Toward a syntactic phylogeny of modern Indo-European languages*

Giuseppe Longobardi[a,b], Cristina Guardiano[c], Giuseppina Silvestri[a,d], Alessio Boattini[e] and Andrea Ceolin[a]
University of Trieste[a] / University of York[b] / University of Modena e Reggio Emilia[c] / University of Pisa[d] / University of Bologna[e]

The Parametric Comparison Method (PCM, Guardiano & Longobardi 2005, Longobardi & Guardiano 2009) is grounded on the assumption that syntactic parameters are more appropriate than other traits for use as comparanda for historical reconstruction, because they are able to provide unambiguous correspondences and objective measurements, thus guaranteeing wide-range applicability and quantitative exactness. This article discusses a set of experiments explicitly designed to evaluate the impact of parametric syntax in representing historical relatedness, and performed on a selection of 26 contemporary Indo-European varieties. The results show that PCM is in fact able to correctly identify genealogical relations even from modern languages only, performing as accurately as lexical methods, and that its effectiveness is not limited by interference effects such as 'horizontal' transmission. PCM is thus validated as a powerful tool for the analysis of historical relationships not only on a long-range perspective (as suggested by Longobardi & Guardiano 2009), but even on more focused, though independently well-known domains.

Keywords: Parametric Comparison Method, historical reconstruction, syntactic distances, quantitative phylogenies, nominal domain

1. Introduction

The historical classification of Indo-European languages is traditionally based on the inspection of **lexical** entities (roots and grammatical morphemes) and of sound laws connecting cognate words crosslinguistically. In contrast, syntactic structures have rarely been regarded as encoding relevant genealogical information. Such supposed lack of parallelism between syntactic and lexical classification was most explicitly challenged in Longobardi & Guardiano (2009). That article used a

Journal of Historical Linguistics 3:1 (2013), 122–152. DOI 10.1075/jhl.3.1.07lon
ISSN 2210-2116 / E-ISSN 2210-2124

new taxonomic technique crucially based only on syntax, namely the Parametric Comparison Method (PCM), which is built on the insights of modern generative theory. Its main claim is that even syntax codes for phylogenetic history once a sufficient number of characters is considered.

The possibility of a Parametric Comparison Method was first exposed in Longobardi (2003). In Guardiano & Longobardi (2005), it was shown that the analysis of a few languages in terms of PCM could satisfy certain conceptual conditions of adequacy (scattering of language distances, the Anti-Babelic principle). Longobardi & Guardiano (2009) argued that PCM provided language distances and preliminary taxonomies in good empirical agreement with some independent historical evidence, within and outside of the Indo-European family. Bortolussi et al. (2011) argued that the probability of many of the language distances that are empirically calculated by PCM is significant beyond chance, and hence calls for a historical explanation. Finally, Longobardi (2012) argued that some apparently paradoxical instances of parallel developments can be readily explained by the structure of the syntactic theory underlying PCM.

Thanks to its reliance on a detailed theory of Universal Grammar (in particular a universal list of parameters), PCM is in principle capable of comparing any pair or set of natural languages. Actually, it was devised in order to potentially address unsolved long-range genealogical questions; to fully assess and increase its adequacy, however, it is crucial to first test its performance in domains whose genealogy is already known. Furthermore, it is important to test PCM's ability to reconstruct chronologically deep phylogenies using exclusively modern language data, often the only available data outside Eurasia.

This article describes some experiments performed on a selection of 26 contemporary Indo-European varieties belonging to the Romance, Greek, Germanic, Celtic, Slavic, Indic and Iranian families. With respect to previous PCM experiments (cf. Longobardi & Guardiano 2009), the present article relies on:

a. Nine additional contemporary IE languages analyzed parametrically: Sicilian, (Northern) Calabrese, Bovese Greek, Danish, Icelandic, Slovenian, Polish, Farsi, and Marathi.
b. A much more refined grid of nominal parameters. The list and a brief description (the first of its kind) of the 56 parameters used, defining the variability of a single module (the internal structure of nominal phrases), is contained in the electronic support material (see the Appendix).
c. A wider range of statistical procedures (UPGMA trees, SplitsTree package, distribution and correlation tests, Structure, PAUP*).

The experiments produced two major results:

1. The current version of PCM identifies the main subfamilies of Indo-European strikingly well, even from modern languages only, performing genealogically as accurately as lexical methods.
2. 'Horizontal' transmission (interference) does not seem to limit the effectiveness of PCM to the extent of seriously undermining the correct representation of the main 'vertical' relations.

We contend, thus, that generative grammar, and more generally the bio-cognitive framework of which it is a salient part (Roberts to appear), is not only an insightful theory of mind (Chomsky 1975) and of its synchronic (Chomsky 1981 and following work) and diachronic (Lightfoot 1979) variability: it can also become a true historical science, capable of gaining insights into the actual (pre)history of human populations, no less than the successful historical–comparative enterprise of the 19th century. In turn, such historical success of generative parameters brings about new qualitative evidence that parameters are really at work in language transmission through time, hence, ultimately, in the actual processes of language acquisition.

This article is organized as follows: Section 2 provides an outline of the recent debate about the reconstruction of genealogical relationships within Indo-European on a quantitative/statistical basis; Section 3 discusses a class of experiments performed over the novel database of syntactic (parametric) data with the aid of a selection of distance-based taxonomic algorithms of biostatistic derivation, and compares the results with those provided by lexical taxonomic characters (i.e. the cognacy judgments provided in Dyen et al.'s 1992 database); Section 4 proposes two preliminary tests with character-based algorithms, and provides comments about the role of plausible interference effects (i.e. contact) on the taxonomies produced; Section 5 lists the relevant conclusions.

2. Indo-European classifications on a quantitative basis

2.1 Quantitative taxonomies

Many subfamilies of Indo-European, often identified long before the establishment of the classical comparative method, are undisputed. The hypotheses of intermediate units (aggregations across subgroups, position of isolates) are more debated, such as, for instance, the Balto-Slavic hypothesis (Schleicher 1861–1862, Szemerényi 1957, *contra* Meillet 1905, 1924) or the Italo-Celtic one (Meillet 1922, also cf. recently, e.g., Ringe et al. 2002). Even the model of Indo-Iranian as a secondary proto-language simply diverging into Indic and Iranian (and Nuristani), reconstructed mainly from their earliest literary stages of the former two (Meillet 1922), has been occasionally called into question (Lazzeroni 1968).

In recent years, all such questions are being readdressed with the aid of statistical and computational methods. This quantification of relatedness across IE languages has still been mostly based on lexical datasets: the most frequently adopted is Dyen et al.'s (1992) 'Comparative Indo-European data corpus', where a 200-item Swadesh list (Swadesh 1950) is translated into 95 Indo-European languages, with (idealized) binary cognacy judgments provided for all pairs of synonyms.[1]

Phylogenetic computational procedures derive mainly from those of molecular biologists and, according to the nature of the input, fall into two main types: distance-based and character-based. Character-based algorithms deal with matrices of finite-state characters, and the chronological distance of each single taxonomic unit from the root is calculated without assuming any uniform rate of evolution across the tree. Conversely, distance-based algorithms are more flexible with respect to the nature of the input (i.e., they can work with different typologies of data, *a priori* uniformed by the distance metrics), but are subject to rigid restrictions on the form of the output; for instance, most of them assume a constant rate of evolution across the tree, which might produce misleading effects on the topologies (cf. Section 3.3 below).

An extensive list of effective linguistic implementations of algorithms from the classic PHYLIP package (*Neighbour*, *Fitch* and *Kitsch*: Felsenstein 1993), all empirically relying on Dyen et al. (1992), is reported in McMahon & McMahon (2003, 2005), with the purpose of singling out those most appropriate for linguistic data; they suggest that distance-based algorithms are more viable than character-based ones. In the same vein, Nakhleh et al. (2005) discuss various experiments on six reconstruction methods (both character-based and distance-based) and on four different versions of an Indo-European lexical database (for further discussion cf. at least Lohr 1998, van Cort 2001, Rigon 2009, 2012, Barbançon et al. to appear).

