Historical Bio-Linguistics.

A biostatistic approach to the study of linguistic phylogenies and the correlation of genetic, linguistic and geographical data.

Guido Cordoni, Cristina Guardiano, Dimitar Kazakov, Giuseppe Longobardi.

Background
Demographic events often leave traces in languages and genes: this prompted Darwin’s prediction that the evolutionary tree of human populations would provide the best possible phylogeny of language relationships. We tested Darwin’s expectation through long-distance genome-language comparisons across Eurasia, relying on independently assessed quantitative tools on both sides. To do so, we had to resort to a linguistic method able to compare across different families, based on abstract syntactic characters, which proved more apt for long-term historical reconstruction than phonemic ones.

Materials and Methods

Data collection
Linguistics:
- Interviews with native speakers (collection of actual grammaticality judgments)
- Manifestations
- Binary strings for each language

Genetics:
- Sample collection
- DNA extraction
- Binary strings for each DNA collected

Geography
Three types of geographic distances were calculated: Great Circle Distance (GCD), Least Cost Distance (LCD) and CircuitScape Distance (CSD)

Phylogenetic analysis
1) Linguistic data
   - Distance based method (Jaccard-UPGMA) for syntactic and phonetic data (the latter found in literature: Creanza et al., 2015).
   - Character based methods (Bayesian).
2) Genetic data
   - Distance based method (Fixation Index (FST)-UPGMA).

Correlations
Relationships between syntactic, phonetic, genetic and geographic distances were assessed by using Mantel tests and partial Mantel tests.

Results

Linguistics
Distance-based tree from syntactic data
Character-based tree from syntactic data
Heats maps from phonemic data

Genetics
Distance-based tree of genetic data
PCA of genetic data

Mantel correlations

The Mantel correlation for the full dataset of syntactic data:
- Dsyn-Dgen: r=0.53 p=0.001
- Dsyn-Dgeo: r=0.47 p=0.001
- Dgen-Dgeo: r=0.88 p=0.001
- Dsyn/Dgen-Dgeo: r=0.28 p=0.004
- Dsyn/Dgeo-Dgen: r=0.04 p=0.001

The Mantel correlation for the full dataset of phonemic data:
- Dphon-Dgen: r=0.33 p=0.001
- Dphon-Dgeo: r=0.27 p=0.002
- Dphon-Dsyn: r=0.35 p=0.001
- Dphyon/Dgen-Dgeo: r=0.26 p=0.01
- Dphon/Dgen-Dsyn: r=0.20 p=0.02
- Dphon/Dgeo-Dsyn: r=0.04 p=0.05

These correlations were confirmed in strategically chosen subsets and performing an analysis on 9604 randomly chosen subsets of cardinality n=15 (Confidence value= 99% ±1).

Conclusions
We discovered significant congruence between biological and linguistic traits, not fully accounted for by any of the three types of geographic distances: this suggests that in the Old World the grammatical structure of most languages was largely transmitted along with the ancestral speakers’ genes. Indeed, we found that language history is much more vertical and independent of geography than genes, and correlates with geography only as a byproduct of its correlating with genes. Few exceptions to the congruence emerged, and were analyzed into differently motivated types, based on plausible conditions on language diffusion and on gene-language combinations. The patterns observed were summarized into a set of foundational generalizations for a science of long-term glottogenetic history, as emerging from the Eurasian domain.

Acknowledgement - Special thanks go to former and present collaborators of the Langelin project from the Universities of York, Modena and Reggio Emilia, Ferrara, Bologna and Pennsylvania.

Correspondence - guido.cordoni@york.ac.uk