doi 10.4436/jass.94031

JASs Proceeding Paper
Journal of Anthropological Sciences
Vol. 94 (2016), pp. 147-155

Formal linguistics as a cue to demographic history

Giuseppe Longobardi[1,2], Andrea Ceolin[3], Aaron Ecay[1*], Silvia Ghirotto[4], Cristina Guardiano[5], Monica-Alexandrina Irimia[5], Dimitris Michelioudakis[1], Nina Radkevich[1], Davide Pettener[6], Donata Luiselli[6] & Guido Barbujani[4]

1) *Department of Language and Linguistic Science, Vanbrugh College V/C/213, University of York, Heslington, York YO10 5DD, UK*
e-mail: giuseppe.longobardi@york.ac.uk

2) *Dipartimento di Studi Umanistici, Università di Trieste, Via del Lazzaretto Vecchio 6, 34123 Trieste, Italy*

3) *Department of Linguistics, 619 Williams Hall, 255 S 36th Street, Philadelphia, PA 19104-6305, USA*

4) *Dipartimento di Biologia ed Evoluzione, Università di Ferrara, Via Luigi Borsari 46, 44100 Ferrara, Italy*

5) *Dipartimento di Comunicazione ed Economia, Università di Modena e Reggio Emilia, Viale Allegri 9, 42121 Reggio Emilia, Italy*

6) *Dipartimento di Scienze Biologiche, Geologiche e Ambientali, Università di Bologna, Via Selmi 3, 40126 Bologna, Italy*

Summary - *Beyond its theoretical success, the development of molecular genetics has brought about the possibility of extraordinary progress in the study of classification and in the inference of the evolutionary history of many species and populations. A major step forward was represented by the availability of extremely large sets of molecular data suited to quantitative and computational treatments. In this paper, we argue that even in cognitive sciences, purely theoretical progress in a discipline such as linguistics may have analogous impact. Thus, exactly on the model of molecular biology, we propose to unify two traditionally unrelated lines of linguistic investigation: 1) the formal study of syntactic variation (parameter theory) in the biolinguistic program; 2) the reconstruction of relatedness among languages (phylogenetic taxonomy). The results of our linguistic analysis have thus been plotted against data from population genetics and the correlations have turned out to be largely significant: given a non-trivial set of languages/populations, the description of their variation provided by the comparison of systematic parametric analysis and molecular anthropology informatively recapitulates their history and relationships. As a result, we can claim that the reality of some parametric model of the language faculty and language acquisition/transmission (more broadly of generative grammar) receives strong and original support from its historical heuristic power. Then, on these grounds, we can begin testing Darwin's prediction that, when properly generated, the trees of human populations and of their languages should eventually turn out to be significantly parallel.*

Keywords - *Parametric comparison method, Historical syntax, Population genetics, Molecular anthropology, Biolinguistics, Computational linguistics.*

[*] Dr. Ecay collaborated to this work until November 1, 2015.

For nearly two centuries, biological anthropology and historical linguistics have pursued analogous questions, aiming to classify human populations and languages into genealogically significant families, thus explaining their resemblances and charting the paths of their diversification. Beyond the similarity of the problems, emphasized among linguists by Darwin's contemporary and admirer August Schleicher, it is worth remarking that Darwin himself in *The Origin of Species* had already predicted the eventual emergence of matching phylogenetic results for the two disciplines, which could then reinforce each other in their reconstruction of human past:

> *"If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one" (Darwin, 1859, ch. 14).*

Darwin's hypothesis of population-language congruence is *a priori* challenged by the fact that there is no deterministic connection between the biological transmission of genes and the cultural transmission of languages. For this very reason, an average global congruence would provide some of the strongest possible evidence about the peopling of the continents, as it would converge from two quite independent domains; furthermore, it would answer a basic theoretical question of modern anthropology, i.e. to what extent transmission of cultural traits is accompanied by robust demic expansion/migration of specific populations (Bellwood, 2005).

