# Mathematical modeling of grammatical diversity supports the historical reality of formal syntax

A. Ceolin[A], G. Longobardi[B,C], L. Bortolussi[B], C. Guardiano[D], M. A. Irimia[C], D. Michelioudakis[C], N. Radkevich[C], A. Sgarro[B]

[A] University of Pennsylvania
[B] Università di Trieste
[C] University of York
[D] Università di Modena e Reggio Emilia

The Classical Comparative Method has proven to be the only statistically uncontroversial method to study genealogical relationships between languages. However, the fact that the method can no longer be applied when phonetic correspondences are obscured by several thousands of years of language change has inspired the search for alternative methods for long-range comparison.

Longobardi and Guardiano (2009) show that another domain, syntax, is a potential source for cross-family comparison. The Parametric Comparison Method (PCM) uses syntactic parameters (Chomsky 1981, Baker 2001) to study relationships between languages. Parameters are coded as discrete binary values (+ or -). Additionally, the PCM allows for parametric implications, whereby a combination of values for some parameters can allow other parameters to take on only one value. The 'forced' or implied parameter in these cases is given the value 0 (undefined).

A question raised by the PCM framework is whether the results are secure against chance similarities between languages. Bortolussi et al. (2011) attempted to answer this question by using a randomly simulated distribution of parametric distances between languages (which are defined to range between 0 and 1) to perform statistical tests of the hypothesis that the distances observed in the real world are unlikely to arise by chance.

Here we evaluate the statistical significance of the results of PCM. We propose a refinement to Bortolussi et al.'s algorithm to better take into account the linguistic assumptions on syntactic parameters. After we generate a sample of 5000 artificial languages and calculate Jaccard distances among them, we compare the results with distances drawn from a database of 40 languages coded through 75 syntactic parameters (24 Indo-European, 3 Finno-Ugric, 2 Semitic, 2 Altaic, 2 Sinitic, 2 Basque and some isolated languages from Asia, Africa and South-America).

Figure 1 illustrates the difference between the distribution of actual language distances (green) and distances simulated by our algorithm (blue). We checked this difference with Mood's median test, which yielded an infinitesimally small p-value ($2.94 * 10^{-253}$), disconfirming the null hypothesis that the two distributions have equal medians. The difference remains ($p = 3.14 * 10^{-156}$), even after removing from the dataset language pairs that are both drawn from the same family (red).

If this signal were attributable to universal factors, such as the third factor computational pressures, it would not correlate with geographic or anthropological divisions.

Figure 2 shows the proportion of language pairs in our dataset that fall below a critical threshold (defined as the 10^3 quantile of the random distribution of distances). A high

proportion of pairs is exhibited by pairs within the IndoEuropean family. Almost all the missing pairs include an Iranian language (Farsi or Pashto), showing that this sub-family is the one which exhibits the highest distances with other IE languages.
Interestingly, all the pairs between Finno-Ugric (Finnish, Hungarian and Estonian) and Altaic (Turkish and Buryat) languages are below the threshold. While evidence for an Eurasiatic or Nostratic hypothesis is weak, the data seem to suggest the plausibility of a Ural/Altaic cluster. This finding requires further investigation.
These results confirm that syntactic parameters can provide novel information for the study of the prehistory of human languages, and hint at the possibility of aiming toward a greater time depth, given that parameters are part of a universal faculty of language.

REFERENCES


Baker M., The Atoms of Languages. New York, Basic Books, 2001.

Bortolussi L., G. Longobardi, C. Guardiano, A. Sgarro, "How many possible languages are there?", Biology, Computation and Linguistics. eds G. Bel-Enguix, V. Dahl, M. D. Jiménez-Lopez, IOS, Amsterdam, pp.168-179, 2011.

Chomsky N., "Lectures on Government and Binding", Foris, Dodrecht, 1981.

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


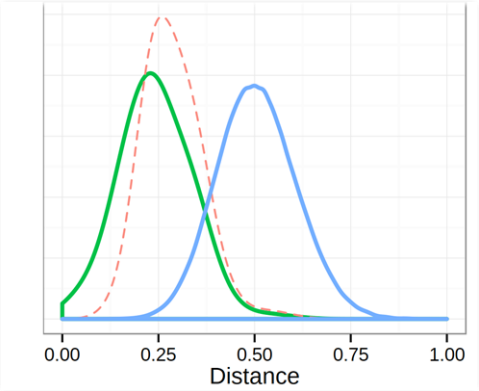


Fig.1

| Class | Table Column Head | | |
|---|---|---|---|
| | *Total Pairs* | *Below Threshold* | *Percentage* |
| IE | 276 | 205 | 74.3% |
| IE/Finno-Ugic | 72 | 23 | 31.9% |
| IE/Altaic | 48 | 4 | 8.3% |
| IE/Basque | 48 | 12 | 25.0% |
| IE/Semitic | 48 | 6 | 12.5% |
| IE/Inuktitut | 24 | 2 | 8.3% |
| Finno-Ugric/Altaic | 6 | 6 | 100% |

Fig.2