In contrast, some other scholars have used algorithms directly, taking linguistic characters as input, e.g. works by Gray, Atkinson and their collaborators, and Ringe et al. (2002). Gray & Atkinson's (2003) input is based on Dyen et al.'s (1992) dataset; their experiments are more explicitly aimed at dating splits and testing the two most disputed hypotheses on the origin of Indo-Europeans, the 'Kurgan expansion' (Gimbutas 1973) and 'Anatolia farming' (Renfrew's 1987 'Neolithic Discontinuity Theory'). Ringe et al.'s (2002) work is notable because it develops a phylogenetic algorithm specifically conceived for the analysis of language relationships, and stresses the practical difficulties in assigning discrete values to lexical correspondences. They adopt taxonomic characters of three types: root correspondences (2002: 333), sound correspondences (2002: 22) and correspondences in inflectional endings (2002: 15), all encoded as binary. Their experiments across 24 Indo-European ancient and modern languages suggest the possibility of constructing a stable evolutionary tree for Indo-European; their conclusions sustain

the Italo–Celtic hypothesis with "respectable support" (Ringe et al. 2002: 112), and Greek–Armenian unity, though with "significantly poorer" evidence (Ringe et al. 2002: 102); no counterevidence is prompted against Balto-Slavic and Indo-Iranian.

2.2 Beyond classical trees

The methods described above are mainly used to produce evolutionary trees, essentially still in the tradition of Schleicher's (1861–1862) *Stammbaum*: they often assume a root, and always one single possible path between each pair of leaves. As is known, such classical graphs only encode genealogical (i.e. **vertical**) transmission and cannot represent secondary convergence (i.e. **horizontal** transmission), to which lexical characteristics are particularly sensitive (Thomason & Kaufman 1988, Thomason 2001, 2004, 2010, among others). In biology, network graphs have been elaborated precisely to deal with "cases where a particular genetic sequence has more than one possible history" (McMahon & McMahon 2005: 140); networks collapse different trees into a single representation of all possible transmission paths. When different possible branchings are in conflict, a tree must choose one, while a network provides a reticulation, i.e. it depicts all the possible branches together. Thus, networks represent historical relations without favoring vertical over horizontal transmission. For this reason, network programs have been advocated in linguistics as well (cf., among many others, McMahon & McMahon 2005, McMahon et al. 2005, McMahon 2010, Noonan 2010, Nelson-Sathi et al. 2010), under the traditional (essentially since Schmidt's 1872 *Wellentheorie*) idea that "the diversification of the Indo-European family must be modeled at least in part as a network rather than a tree" (Ringe et al. 2002: 110).

3. Indo-European classification on a syntactic basis

3.1 The rise of PCM

The computational attempts to analyze relations within Indo-European have been encouraging, though hardly uncontroversial. One difficulty, perhaps, is that the programs adopted are only partially adequate to deal with language history (Nakhleh et al. 2005); but another may be that lexical characteristics are not unconditionally suitable for effective measurement, because many lexical correspondences are only arbitrarily reducible to discrete units. Longobardi & Guardiano (2009) list and briefly exemplify four subcases, in which relevant affinities between lexical items cannot be easily reduced to binary, or perhaps even discretely computable correspondences (partial correspondence of form, partial correspondence

of meaning, correspondence of form with no correspondence of meaning, identical peculiar semantic shifts in the same cultural area with no correspondence of form). Many such difficulties are well represented in the attempt to compute relative distances among IE languages from, say, four unquestionable cognates like German *gießе* 'I pour', Italian *fondo* 'I melt', Sanskrit *juhomi* 'I pour', and Ancient Greek χέω 'I pour', which all variously differ from each other regarding reduplication, *Ablaut* grade, suffix, infixation, ending, or meaning.

Using sound laws (still a broad-sense lexical device, in our terminology, since they are based on the arbitrary distribution of sounds across the vocabulary) circumvents some of these problems, but others may arise. For instance, if a sound law has different subclauses (e.g. the three shifts of Grimm's law), how many identities/differences should be counted between two languages with respect to it? The problem is especially acute for, e.g., microvariation in the German second consonant shift, which is dialectally quite non-uniform.

In any case, some non-trivial arbitrary manipulation seems inevitable to turn lexical data into input characters for precise quantitative experiments. Therefore, the idea of considering taxonomic characters beyond the vocabulary has often been advocated (Nichols 1992, Longobardi 2003, Heggarty 2004, Dunn et al. 2005, among others). Longobardi (2003), developing some ideas from Roberts (1998), suggested that syntactic parameters are particularly suited to function as comparanda for phylogenetic reconstruction, because they require little arbitrary idealization, since they are endowed by definition with formal properties (discreteness and universality) that guarantee longest-range comparison, quantitative exactness, and no ambiguity of settings (Chomsky 1981, Lightfoot 1991, Baker 1996, Kayne 2000, Biberauer 2008, in a vast literature).

While parameter values can in principle measure syntactic distances within any set of languages, no single binary parameter can, of course, prove kinship between two languages; however, since each parametric comparison yields a clear-cut answer, one can calculate probabilistic thresholds (Bortolussi et al. 2011). Longobardi (2003) also suggested, as a strategy of realistic size, to study relatively many parameters in relatively many languages, at the acceptable cost of focusing on one compact module of grammar (Modularized Global Parametrization, MGP). Among its advantages, MGP allows for a better identification of cross-parametric implications and for a probabilistically sounder sample/population ratio.

3.2 The parametric database

An updated database has been used in the parametric experiments presented in this article. It consists of 56 binary parameters, all, in agreement with MGP, defining properties of nominal structures, identified on the basis of the existing

literature (Longobardi 1994, 2001, 2005, 2008, Plank 2003, Alexiadou et al. 2007, Ghomeshi et al. 2009, Keenan & Paperno 2012, among many others) and ongoing investigation. A salient feature of this type of datasets, stressed in Longobardi & Guardiano (2009), is that in many cases the value of a parameter is entirely