To test Darwin's claim, one needs a broad amount of information on both linguistic and genetic diversity over the world. After more than a century, Sokal (1988) showed that there are indeed variable degrees of positive correlation between genetic, geographic and linguistic distances in Europe. In the same year, Cavalli-Sforza *et al*. (1988) claimed to have found a general resemblance between evolutionary trees inferred, respectively, from genetic and linguistic similarities between populations. Their work has remained controversial, especially among linguists. Doubts and structured objections have been raised on at least three major points:

1) the very possibility of reasonably expecting a parallelism between the outputs of so different phenomena as biological evolution and culturally transmitted linguistic history (Bateman *et al.*, 1990);
2) the somewhat unclear task of identifying a unitary population in genetic and linguistic terms;
3) finally but most importantly, the reliance on highly controversial linguistic superphyla, such as Nostratic or Amerind, produced by quantitatively unsupported and non-replicable taxonomic practices (especially cf. Ringe's 1996 severe criticism, among others).

However, by virtue of recent advances, especially in linguistic theory, we are now in a much better position to overcome these objections and successfully readdress the issue. This claim is grounded in some considerations which seem to reply to each of the previous objections.

As for point 1, the idea that there is no reason to expect congruence between linguistic and genetic diversity now seems obsolete, after two decades of studies have proven that there is some empirically demonstrable degree of association between linguistic and genetic diversity (see also Cavalli-Sforza *et al*., 1994; Levinson & Gray, 2009), although it cannot be taken for granted for all regions of the planet. When choosing a partner, humans do not easily cross barriers, be they part of their physical or cultural environment. The consequences of this mating behavior have been shown to be substantial. In Europe, for instance, linguistic boundaries show increased rates of allele-frequency change (Barbujani & Sokal, 1990, among others), to the point that several inheritable diseases differ, in their incidence, between geographically close populations speaking different languages (de la Chapelle, 1993). In principle, any process of population admixture and language contact is expected to weaken the correspondence between genetic and linguistic diversity; therefore, it is all the more remarkable that in a large number of case studies such patterns appeared locally well correlated (Barbujani, 1991; Belle & Barbujani, 2007; Tishkoff *et al.*, 2009), confirming that quite often genetic and linguistic changes have occurred in parallel. The empirical question, then, is to what extent, in which parts of the world, and as a consequence of which demographic phenomena, linguistic and genetic features are correlated.

We contend that the objection in point 2 can be overcome to a large extent by a more accurate and purposeful choice of population/language pairs, selecting populations that can be unambiguously classified at both the linguistic and the anthropological level (Cavalli-Sforza *et al.*, 1988 drew genetic data and linguistic classifications from completely unrelated and poorly matching sources, without undertaking a dedicated joint venture with linguists).

In our view problem 3 is the crucial one. It is fair to say that, at the present stage of research, virtually no professional historical linguist unconditionally subscribes to the wide-range language trees used as matches in Cavalli Sforza's experiments, and most of them deny the very possibility of a reliable global classification of languages for serious methodological reasons. These reasons - as we claim below - can only now be successfully addressed by theoretical linguistics. The principal cause of the large-scale congruence debate having arrived at a virtual dead-end in recent years depends precisely on the methodological gap between the disciplines of linguistics and biology. Specifically, genetics has been able to assess similarities/differences on a global basis, i.e. among very distant populations, and to draw phylogenies by means of exact biostatistical methods; instead, no solid long-range comparison has so far been possible in phylogenetic linguistics and quantitative tools have only recently been adopted.

In biology, phylogenetic investigation has undergone a successful revolution owing to the progress of theoretical and experimental research, which has led to the discovery of a wealth of deep genetic polymorphisms, more likely to retain historical information than external morphological traits and hence able to generate robust phylogenies. In linguistics, no comparable phylogenetic exploitation of recent theoretical progress has been systematically attempted.

