investigation of macro-families, something that the Classical Comparative Method cannot pursue by definition.

- Focusing on Eurasia, there is evidence for a Uralo/Altaic group that emerges using only syntactic data. Some syntactic properties of the DP seem to be property of the entire family. More syntactic investigation is needed to confirm or reject the hypothesis.

- These findings, if compared to the genetic landscape of the populations, might provide new insights to study historical migrations in Eurasia. The method can be potentially extended to a world wide level.

THANKS!

Guido Cordoni
Aaron Ecay

Selected references:

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science, 332(6027), 346-349.

Chomsky, N. 1981. Lectures on Government and Binding. Dordrecht, Foris.

Creanza, N., Ruhlen, M., Pemberton, T. J., Rosenberg, N. A., Feldman, M. W., & Ramachandran, S. (2015). A comparison of worldwide phonemic and genetic variation in human populations. Proceedings of the National Academy of Sciences, 112(5), 1265-1272.

Dyen, I., Kruskal, J., Black, P. 1992. An Indo-European classification: a lexicostatistical experiment. Transactions of the American Philosophical Society, 82 (5).

Longobardi, G., Guardiano, C. 2009. Evidence for syntax as a signal of historical relatedness. Lingua, 119 (11): 1679-1706.

Longobardi, G. et al., Across Language Families: Genome diversity mirrors linguistic variation within Europe, American Journal of Physical Anthropology, 157(4):630-640, 2015.

Ringe, D., Warnow, T., Taylor, A. 2002. Indo-European and computational cladistics. Transactions of the Philological Society, 100 (1): 59-129.