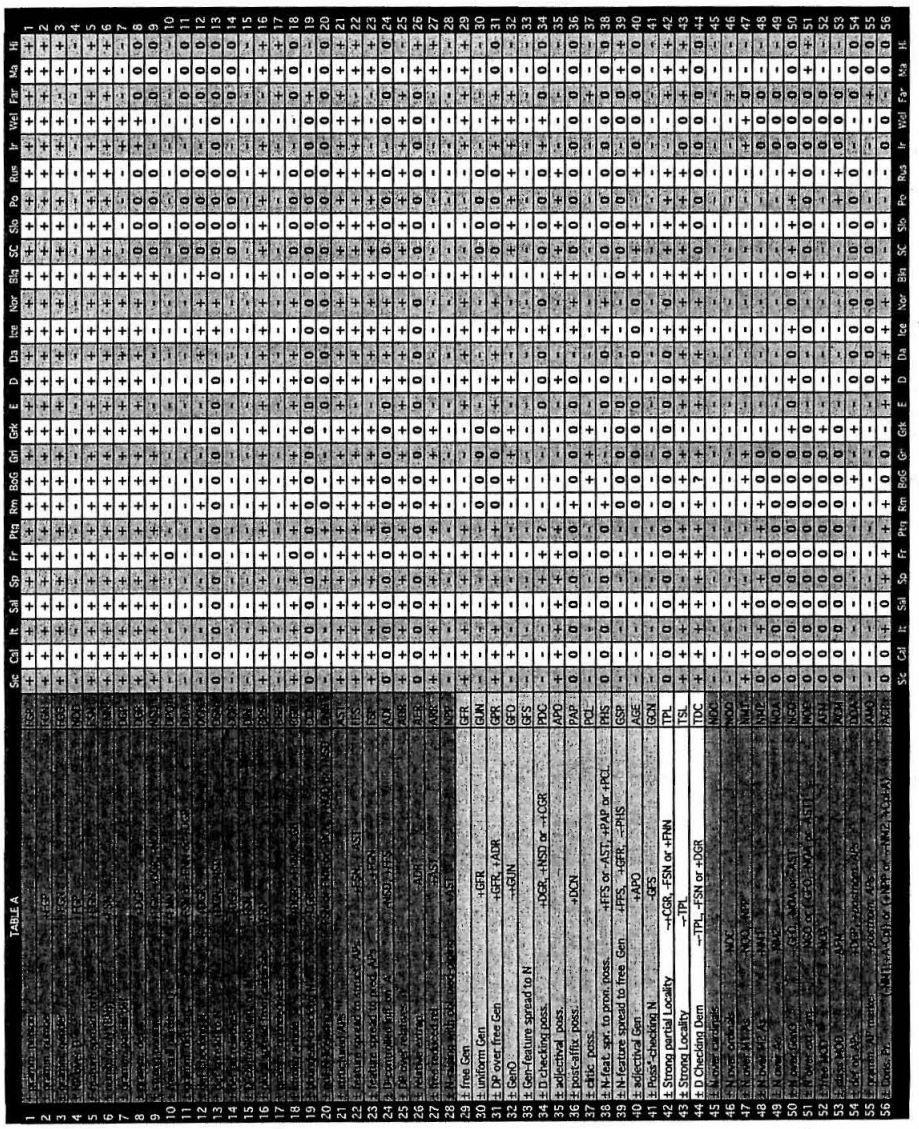


Figure 1. Table A (56 nominal parameters and their settings in 26 contemporary IE languages)

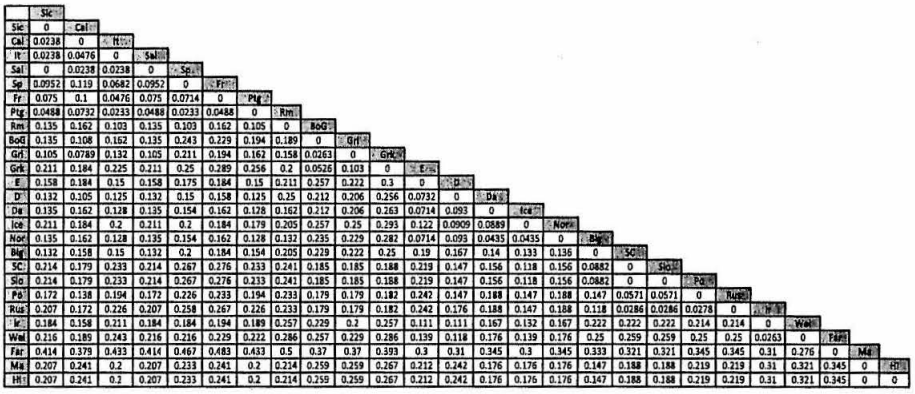
predictable from those of other parameters (also cf. Baker 2001), and thus represents information to be completely disregarded for taxonomic purposes. Figure 1 presents the parameters and their settings, in the format introduced in Longobardi & Guardiano (2009). The alternative states are encoded as '+' and '–', while the neutralized states resulting from implications across parameters are coded as '0'; the conditions determining 0 for a parameter are indicated next to the name of that parameter in the second column. A few empirically uncertain states, indicated by '?', are also treated as 0s for the purposes of taxonomic computations. A brief description of the 56 nominal parameters, as well as the illustration of their implications, is available in the electronic support material (see the Appendix).

The phylogenetic experiments have been performed on two distinct sets of Indo-European languages. One includes 26 contemporary varieties (Sicilian: Sic, Northern Calabrese: Cal, Italian: It, Salentino: Sal, Spanish: Sp, French: Fr, Portuguese: Ptg; Rumanian: Rm; Bovese Greek of Southern Calabria: BoG; Grico, i.e. Greek of Salento: Gri, standard Greek: Grk, English: E, German: D, Danish: Da, Icelandic: Ice, Norwegian: Nor, Bulgarian: Blg, Serbo-Croatian: SC, Slovenian: Slo, Polish: Po, Russian: Rus, Irish: Ir, Welsh: Wel, Farsi: Far, Marathi: Ma, Hindi: Hi). The other consists of the subset of 21 such varieties overlapping with those of Dyen et al.'s (1992) database (which does not include the five Greek and Romance dialectal varieties of Southern Italy).

3.3 Phylogenetic algorithms

Among phylogenetic algorithms, character-based ones, though more precisely representing language history (Rigon 2009, Barbançon et al. to appear), are less suitable in our context for two reasons. First, the implications among parameters, very pervasive within the same module, violate their crucial assumption of character independence. Second, character-based algorithms produce more informative results when the input is supplemented with additional information, especially regarding the directionality of change (e.g. the markedness theory assumed in Ringe's experiments) or the characters' stability (i.e. the possibility of assigning each character a weight, according to how resistant it appears to change). At the present stage, we cannot yet rely on a solid theory of parameter resetting, therefore no such information was included in the input. Thus, distance-based methods appear more readily applicable: they minimize the effects of internal dependencies across parameters, and allow phylogenetic hypotheses even in the absence of a refined theory of diachronic change.

Following Longobardi & Guardiano (2009), the number of parametric identities and differences for each pair of languages was first represented as an ordered pair of non-negative numbers $\langle i;d \rangle$ and then normalized to a monadic figure, the



	Sic	Cal	It	Sal	Sp	Fr	Ptg	Rm	BoG	Grl	GrM	E	D	Da	Ice	Nor	Big	SC	Slo	Pol	Rus	It	Wel	Fär	Mlt	Hit
Sic	0																									
Cal	0.0238	0																								
It	0.0238	0.0476	0																							
Sal	0	0.0238	0.0238	0																						
Sp	0.0952	0.119	0.0682	0.0952	0																					
Fr	0.075	0.1	0.0476	0.075	0.0714	0																				
Ptg	0.0488	0.0732	0.0233	0.0488	0.0233	0.0488	0																			
Rm	0.135	0.162	0.103	0.135	0.103	0.162	0.105	0																		
BoG	0.135	0.108	0.162	0.135	0.243	0.229	0.194	0.189	0																	
Grl	0.105	0.0789	0.132	0.105	0.211	0.194	0.162	0.158	0.0263	0																
GrM	0.211	0.184	0.225	0.211	0.25	0.289	0.256	0.2	0.0526	0.103	0															
E	0.158	0.184	0.15	0.158	0.175	0.184	0.15	0.211	0.257	0.222	0.3	0														
D	0.132	0.105	0.125	0.132	0.15	0.158	0.125	0.25	0.212	0.206	0.256	0.0732	0													
Da	0.135	0.162	0.128	0.135	0.154	0.162	0.128	0.162	0.212	0.206	0.263	0.0714	0.093	0												
Ice	0.211	0.184	0.2	0.211	0.2	0.184	0.179	0.205	0.257	0.25	0.293	0.122	0.0909	0.0889	0											
Nor	0.135	0.162	0.128	0.135	0.154	0.162	0.128	0.132	0.235	0.229	0.282	0.0714	0.095	0.0435	0.0435	0										
Big	0.132	0.158	0.15	0.132	0.2	0.184	0.154	0.205	0.229	0.222	0.25	0.19	0.167	0.14	0.133	0.136	0									
SC	0.214	0.179	0.233	0.214	0.267	0.276	0.233	0.241	0.185	0.185	0.188	0.219	0.147	0.156	0.118	0.156	0.0882	0								
Slo	0.214	0.179	0.233	0.214	0.267	0.276	0.233	0.241	0.185	0.185	0.188	0.219	0.147	0.156	0.118	0.156	0.0882	0	0							
Pol	0.172	0.138	0.194	0.172	0.226	0.233	0.194	0.233	0.179	0.179	0.182	0.242	0.147	0.188	0.147	0.188	0.147	0.0571	0.0571	0						
Rus	0.207	0.172	0.226	0.207	0.258	0.267	0.226	0.233	0.179	0.179	0.182	0.242	0.176	0.188	0.147	0.188	0.118	0.0286	0.0286	0.0278	0					
It	0.184	0.158	0.211	0.184	0.184	0.194	0.189	0.257	0.229	0.2	0.257	0.111	0.111	0.167	0.132	0.167	0.222	0.222	0.222	0.214	0.214	0				
Wel	0.216	0.189	0.243	0.216	0.216	0.229	0.222	0.286	0.257	0.229	0.286	0.139	0.118	0.176	0.139	0.176	0.25	0.259	0.259	0.25	0.25	0.0263	0			
Fär	0.414	0.379	0.433	0.414	0.467	0.483	0.433	0.5	0.37	0.37	0.393	0.3	0.31	0.345	0.3	0.345	0.333	0.321	0.321	0.345	0.345	0.31	0.276	0		
Mlt	0.207	0.241	0.2	0.207	0.233	0.241	0.2	0.214	0.259	0.259	0.267	0.212	0.242	0.176	0.176	0.176	0.147	0.188	0.188	0.219	0.219	0.31	0.321	0.345	0	
Hit	0.207	0.241	0.2	0.207	0.233	0.241	0.2	0.214	0.259	0.259	0.267	0.212	0.242	0.176	0.176	0.176	0.147	0.188	0.188	0.219	0.219	0.31	0.321	0.345	0	0