Indeed, virtually all attempts to systematically classify languages into families have so far been based on comparisons of words with similar meanings and similar forms. Such broadly 'lexical' methods of comparison, however, cannot safely establish etymological cognacy of words against chance across obvious and relatively shallow language families (Nichols, 1996; Ringe, 1996; Longobardi, 2003; Heggarty, 2006; Guardiano & Longobardi, 2005), and are especially unsuitable to precisely compute mathematical distances between two or more languages as well as the probability of their relatedness. An emerging body of pioneering research attempts to overcome this problem in Indo-European (e.g. Ringe *et al.*, 2002; Gray & Atkinson, 2003; MacMahon & MacMahon, 2005) and even across language families (Jäger, 2013). While promising results have been obtained, the fundamental obstacles are far from eliminated (Guardiano & Longobardi, 2016).

However, over the past decades, significant theoretical progress has revolutionized our understanding of diversity even in linguistics, focusing on a new domain of evidence, much more articulated and abstract than the lexicon, namely *grammatical* structure. Typological approaches (at least since Greenberg, 1963; Hawkins, 1983; Comrie, 1989, among many others), have uncovered presumable universals or constrained co-variation in grammar over an impressive amount of languages. Concurrently, the neurocognitive framework of the "biolinguistic" paradigm developed a deeper and more abstract study of grammatical structures (Chomsky, 1959; Lenneberg, 1967; Hauser *et al.*, 2002; Boeckx & Piattelli Palmarini 2005; Lightfoot, 2006; Fitch, 2010; Biberauer *et al.*, 2010; Di Sciullo & Boeckx, 2011). This approach has argued for the necessity of some species-invariant language properties (so-called Universal Grammar). However, most importantly from our perspective, the paradigm

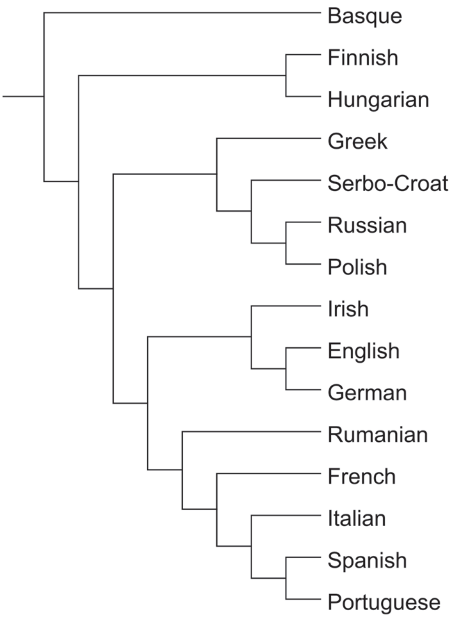

Fig. 1 - A UPGMA tree computed on the matrix of syntactic distances, from Longobardi* et al. *(2015).

centred on generative grammar has increasingly focused on species-internal, i.e. crosslinguistic, diversity. The fine structure of a few minimal grammatical contrasts between languages has been studied during the past few decades leading to the development of *parametric* models (cf. Chomsky, 1981; Lightfoot, 1991; Baker, 2001; Kayne, 2000; Biberauer, 2008, within a vast literature). The basic intuition of parametric theories of generative grammar is that the majority of observable syntactic differences among languages are derived, usually through complex deductive chains, from a smaller number of more abstract contrasts, drawn from a *universal* list of *discrete* (normally binary) options, called parameters.

Constructing an analogy between these parameters and genetic markers, by way of some inspiring hints from Roberts (1998), Longobardi (2003) proposed the use of parameters as taxonomic characters and suggested the possibility of a new phylogenetic research program. Since the core grammar of every natural language can in principle be represented by a string of binary symbols coding the value of a succession of parameters, such strings can easily be collated and used to define exact correspondence sets.

Owing to their *universality* and *discreteness*, parameter values as taxonomic characters can, in principle, precisely measure syntactic distances, even beyond firmly established language families, serving as a perfect input for computationally testable and replicable clustering hypotheses of the sort needed for a wide-scale comparison with genetic classifications.

No single shared binary parameter value can ever conclusively demonstrate kinship between two languages, of course: however, since parametric comparisons yield clear-cut answers (owing to discreteness), one can calculate probabilistic thresholds (Bortolussi *et al.*, 2011), an unrealistic objective for most lexical comparisons.

