Syntactic theory and the science of (language) history

Giuseppe Longobardi, Andrea Ceolin^, Guido Cordoni*, Cristina Guardiano°, Monica-Alexandrina Irimia°, Dimitar Kazakov*, Shin-Sook Kim*, Dimitris Michelioudakis*, Nina Radkevich**
*University of York, ^University of Pennsylvania, °Università di Modena e Reggio Emilia

Through the pioneering work of Ringe, Taylor and Warnow (2002), and Gray and Atkinson (2003), mathematical procedures of evolutionary biology have been introduced into historical and taxonomic linguistics and applied to its more traditional phonological and lexical data. These quantitative developments are a crucial step toward a scientific approach to the study of history. In fact, the rise of modern natural sciences in the XVII century has been supported not only by the adoption of a quantitative and experimental view of research, but also by a style of inquiry based on idealization and deduction of observed phenomena from very general abstract hypotheses. This line of inquiry has been tentatively reproduced with some success in synchronic linguistics by trends in cognitive science and formal grammatical theory, for instance representing crosslinguistic diversity as descending from a limited set of deep and interacting syntactic differences (parameters of universal grammar). Since Longobardi and Guardiano (2009), methods for importing the deductive style of formal grammar into the study of language phylogenies and for assessing the historical value of controversial parametric hypotheses have been explored (Longobardi et al. 2013).
In this work, we present a database of over 80 binary syntactic parameters set in 50+ languages from 4 continents, encoding grammars of nominal phrases in such languages, each in the form of a string of binary values (or of null states predicted indeed by the deductive structure of the hypotheses). This database is meant to preserve the deductive depth and predictive intricacy of highly axiomatized syntactic analyses, while improving on the empirical coverage by a full order more than one full order of magnitude (Baker 2012 calculated that the average number of languages compared in formal analyses involving parameters rarely exceeds 4 per article) with respect to standard practices of generative literature. The same dataset and associated hypotheses will be evaluated against two standards of adequacy:

A) Crosslinguistic descriptive adequacy. A typologically wide and high-resolution coverage can be achieved on a specific module of grammar (DP-structure) in terms of relatively few abstract syntactic characters
B) Historical adequacy. These characters achieve a high degree of historical adequacy by returning plausible phylogenies for four different domains:
1. IE languages
2. Macro-Uralic languages
3. Macro-Altaic languages
4. Microvariation samples (with sociolinguistic interactions) in the Northern and Eastern Mediterranean areas

Then, it will be shown how the insights obtained from the database can be used to explore historical issues beyond independently known language taxonomies. We will present:

i) Significant cross-family gene-language correlations among Eurasian populations.
ii) The possibility of formally evaluating the relative position of the IE, Uralic and Altaic phylogenies above within a wider set of other Eurasian and American languages.

On these grounds, we will try to argue for two conclusions: 1) capturing phylogenetic signals with a model embodying a high degree of universal hypotheses about language is not only possible, but also recommendable for the purposes of statistical reliability; 2) deeply deductive approaches advocated by some modern cognitive theories are useful for the scientific foundation of the study of human history.