Figure 2. Matrix of parametric (syntactic) distances

Jaccard-Tanimoto distance ($d/(d+i)$), to neutralize most drawbacks of the cross-parametric dependencies. Hence, two identically set languages will have distance 0, two with all opposite settings distance 1, all other cases falling in between. Such distances are represented in Figure 2.

The algorithms that turned out to be most effective with parametric data (Rigon 2009, Longobardi & Guardiano 2009), i.e. UPGMA (Sokal & Michener 1958) and *Kitsch* (from the PHYLIP package, Felsenstein 1993), have the further property of producing 'ultrametric trees', i.e. trees complying with the 'molecular clock hypothesis' (Bromham & Penny 2003, Felsenstein 1993, 2004): in the rooted trees produced, the length of each branch, that is the distance between the root and every single leaf, is uniform. This requires all the languages to be chronologically equidistant from the root; the method treats all taxonomic units as leaves, i.e. it does not place any language on non-terminal (ancestral) nodes; a tree from languages of different chronological attestation is only possible under the risky idealization that they can be all viewed as coeval. For the purpose of argument rigor, this idealization has not been pursued here; therefore, the five ancient varieties of Longobardi & Guardiano's (2009) database (i.e., Latin, Classical Greek, New Testament Greek, Gothic and Old English) have not been included in the present experiments.

3.4 Distance-based trees and networks

The rooted trees in Figures 3–4 have been obtained from the syntactic distances through UPGMA and *Kitsch*, after bootstrapping the input with 1,000 resamples. Bootstrapping is a resampling procedure aimed at testing the solidity of the classification (the robustness of each single branching and of the tree as a whole): the data are resampled sequentially, i.e. a given percentage of the characters is

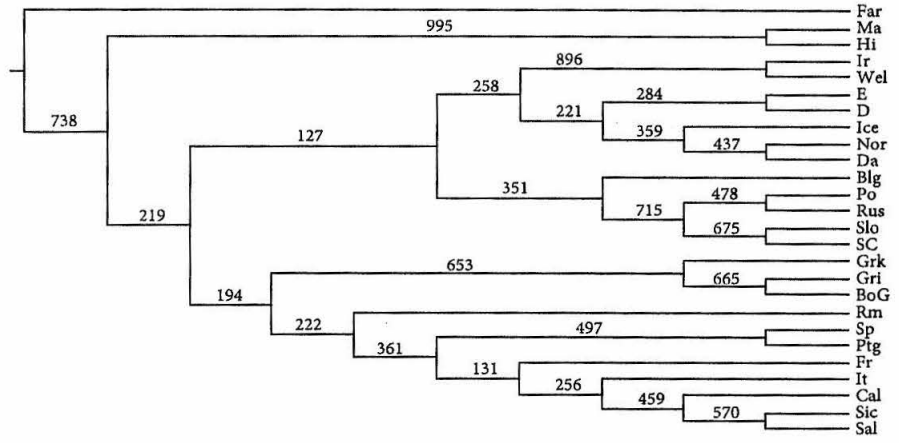


Figure 3. UPGMA (bootstrapped) tree from the syntactic distances in Figure 2

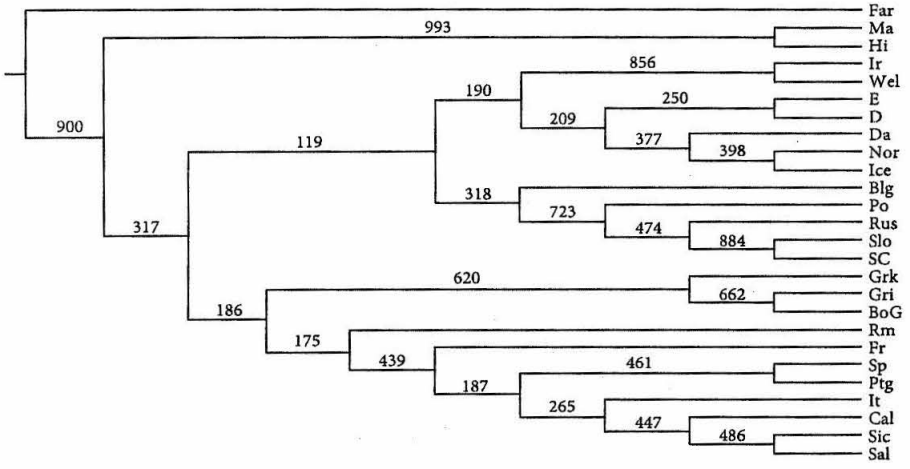


Figure 4. *Kitsch* (bootstrapped) tree from the syntactic distances in Figure 2

removed and replaced by others randomly chosen within the same corpus, so that new matrices are produced and in turn used as input for further tree-experiments. The procedure can be repeated thousands of times, according to the size of the database. The final result is a tree where each non-terminal node is labeled with the number of the iterations supporting that particular branching.

The corresponding networks (Figures 5–6) have been generated from the algorithms *ConsensusTree* and *Equalangle* (contained in the 'SplitsTree' package, Huson & Bryant 2006), using as an input the same 1,000 trees resampled through UPGMA and *Kitsch*, respectively.