The hypothesis of Longobardi (2003) was precisely that the examination of relatively large sets of parameter values together, rather than of few at a time, could completely recall into question the traditional assumption that syntax is phylogenetically less informative than the vocabulary, and eventually demonstrate its congruence with the historical information carried by lexical comparison in the cases where the latter is possible and efficient.

Longobardi & Guardiano (2009) implemented the suggestions of Longobardi (2003) into a systematic procedure termed *Parametric Comparison Method* (henceforth PCM), showing that the distribution of calculated language distances is clearly non-random (calls for an explanation and lends itself to a historical one). As the next testing step, in Longobardi *et al.* (2013), the PCM has been applied to the classification of 26 modern Indo-European languages, recovering the correct internal articulation of the best known language family. Furthermore, the probability of relatedness for the closest language pairs returned by the PCM has been confirmed as statistically significant by a first large-scale computer-run experiment (Bortolussi *et al.*, 2011). Finally, it is also important to stress that the findings of such works about the role of parameters in reconstructing history provide further support of the validity of the generative approach to language variation, which embodies a notion of Universal Grammar at its core: in the case of a list of parameters the latter notion is minimally represented by the list itself (as distinct from their language-specific values) and by the implicational constraints connecting the parameters of such a list to each other.

At this point, the ongoing success of the PCM makes it possible to test its effectiveness and utility for calculating correlations between its syntactic distances and genetic distances inferred from a broad genetic dataset: indeed, through the PCM it is conceivable to address Darwin's congruence issue both beyond the limits of traditional language families and in a mathematically more precise and linguistically sophisticated way. To make Darwin's hypothesis empirically testable, it is important to split the issue into (at least) two questions: (1) can a gene-language parallelism indeed be identified? (2) if so, how much of it depends on common demic processes shaping genetic and linguistic diversity together?

To address these problems, we will examine some results obtained in Longobardi *et al.* (2015): 15 European languages (see the tree in Figure 1) from 3 different families (Indo-European, Finno-Ugric and Basque) were selected and analyzed in terms of the PCM methods and the same parametric database used and described in Longobardi *et al.* (2013), calculating their syntactic distances. A genetic distance matrix for the 15 corresponding populations (see the tree in Figure 2), based on 178,000 SNPs, was then calculated, along with a matrix of geographical (great circle) distances and one of lexical distances drawn from the recent IELex database (Bouckaert *et al.*, 2012). Since the IELex database targets only Indo-European languages (as its name hints), the maximum distance was assigned by default to cross-family language pairs.

Correlations between these four distance matrices were computed according to the Mantel (1967) procedure. The two linguistic matrices (syntactic and lexical) were highly correlated (r=0.850), and, as a consequence, showed very similar levels of correlation with genetic distances (0.60 for syntax and 0.54 for lexicon), both much higher than what we find between genes and geography (0.30). The results hold even when we control for geography: the partial Mantel correlation between syntax and genes is still strong (0.57) if we hold geography constant, while the genes-geography correlation is saliently lower (0.20), with syntax held constant.

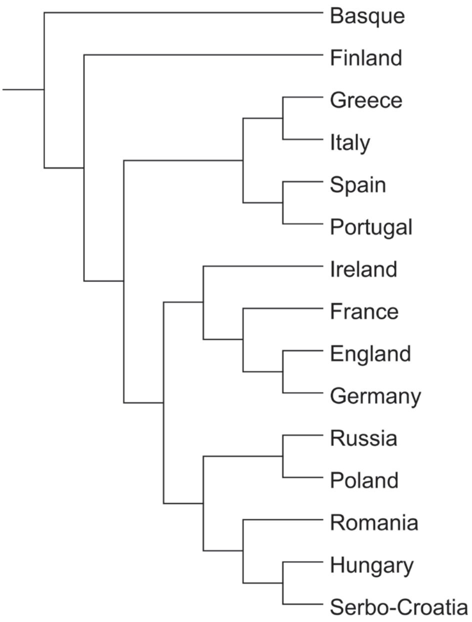

Fig. 2 - A UPGMA tree computed on the matrix of genetic distances, from Longobardi **et al.** ***(2015).***

Thus, a gene/language congruence seems to hold at a continent-wide scale (Europe), and to be independent of geographical distances: actually, language, once properly modeled through quantitative tools, proved a better predictor of European genetic diversity than geography (Novembre *et al.*, 2008).

