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 19\textsuperscript{th} 2016, FWA\textsuperscript{V}3, 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: AIA3BO, MNS, RH, P, LU, K, FY, Jk, Dj, HP, TP, GC, Le, F, P, PEPA, PEPB, PEPC, AG, HLA-A (12 alleles), HLAB (17 alleles), Pi, CF, ACP, PGD, PGM1, MDH, ADA, PTC, E1, SODA, GPT, PGK, GJ, SE, ESID, GLO, KM, BF, LAD, E2, GM, and PG.
Creanza et al. (2015)

Atkinson (2011)

A comparison of worldwide phonemic and genetic variation in human populations

Nicole Creanza*, Merritt Ruhlen1, Trevor J. Pemberton*, Noah A. Rosenberg*, Marcus W. Feldman*,1, and Sohini Ramachandran1,2,3

*Department of Biology and 1Department of Anthropology, Stanford University, Stanford, CA 94305; 2Department of Biochemistry and Medical Genomics, University of Manitoba, Winnipeg, MB, Canada R3E 0W3; and 3Department of Ecology and Evolutionary Biology and 4Center 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 (201) 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 phoneme inventories can provide evidence of this lost information.
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
both........

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.**
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)
Across language families: Genome diversity mirrors linguistic variation within Europe

Giuseppe Longobardi¹,², Silvia Ghiroto³, Cristina Guardiano⁴, Francesca Tassi³, Andrea Benazzo³, Andrea Ceolin¹ and Guido Barbujani¹

American Journal of Physical Anthropology
Volume 157, Issue 4, pages 630–640, August 2015

DOI: 10.1002/ajpa.22758
© 2015 Wiley Periodicals, Inc.

Table A (Longobardi et al. 2015)
Languages are encoded as lists of binary parameters (+,-)

Grammaticalized definiteness and enclitic articles

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Norwegian</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>p10: gramm. Def. (articles)</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>p14: enclitic articles (+p10)</td>
<td>-</td>
<td>+</td>
<td>?</td>
</tr>
</tbody>
</table>
Languages are encoded as lists of binary parameters (+,-)

Grammaticalized definiteness and enclitic articles

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Norwegian</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>p10: gramm. Def. (articles)</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>p14: enclitic articles (+p10)</td>
<td>-</td>
<td>+</td>
<td>0</td>
</tr>
</tbody>
</table>
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:

\[ \delta(A,B) = \frac{d(A,B)}{d(A,B) + i(A,B)} \]

= differences / identities + differences

E.g.: Italian-English: (35 id., 6 diff.) \( \delta = \frac{6}{41} = 0.146 \)
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)

<table>
<thead>
<tr>
<th>Language Family</th>
<th>Under threshold</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indoeuropean</td>
<td>235</td>
<td>253</td>
<td>92.89%</td>
</tr>
<tr>
<td>Finno-Ugric</td>
<td>10</td>
<td>10</td>
<td>100.00%</td>
</tr>
<tr>
<td>Altaic</td>
<td>10</td>
<td>10</td>
<td>100.00%</td>
</tr>
<tr>
<td>Basque</td>
<td>1</td>
<td>1</td>
<td>100.00%</td>
</tr>
<tr>
<td>Chinese</td>
<td>1</td>
<td>1</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
## Family-external pairs (167)

<table>
<thead>
<tr>
<th>Pair</th>
<th>Under threshold</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>IE/Finno-Ugric</td>
<td>63</td>
<td>115</td>
<td>54.78%</td>
</tr>
<tr>
<td>IE/Altaic</td>
<td>39</td>
<td>115</td>
<td>33.91%</td>
</tr>
<tr>
<td>Finno-Ugric/Altaic</td>
<td><strong>25</strong></td>
<td>25</td>
<td><strong>100.00%</strong></td>
</tr>
<tr>
<td>IE/Basque</td>
<td>23</td>
<td>46</td>
<td>50.00%</td>
</tr>
<tr>
<td>IE/Chinese</td>
<td>8</td>
<td>46</td>
<td>17.39%</td>
</tr>
<tr>
<td>Altaic/Yukaghir</td>
<td>4</td>
<td>5</td>
<td>80.00%</td>
</tr>
<tr>
<td>Finno-Ugric/Chinese</td>
<td>3</td>
<td>12</td>
<td>25.00%</td>
</tr>
<tr>
<td>Finno-Ugric/Yukaghir</td>
<td>2</td>
<td>5</td>
<td>40.00%</td>
</tr>
<tr>
<td>Japanese/Korean</td>
<td>1</td>
<td>1</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
The $10^{-4}$ quantile falls at $d = 0.167$

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

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Under threshold</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indoeuropean</td>
<td>149</td>
<td>253</td>
<td><strong>58.89%</strong></td>
</tr>
<tr>
<td>Finno-Ugric/Altaic</td>
<td>21</td>
<td>25</td>
<td><strong>84.00%</strong></td>
</tr>
<tr>
<td>IE/Finno-Ugric</td>
<td>15</td>
<td>115</td>
<td><strong>13.04%</strong></td>
</tr>
<tr>
<td>IE/Altaic</td>
<td>11</td>
<td>115</td>
<td><strong>9.57%</strong></td>
</tr>
<tr>
<td>Finno-Ugric</td>
<td>10</td>
<td>10</td>
<td><strong>100.00%</strong></td>
</tr>
<tr>
<td>Altaic</td>
<td>9</td>
<td>10</td>
<td><strong>90.00%</strong></td>
</tr>
<tr>
<td>Altaic/Yukaghir</td>
<td>2</td>
<td>5</td>
<td><strong>40.00%</strong></td>
</tr>
<tr>
<td>Altaic/Basque</td>
<td>2</td>
<td>10</td>
<td><strong>20.00%</strong></td>
</tr>
<tr>
<td>Finno-Ugric/Basque</td>
<td>2</td>
<td>10</td>
<td><strong>20.00%</strong></td>
</tr>
<tr>
<td>IE/Basque</td>
<td>2</td>
<td>46</td>
<td><strong>4.35%</strong></td>
</tr>
<tr>
<td>Basque</td>
<td>1</td>
<td>1</td>
<td><strong>100.00%</strong></td>
</tr>
<tr>
<td>Chinese</td>
<td>1</td>
<td>1</td>
<td><strong>100.00%</strong></td>
</tr>
<tr>
<td>Japanese/Korean</td>
<td>1</td>
<td>1</td>
<td><strong>100.00%</strong></td>
</tr>
</tbody>
</table>
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.
Selected references:


Guido Cordoni
Aaron Ecay