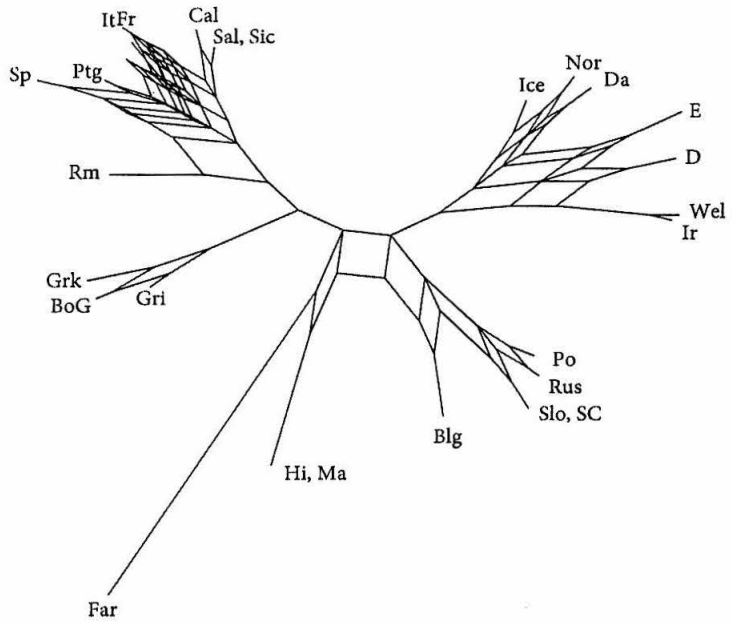


Figure 5. Network from 1,000 trees resampled by UPGMA

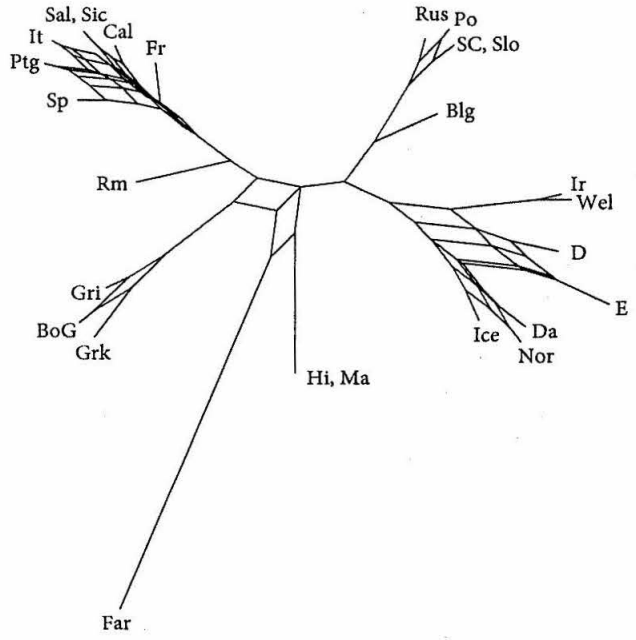


Figure 6. Network from 1,000 trees resampled by *Kitsch*

The trees, as well as the networks, are remarkably consistent with each other, and with the suggestions of traditional scholarship.

The Romance languages have been grouped together in all experiments, though bootstrapping values in the trees are lower than 500 for most of the nodes.[2] Rumanian always appears as the outlier; Iberian unity is well-represented, and the Italian varieties are appropriately placed, with the three Southern dialects (Northern Calabrese, Sicilian and Salentino) clustered together, and their internal articulation reflecting the traditional separation (Pellegrini 1977) between 'Extreme' Southern dialects (here Salentino and Sicilian) and 'Intermediate' Southern ones (Northern Calabrese). Only the position of French is slightly variable: under several replications, in the trees it shows up as an outlier of the Italian group, of the Iberian cluster, or even of a larger cluster including both.[3] Rumanian is never attracted towards any other group in particular; its isolation, and its relations with the other languages of the area, may have produced some divergence from the core of the ancestral family, but actual convergence with the surrounding Balkan languages is limited (cf. Section 4.3 below).[4]

The three Greek varieties (standard Greek, Grico and Bovese Greek) are included under one and the same node, with significant bootstrapping values (653 UPGMA, 620 *Kitsch*). The two of Southern Italy are clustered apart (665 UPGMA, 662 *Kitsch*). Thus, the very obvious contact between Romance and Greek in Southern Italy (Guardiano & Stavrou to appear) is not overrepresented: the two Greek dialects are never misclustered with any of the Romance languages rather than with homeland Greek.

Slavic unity is also well recognized in trees and networks. One disparity between PCM and traditional evidence concerns its internal articulation, namely, only two of the three Southern varieties, i.e. Serbo-Croatian and Slovenian, are clustered together (with high bootstrapping values in both trees: 675 UPGMA, 884 *Kitsch*), while Bulgarian systematically shows up as the outlier of the whole family. This is a taxonomic **factual mistake** (Guardiano & Longobardi 2005: 166): given three languages X, Y and Z, X turns out to be closer to Y than to Z (or vice versa), while the opposite is true according to independently well-established taxonomies. The cluster of all the Slavic languages, excluding Bulgarian, is supported by a significant bootstrapping index (715 UPGMA, 723 *Kitsch*); on the contrary, the node including Bulgarian is associated with smaller values. Such a position of Bulgarian is also replicated in the two networks. Here, all the Slavic varieties are grouped together, while Bulgarian appears on a separate branch. Yet, precisely as with Rumanian, no evidence of systematic attraction from one external group is visible; indeed, no particular branch (or reticulate of branches) connects Bulgarian to a non-Slavic language. Actually, Bulgarian contrasts with the rest of Slavic in three parameters. In p7 (DGP) and p29 (GFR), Bulgarian exhibits a '+' (like many

non-Slavic languages of the sample), while the other Slavic languages homogeneously select '–'.[5,6] In p32 (GFO), Bulgarian exhibits a '–', while the rest of Slavic homogeneously selects '+'. Precisely two of the mentioned parameters (p7, DGP and p29, GFR) are responsible for intricate cascades of implications, so that their values in the Slavic languages other than Bulgarian trigger a high number of zeros. Now, as noted, the distance measure adopted here successfully normalizes the zeros in the assessment of similarities/differences, but cannot avoid their reducing the potential for comparison and the overall balance of differences and identities. In any case, contrary to some recent literature (for recent contributions, see at least Joseph 1999, Thomason 2000, Tomić 2003), which suggests that the grammar is even more sensitive to contact than the lexicon within the Balkan *Sprachbund*, PCM misplaces none of the languages of the area outside its genealogical family.

The two Celtic languages are systematically clustered together, with remarkably high bootstrapping values (896 UPGMA, 856 *Kitsch*).

The Germanic group is recognized as well, and its internal articulation into Northern and Western varieties is acknowledged, though less significantly supported by bootstrapping values.

Celtic and Germanic are then clustered together in the two trees: even though such a higher cluster is not supported by significant bootstrapping values, it recurs in the experiments performed. The two groups also appear in one and the same region of the networks, which (especially the one produced through UPGMA) confirm the existence of the two families but also their closeness, particularly affecting the Western branch of Germanic. Although between English and Welsh convergence effects have unquestionably existed (Willis 2008), within Figure 1 only the value of parameter 23 (FSP) seems identical in Celtic and West Germanic to the exclusion of their neighbors or kin (it is only further shared with remote Farsi) and thus may represent a common secondary innovation; for the rest, the whole extent of Celto–Germanic similarity is not obviously traceable to particular points of our dataset. No experiments, on the contrary, provide any evidence for an Italo–Celtic unit.

The most interesting point concerns the Indic and Iranian languages. In the trees, the two Indic languages, displaying identical nominal syntax, correctly cluster under the same node, with the highest bootstrapping values (995 UPGMA, 993 *Kitsch*). Farsi, instead, is the outlier of the whole tree in both topologies, a position corroborated by high values of bootstrapping, especially in the tree produced by *Kitsch* (900; 738 UPGMA). Thus, no Indo-Iranian unity is recognized when a strict genealogical model is assumed. Their relative proximity, however, shows up moderately in the networks, where the three varieties are grouped in one separate section. Here, Farsi is on a branch longer than any other of the network, thus suggesting a higher number of mutations, in principle due either to isolation, which is

historically implausible, or to mixing with external varieties. More forcefully, the relation between Farsi and Indic is suggested by the particular additional experiment attempted in Section 4.1. below.

However, it is most remarkable that the failure of the trees in perceiving Indo-Iranian as a genealogical unit is exactly shared with Dyen et al.'s (1992) lexicostatistical experiment. In their results, while the highest percentage of cognates exhibited by Iranian is indeed with the Indic cluster (18.1), Indic exhibits a higher percentage of cognates with the so-called 'Mesoeuropeic' set, i.e. Central European IE languages (18.6): as a consequence, Indo-Iranian fails their subgrouping method, which would require a cognacy percentage of 21.1 (= 18.6 + 2.5).[7] Second, the percentage of cognates between an Indo-Aryan and an Iranian language ranges from 9.4 (Singhalese–Afghan) to 25.4 (Panjabi–Baluchi), indeed suggesting some relevance of areal closeness/distance.[8] For Dyen et al., a range of 16 percentage points of internal variability is too high to characterize a well-defined group. They justify the discrepancy between their lexicostatistical classification and the traditional Indo-Iranian hypothesis with the self-imposed restriction to contemporary word lists: they suppose that the method, if applied to older stages of Indic (i.e. Sanskrit) and Iranian (i.e. Avestan, Old Persian) evidence, might perhaps identify an Indo-Iranian group. Meillet's (1922: 25–30) and others' conclusion that Indo-Aryan and Iranian continue a specific proto-language is based precisely on evidence from Old Persian, Avestan, and Sanskrit. Thus, the failure of the lexicostatistical (as well as the syntactic) experiments in confirming Indo-Iranian may result from the loss of evidence in the modern languages: "this erosion is conceivably great enough that contemporary languages fail clearly to reveal the group, whether the data used are lexicostatistical or those of common innovations" (Dyen et al. 1992: 48).