Having reached a preliminary conclusion about the first Darwinian subquestion and gained some promising insight into the second, one can proceed to look in more detail at the demic history of European populations. Actually, in the comparison of the syntactic and biological phylogenetic trees (Figs. 1 and 2 respectively), the latter meets all the historical linguistic expectations, singling out correctly the three ancestral components (Basque, Finno-Ugric, Indo-European, and the subfamilies within it); the main elements of disagreement in the genomic one appeared to be the positions of Hungarians and Romanians, which cluster genetically with speakers of Serbo-Croatian despite being highly differentiated syntactically. The mismatch is especially salient for Hungarian, whose linguistic distance from IE languages as measured by the PCM is expectedly large.

The gene/language mismatch of Hungarians was noticed by Cavalli Sforza *et al.* (1994) and Greenhill (2011), even though it could not be quantified without a tool such as the PCM. Both ancient and modern genetic evidence presented in the literature has been invoked in Longobardi *et al.* 2015 to suggest precisely that no substantial demographic replacement occurred (Nadasi *et al.*, 2007; Tömöry *et al.*, 2007; Hellenthal *et al.*, 2014) when a Finno-Ugric language was introduced to what is now Hungary by a group of Asian immigrants during the historical 9th century invasion. On the contrary, the same case cannot be easily made for Basques and Finns, for which, to our knowledge, no available evidence suggests a similar model of partial demographic replacement associated with language replacement (Nelis *et al.*, 2009).

Unsurprisingly, then, the recalculation of the correlations after removing the apparent exception represented by Hungarians led Longobardi *et al.* (2015) to notice a further increase of that between genetic and syntactic distances (r=0.74), while the correlation with geography remained low (r=0.28). The skew became even sharper in partial Mantel tests (r=0.72 for gene/syntax with geography held constant, while it is r=0.093 for genes/geography with syntax held constant), providing the clearest demonstration to date of a language/biology correlation for the core of Europe.

In this article, we proceed to a new test of the ability of our tools to detect such special events of demographic history: we used the Distatis program, plotting syntactic, genetic and geographic distances into a single multidimensional scaling graph to identify some correlation patterns. The results are presented in Figure 3.

From a first analysis of the graph, it is clear that both genes and syntax are able to identify the two most obvious outliers, Finns and Basques, while their geographical distances ("G" on the graph) put them closer to their European neighbors. The same appears to be true, to a more reduced extent, for Greeks. Greek, basically an isolate among IE subfamilies, turned out in fact as the linguistic outlier of the IE languages of Europe also in previous experiments (Gray & Atkinson, 2003).

Most interestingly, even this test confirms the salient exception exhibited by Hungarians: based on syntactic distances ("S" on the graph), they fall close to the other Finno-Ugric population, i.e. Finns, and very far from those speaking Indo-European languages, while, from the viewpoint of biological anthropology ("B" on the graph), Hungarians appear to be indistinguishable from the other Central/Eastern European populations.

The further confirmation of these results now raises the question of explaining this exception: it is reasonable to speculate about the possibility of a non-accidental connection between the gene/language mismatch effects so discovered and the fairly recent (end of the 9th century) dating of the settlement of prince Árpád's tribes in present-day Hungary: the latter is the most recent among all those of the other European populations considered, whether IE, Finnic, or Basque.

The gene/language mismatch in Hungary suggests that language replacement there was not due to a radical population replacement. Phenomena of that kind were possible in prehistoric times, but unlikely in more densely populated medieval Europe. Such a language shift is compatible with the immigration of a limited group of (presumably male) individuals, whose arrival caused a major cultural shift, without deeply affecting the genetic makeup of the population, a situation certainly different from that of more ancient migrations into Europe, irrespective of their controversial dating and routing. Language replacements of this kind have been termed "elite dominance" by Colin Renfrew (1992).