Speaking of secondary interference effects, the percentage of cognates has been considerably deflated by the massive borrowing into Iranian of lexical items coming from languages belonging to non-native populations that, since the first millennium AD, have dominated in the area, e.g. Arabic.

In syntactic terms, Farsi exhibits at least four cases of parameter values which are not shared with any other Indo-European variety of our sample, i.e. p3 (FGG), p19 (CCN), p46 (NOO) and p55 (AMO), plus two further parameters (p16, CPS, and p22, FFS), whose value (–), within Indo-European, is shared with Welsh (p16, CPS) and English (p22, FFS) only. Of these parameters, the values of p3 (FGG) and p16 (CPS) are shared by Uralic and Altaic languages: this could be taken as a marker of Central Asian influence. Finally, p46 (NOO) affects the position of nouns with respect to adjectives, and its value is partly shared with Semitic languages, as the possible correspondent to Arabic loanwords in the lexicon. However, only an analysis of more Iranian and Indic languages could sway syntactic evidence either in favor or against skepticism about Indo-Iranian.

3.5 Syntax and lexicon

The parallelism of syntactic and lexical analyses in a critical case (the Indo-Iranian question) suggests the opportunity of pursuing such comparison more generally, refining an enterprise inaugurated in Longobardi & Guardiano (2009).

With Dyen et al.'s (1992) database, distances can be computed as the complement of the percentage of cognates shared between any two languages in a pair. By way of illustration, the percentage of cognate forms between Italian and French amounts to 80.3%; thus, if measured on a scale ranging from 0 to 100, their distance turns out to be 19.7 (0.197 on a scale ranging from 0 to 1). As mentioned, 21 languages of our sample overlap with those of Dyen et al. (which does not include the five Greek and Romance dialectal varieties of Southern Italy), allowing for comparisons.

First, the two distance matrices (Figure 7) reveal that syntactic distances are significantly smaller than lexical ones. For most pairs, the ratio between the two, attended around 3:4, is considerably smaller, varying between 1:3 and 1:4 (mean = 1:3.6).[9] Such a difference hints at the formal demonstration that the composition of the basic vocabulary has changed more rapidly than (at least nominal) syntax in the history of these 21 Indo-European languages.

The distribution of syntactic distances in the matrix can be assimilated to a Normal: this is shown in the histogram in Figure 8. On the contrary, lexical distances (Figure 9) display a Bimodal distribution, with a first peak for the interval 0.3–0.4 and a second at 0.8–0.9. In other words, it seems that lexical distances are able to clearly detect taxonomic signals up to 0.5; on higher values, they tend to saturate, i.e. to become uninformative because they are no longer able to scatter on different degrees of relationship. Notice that all syntactic distances distribute

SYNTACTIC DISTANCES

	It	Sp	Fr	Ptg	Rm	Grk	E	D	Da	Ice	Nor	Blg	SC	Slo	Po	Rus	Ir	Wel	Far	Ma	Hi
It	0	0,0682	0,0476	0,0233	0,103	0,225	0,15	0,125	0,128	0,2	0,128	0,15	0,233	0,233	0,194	0,226	0,211	0,243	0,433	0,2	0,2
Sp	0,212	0	0,0714	0,0233	0,103	0,25	0,175	0,15	0,154	0,2	0,154	0,2	0,267	0,267	0,226	0,258	0,184	0,216	0,467	0,233	0,233
Fr	0,197	0,266	0	0,0488	0,162	0,289	0,184	0,158	0,162	0,184	0,162	0,184	0,276	0,276	0,233	0,267	0,194	0,229	0,483	0,241	0,241
Ptg	0,227	0,126	0,291	0	0,105	0,256	0,15	0,125	0,128	0,179	0,128	0,154	0,233	0,233	0,194	0,226	0,189	0,222	0,433	0,2	0,2
Rm	0,34	0,406	0,421	0,371	0	0,2	0,211	0,25	0,162	0,205	0,132	0,205	0,241	0,241	0,233	0,233	0,257	0,286	0,5	0,214	0,214
Grk	0,822	0,833	0,843	0,833	0,843	0	0,3	0,256	0,263	0,293	0,282	0,25	0,188	0,188	0,182	0,182	0,257	0,286	0,393	0,267	0,267
E	0,753	0,76	0,764	0,76	0,773	0,838	0	0,0732	0,0714	0,122	0,0714	0,19	0,219	0,219	0,242	0,242	0,111	0,139	0,3	0,212	0,212
D	0,735	0,747	0,756	0,753	0,751	0,812	0,422	0	0,093	0,0909	0,093	0,167	0,147	0,147	0,147	0,176	0,111	0,118	0,31	0,242	0,242
Da	0,737	0,75	0,759	0,75	0,763	0,817	0,407	0,293	0	0,0889	0,0435	0,14	0,156	0,156	0,188	0,188	0,167	0,176	0,345	0,176	0,176
Ice	0,755	0,763	0,772	0,763	0,777	0,802	0,454	0,409	0,221	0	0,0435	0,133	0,118	0,118	0,147	0,147	0,132	0,139	0,3	0,176	0,176
Nor	0,754	0,761	0,77	0,761	0,786	0,821	0,452	0,367	0,146	0,191	0	0,136	0,156	0,156	0,188	0,188	0,167	0,176	0,345	0,176	0,176
Blg	0,769	0,782	0,791	0,781	0,798	0,811	0,772	0,769	0,76	0,775	0,773	0	0,0882	0,0882	0,147	0,118	0,222	0,25	0,333	0,147	0,147
SC	0,755	0,768	0,772	0,766	0,778	0,821	0,766	0,764	0,749	0,768	0,772	0,291	0	0	0,0571	0,0286	0,222	0,259	0,321	0,188	0,188
Slo	0,76	0,772	0,782	0,781	0,79	0,821	0,751	0,733	0,733	0,763	0,762	0,385	0,316	0	0,0571	0,0286	0,222	0,259	0,321	0,188	0,188
Po	0,764	0,772	0,781	0,776	0,784	0,837	0,761	0,754	0,749	0,758	0,762	0,369	0,32	0,367	0	0,0278	0,214	0,25	0,345	0,219	0,219
Rus	0,761	0,769	0,778	0,773	0,781	0,832	0,758	0,755	0,74	0,754	0,758	0,365	0,325	0,386	0,266	0	0,214	0,25	0,345	0,219	0,219
Ir	0,8	0,805	0,812	0,817	0,837	0,859	0,817	0,806	0,817	0,755	0,836	0,818	0,796	0,809	0,8	0,782	0	0,0263	0,31	0,31	0,31
Wel	0,793	0,813	0,81	0,804	0,812	0,867	0,841	0,82	0,825	0,82	0,849	0,838	0,821	0,838	0,822	0,818	0,645	0	0,276	0,321	0,321
Far	0,859	0,86	0,859	0,854	0,864	0,875	0,86	0,86	0,86	0,862	0,874	0,853	0,842	0,849	0,849	0,844	0,888	0,899	0	0,345	0,345
Ma	0,833	0,839	0,839	0,834	0,838	0,869	0,855	0,849	0,855	0,856	0,859	0,811	0,822	0,817	0,827	0,817	0,879	0,893	0,834	0	0
Hi	0,818	0,819	0,824	0,813	0,827	0,874	0,854	0,853	0,843	0,855	0,852	0,801	0,805	0,8	0,799	0,8	0,878	0,876	0,815	0,436	0