In sum, the use of a new powerful linguistic tool, able to work across linguistic families and to provide quantitatively measurable conclusions, shows that populations speaking similar languages also tend to resemble each other at the genomic level, suggesting that cultural change and biological divergence have mostly proceeded in parallel in Europe. The partial correlation tests of Longobardi *et al.* (2015) and the Distatis results presented here confirm that populations speaking similar languages also tend to be genetically closer than expected on the sheer basis of their geographic location, so that language offers a better prediction of genomic distances than geography.

A number of previously unaddressable questions can now be put on the research agenda, such as the following:

a) do such gene/language correlations equally hold on more focused microareas (such as smaller areas of Europe or the Mediterranean), or on even wider geographical spaces (e.g. the whole of Eurasia), or, finally, on continents with radically different temporal and social peopling scales (say, the Americas)?
b) is there a detectable difference between the diffusion of syntax and that of lexicon in space and time?
c) do various linguistic (syntax, phonology, lexicon) and genetic (subportions of the genome, uniparental markers etc.) entities correlate with each other in different ways?

All these issues are currently under investigation. Of course, for many such inquiries the novel tool offered by the historical use of syntax and its potentially universal background hypotheses (as developed in the formal biolinguistic paradigm used here since at least Chomsky, 1981) has been, and will increasingly be, crucial in allowing comparisons of languages from different families, which by definition are supposed to share no common lexical etymology.

In everyday life, words are more superficially salient than grammatical rules, but the latter may provide more informative and explanatory tools, exactly as molecular markers in biology proved to do, despite being less pre-theoretically observable than external characters. The parallel moves toward abstract theoretical entities in these two domains of historical inquiry reflect their maturation as scientific disciplines, increasing the possibilities of their fruitful combination.

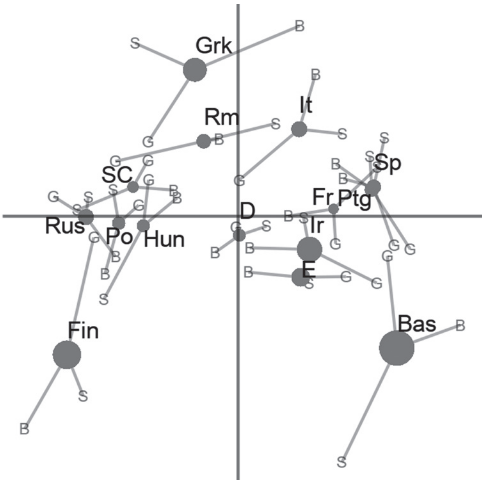

Fig. 3 - A Distatis plot of the (G)eographic, (B) iological/genetic, and (S)yntactic distance matrices from Longobardi* et al. *(2015). The languages are Greek, Romanian, Italian, Russian Serbo-Croatian, Polish, German (= D), French, Irish, English, Spanish, French, and Portuguese (all Indo-European); Finnish and Hungarian (Finno-Ugric) and Basque (a linguistic isolate).

Acknowledgements

This article was supported by ERC Adv. Grant 295736 LANGELIN. We are very grateful to Dr. Eugenio Bortolini for first drawing our attention to the Distatis method as a useful tool for visualizing the three-way correlations.

References

Baker M.C. 2001. *The atoms of language: the mind's hidden rules of grammar*. Basic Books, New York.

Barbujani G. 1991. What do languages tell us about human microevolution? *Trends Ecol. & Evol.,* 6:151-156.

Barbujani G. & Sokal R.R. 1990. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc. Natl. Acad. Sci. USA*, 87: 1816-1819.

Bateman R., Goddard I., O'Grady R., Funk V.A., Mooi R., Kress W.J. & Cannell P. 1990. Speaking with forked tongues: the feasibility of reconciling human phylogeny and the history of language. *Curr. Anthropol.,* 31: 1-24.