LEXICAL DISTANCES

Figure 7. Matrix of syntactic and lexical distances (21 IE languages)

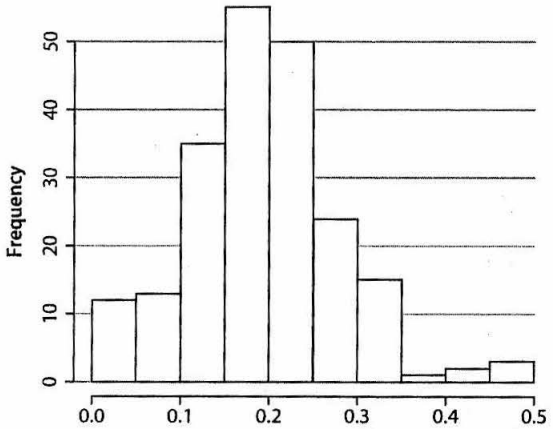


Figure 8. Histogram representing the distribution of the syntactic distances in Figure 7

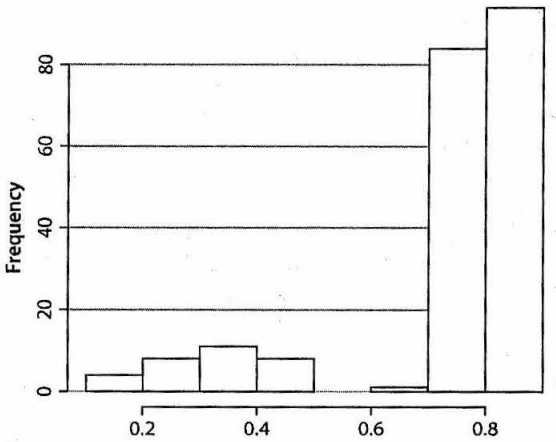


Figure 9. Histogram representing the distribution of the lexical distances in Figure 7

within 0.5 (perhaps in part also as a reflex of the Anti-Babelic principle, cf. fn. 9). A correlation test between the two distance matrices (Mantel test) shows that lexical and syntactic distances are highly correlated ($r = 0.7285$; $p < 0.001$; 9,999 permutations), thus proving wrong the hypothesis of orthogonality between lexicon and syntax. However, this does not provide any evidence that their relationship is linear, i.e. that they return fully overlapping information. The proof against the overlapping hypothesis is well visible in Figure 10, where lexical distances are plotted against syntactic ones. The graph neatly identifies two clouds: the bottom-left one shows an essentially proportional increase of the two distances, thus a linear covariation; the top cloud has a more compact distribution: lexical distances show

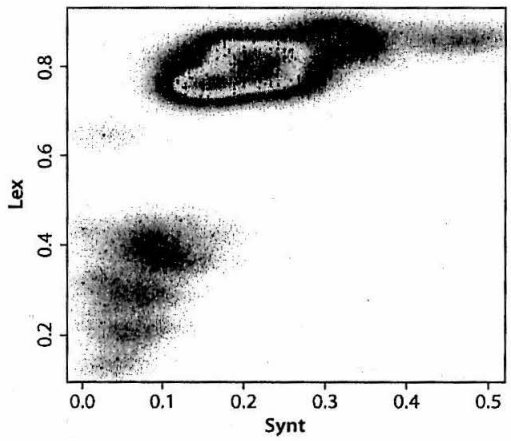


Figure 10. ScatterPlot of the syntactic and lexical distances in Figure 7

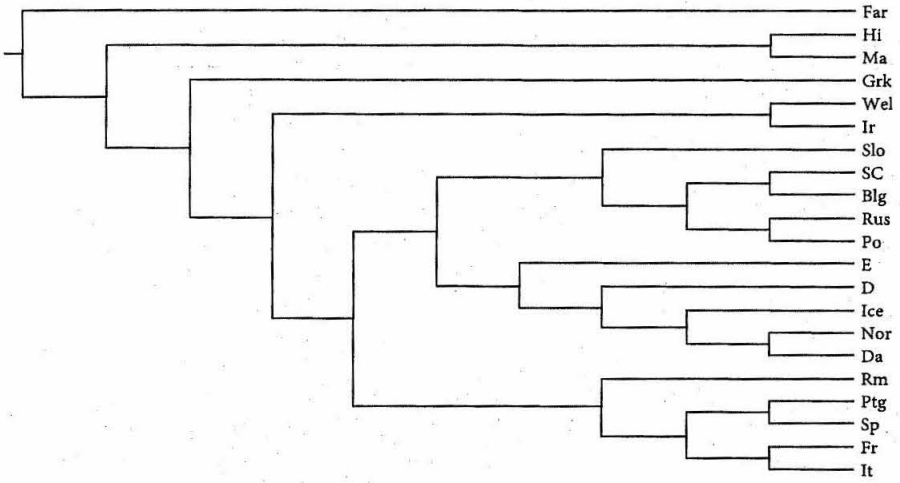


Figure 11. *Kitsch* tree from the lexical distances in Figure 7

little variation, as expected the more they approximate to 1, while syntactic distances remain more scattered. The language pairs corresponding to the dots on the first cloud are all and only those where both languages belong to the same subfamily. This suggests that lexical distances minimize the internal relations within the same subgroup, while maximizing those across distinct subgroups.

In any case, the tree topologies in Figures 11–12, drawn from the lexical and the syntactic distance matrices respectively, largely overlap, with a few low-level differences.[10] Within Romance, the position of French is variable: the lexical tree suggests some unity with Italian, while the syntactic tree separates them. Within

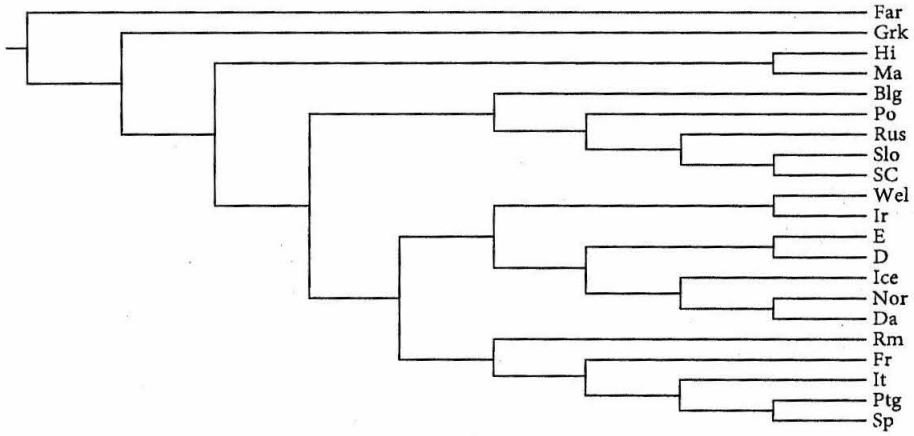


Figure 12. *Kitsch* tree from the syntactic distances in Figure 7

Germanic, both trees recognize the Northern group (with Icelandic as the outlier), though the syntactic tree identifies the Western unit as well, not detected in the lexical one. The superordinate node that groups Germanic and Celtic together only appears in the syntactic tree: the lexical one separates the two groups. Within Slavic, the lexical tree captures some relation between Bulgarian and Serbo-Croatian, but, again, does not properly recognize Southern Slavic unity, as Slovenian shows up as the outlier of the whole group. The position of Russian and Polish, instead, varies slightly in the two trees.