Belle E.M.S. & Barbujani G. 2007. A worldwide analysis of multiple microsatellites suggests that language diversity has a detectable influence on DNA diversity. *Am. J Phys. Anthropol.*, 133: 1137-1146.

Bellwood P. 2005. *First farmers. The origins of agricultural societies*. Blackwell, Oxford.

Biberauer T. (ed) 2008. *The Limits of Syntactic Variation*. Benjamins, Amsterdam.

Biberauer T., Holmberg A., Roberts I. & Sheehan M. (eds) 2010. *Parametric Variation: Null Subjects in Minimalist Theory*. Cambridge University Press, Cambridge.

Boeckx C. & Piattelli Palmarini M. 2005. Language as a natural object; Linguistics as a natural science. *The Linguistic Review*, 22: 447-466.

Bortolussi L., Longobardi G., Guardiano C. & Sgarro A. 2011. How many possible languages are there? *"In G. Bel-Enguix, V. Dahl & M.D. Jimenez-Lopez (eds): Biology, Computation and Linguistics,* pp. 168-179. *IOS Press,* Lansdale.

Bouckaert R., Lemey P., Dunn M., Greenhill S.J., Alekseyenko A.V., Drummond A.J., Gray R.D., Suchard M.A. & Atkinson Q.D. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337: 957-960.

Cavalli Sforza L.L., Menozzi P. & Piazza A. 1994. *The History and Geography of Human Genes*. Princeton University Press, Princeton.

Cavalli Sforza L.L., Piazza A., Menozzi P. & Mountain J. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. USA*, 85: 6002-6006.

Chomsky N. 1959. A review of B.F. Skinner's Verbal Behavior. *Language,* 35: 26-58.

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

Comrie B. 1989. *Language universals and linguistic typology*. University of Chicago Press, Chicago.

Darwin C. 1859. *On the Origin of Species*. John Murray, London.

de la Chapelle A. 1993. Disease gene mapping in isolated human populations: the example of Finland. *J. Med. Gen.*, 30: 857-865.

Di Sciullo A. M. & Boeckx C. (eds) 2011. *The Biolinguistic Enterprise. New Perspectives on the Evolution and Nature of the Human Language Faculty*. Oxford University Press, Oxford.

Fitch W.T. 2010. *The Evolution of Language*. Cambridge University Press, Cambridge.

Gray R.D. & Atkinson Q.D. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426: 435-439.

Greenhill S.J. 2011. Levenshtein distances fail to identify language relationships accurately. *Comput. Linguist.*, 37: 689-698.

Greenberg J. 1963. Some universals of grammar with particular reference to the order of meaningful elements. Universals of language. In J. Greenberg (ed): *Universals of Grammar,* pp. 73-113. Massachusetts Institute of Technology Press, Cambridge, MA.

Guardiano C. & Longobardi G. 2005. Parametric Comparison and Language Taxonomy. In: M. Batllori, M.-L. Hernanz, C. Picallo & F. Roca (eds): *Grammaticalization and Parametric Variation*, pp.149-174. Oxford University Press, Oxford.

Guardiano C., & Longobardi G. 2016. Formal syntax as a phylogenetic method. In: R.D. Janda, B. Joseph, B. Vance (eds): *Blackwell's handbook of historical linguistics Volume II*. Wiley/Blackwell, Hoboken.

Hauser M., Chomsky N. & Fitch T. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298: 1569-1579.

Hawkins J. 1983. *Word Order Universals*. Academic Press, New York.

Heggarty P. 2006. Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data - and to dating language? In J. Clackson, P. Forster & C. Renfrew (eds): *Phylogenetic methods and the prehistory of languages,* pp. 183-194. McDonalds Institute for Archaeological Research, Cambridge.

Hellenthal G., Busby G.B.J., Band G., Wilson J.F., Capelli C., Falush D. & Myers S. 2014. A genetic atlas of human admixture history. *Science*, 343: 747-751.

Jäger G. 2013. Lexikostatik 2.0. In A. Plewnia & A. Witt (eds): *Sprachverfall? Dynamik - Wandel - Variation,* pp. 197-216. Jahrbuch 2013 des Instituts für Deutsche Sprache, de Gruyter, Berlin.

Kayne R. 2000. *Parameters and Universals*. Oxford University Press, Oxford.

Lenneberg E. 1967. *Biological Foundations of Language*. John Wiley, New York.

Levinson S.C. & Gray R.D. 2012. Tools from evolutionary biology shed new light on the diversification of languages. *Trends Cogn. Sci.*, 16: 167-173.

Lightfoot D. 1991. *How to Set Parameters: Arguments from Language Change.* Massachusetts Institute of Technology Press, Cambridge, MA.

Lightfoot D. 2006. *How new languages emerge*. Cambridge University Press, Cambridge.

Longobardi G. 2003. Methods in Parametric Linguistics and Cognitive History. *Linguistic Variation Yearbook*, 3: 101-138.

Longobardi G. & C. Guardiano. 2009. Evidence for Syntax as a Signal of Historical Relatedness. *Lingua*, 119: 1679-1706.

Longobardi G., Guardiano C., Silvestri G., Boattini A. & Ceolin A. 2013. Toward a syntactic classification of Indo-European languages. *Journal of Historical Linguistics*, 3: 122-152.

Longobardi G., Ghirotto S., Guardiano C., Tassi F., Benazzo A., Ceolin A. & Barbujani G. 2015. Across Language Families: Genome diversity mirrors linguistic variation within Europe. *Am. J. Phys. Anthropol.*, 157/4: 630-640.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27: 209-220.

McMahon A. & McMahon R. 2005. *Language Classification by Numbers*. Oxford University Press, Oxford.

Nadasi E., Gyurus P., Czako M., Bene J., Kosztolanyi S., Fazekas S., Dömösi P. & Melegh B. 2007. Comparison of mtdna haplogroups in Hungarians with four other European populations:a small incidence of descents with Asian origin. *Acta Biol. Hungarica*, 58: 245-256.

Nelis M., Esko T., Mägi R., Zimprich F., Zimprich A., Toncheva D., Karachanak S., Piskackova T., Balascak I., Peltonen L. *et al*. 2009. Genetic structure of Europeans: a view from the north-east. *PLoS One*, 4: e5472.

Nichols J. 1996. The Comparative Method as Heuristic. In M. Durie & M. Ross (eds): *The Comparative Method Reviewed: Regularity and Irregularity in Language Change,* pp. 39-71. Oxford University Press, Oxford.

Novembre J., Johnson T., Bryc K., Kutalik Z., Boyko A.R., Auton A., Indap A., King K.S., Bergmann S., Nelson M.R., Stephens M. & Bustamante C.D. 2008. Genes mirror geography within Europe. *Nature*, 456: 98-101.

Renfrew C. 1992. Archaeology, genetics and linguistic diversity. *Man*, 27: 445-478

Ringe D. 1996. The mathematics of Amerind. *Diachronica*, 13: 135-154.

Ringe D., Warnow T. & Taylor A. 2002. Indo-European and computational cladistics. *Trans. Philol. Soc.*, 100: 59-129.

Roberts I. 1998. Review of Harris and Campbell 1995. *Romance Philology*, 51: 363-370.

Sokal R.R. 1988. Genetic, geographic, and linguistic distances in Europe. *Proc. Natl. Acad. Sci. USA*, 85:1722-1726.

Tishkoff S.A., Reed F.A., Friedlaender F.R., Ehret C., Ranciaro A., Froment A., Hirbo J.B., Awomoyi A.A., Bodo J-M. & Doumbo O. *et al.* 2009. The genetic structure and history of Africans and African Americans. *Science*, 324: 1035-1044.

Tömöry G., Csanyi B., Bogacsi-Szabo E., Kalmar T., Czibula A., Csosz A., Priskin K., Mende B., Lango P., Downes C.S. *et al*. 2007. Comparison of maternal lineage and biogeographic analyses of ancient and modern Hungarian populations. *Am. J Phys. Anthropol.*, 134: 354-368.