4. Further testing

4.1 Multiple ancestors: A character-based experiment

In the past years, biostatisticians have developed clustering methods that are neither based on phylogenetic models nor assume tree-like representations. These methods are of great interest for molecular anthropologists, because they allow taxonomic classifications not dependent on predefined evolutionary paths. The major difference with tree-like representations is that taxonomic units are treated as sets of elements coming from different ancestors, whose impact can be evaluated separately, rather than as single-fathered entities. One such algorithm, *Structure* (Pritchard et al. 2000), has recently been argued to be appropriate for the purposes of language classification (Reesink et al. 2009). It implements a Bayesian clustering technique that assumes a model with a number (call it K) of unspecified or unknown ancestral populations, each characterized by a salient set of properties

(e.g. allele frequencies at each DNA locus). As a result, individuals, on the basis of their genotypic profile, are assigned either to one, or jointly to two or more, of the K populations; in the latter case, they are considered as the result of admixture between different ancestral populations.

Structure is character-based. As such, it assumes characters to be unlinked to each other, and thus cannot simply use the parametric matrix as an input, because of the noted pervasive interdependencies across parameters. Thus, to explore how syntactic data behave in such a non-phylogenetic model, we manipulated the parameter grid, trying to minimize the impact of the implications and treat parameters as ideally independent: the subset of parameters exhibiting less than ten blanks ('0' or '?') across all languages was singled out.[11] In this way, the amount of information suitable for the experiment was inevitably reduced: only 37 out of the 56 parameters could indeed be selected. Notice that the reduction of the comparanda obviously makes taxonomic hypotheses less stable, increasing the probability of mistakes, as already experienced in studies on DNA polymorphisms.[12] Furthermore, our selection was not based on qualitative criteria of genealogical stability, adding to the risk of accidentally overestimating or underestimating differences and identities. In fact, any attempt to adapt the parameter probe to character-based algorithms involves violating some constraints, and this could undermine the reliability of the results. Ideally, one should be able to devise an appropriate character-based algorithm able to deal with the complex implicational structure of the whole set of parameters, i.e. of truly treating zeros as implied information rather than as missing values. No such situation seems to have arisen with molecular-genetic databases, so that no biostatistical program has so far been designed for these purposes.

To test the basic reliability of the subset, we first built phylogenetic trees from the corresponding distances (Figures 13–14) and checked them against the ones based on all the parameters. Beyond less relevant readjustments in the subarticulation of Romance and Slavic, there are no differences with the trees generated from the full dataset. The trees were generated without bootstrapping, owing to the inevitably small number of replaceable characters in the reduced dataset.

We ran *Structure* for K values from 2 to 8, K = 5 being the one with the lowest average log-likelihood value (hence the 'best', Figure 15). The results are represented in the form of barplots, each bar corresponding to a language, and colors representing the contribution of each of the supposed K ancestral 'populations'. The experiments with K = 5 were replicated ten times, and all the replications produced the same pattern. Such a result independently attests the statistical solidity of the subset of parameters selected.

Empirically, the clusters in Figure 15 match the taxonomic aggregations of the distance-based phylogenetic trees in several respects. Notice that five optimal

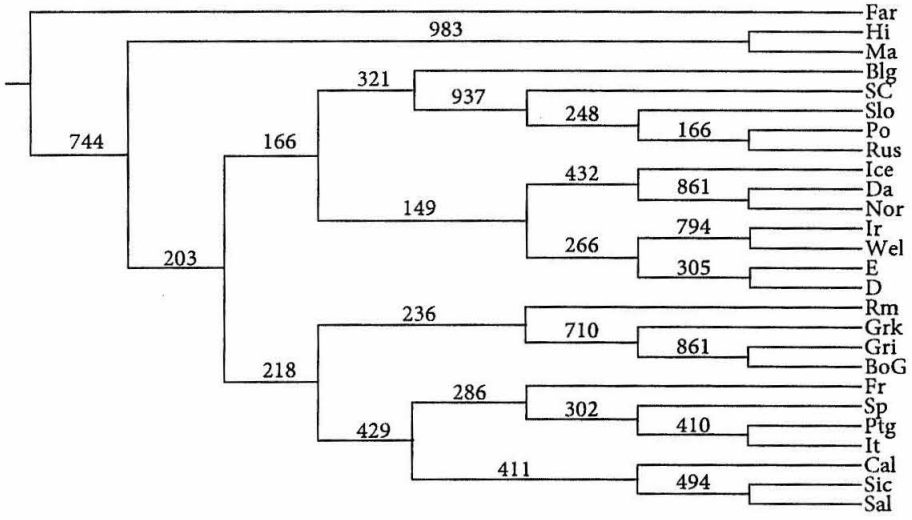


Figure 13. UPGMA tree from the reduced parameter grid (37 parameters)

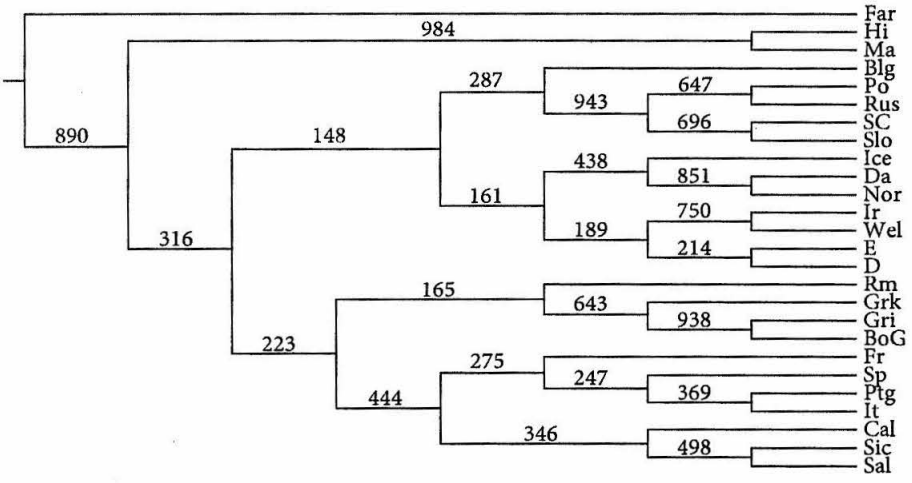


Figure 14. *Kitsch* tree from the reduced parameter grid (37 parameters)

ancestors are actually one fewer than the traditionally well recognized families for the languages of our sample (Romance, Slavic, Greek, Celtic, Germanic, and Indo-Iranian), inevitably anticipating one forced clustering (see below).

Two languages (Rumanian and Bulgarian) immediately appear as likely to have very mixed ancestry. In the others, one component is always saliently prevailing. The red component dominates in the Romance languages, confirming their recent

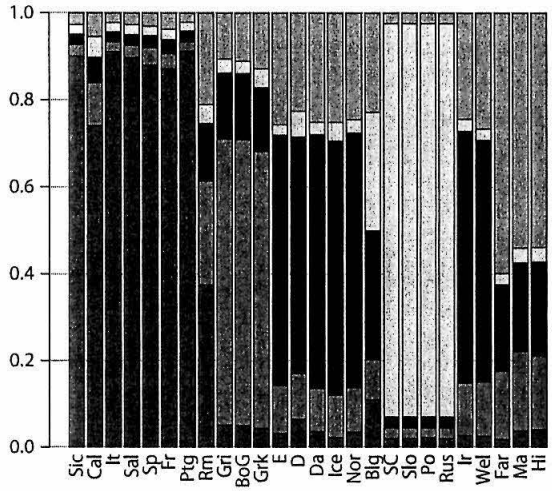


Figure 15. *Structure* graph (K = 5) from the reduced parameter grid (37 parameters)

common ancestry; the sole exception is indeed Rumanian, which displays significant manifestations of the green, orange, and blue components. Thus, Rumanian is weakly assigned to the varieties characterized by the red component, but is not distinctly associated to any of the other potential ancestors; this is coherent with the information provided in the networks obtained from the complete dataset.

The three Greek varieties are all overwhelmingly marked by the green component. The red (Romance) component does not show up significantly in the two varieties of Southern Italy, in any case no more there than in Modern Greek; this confirms that contact with Romance in Southern Italy has not obscured their genealogical origin.

The Slavic varieties, with the exception of Bulgarian, show an almost complete absence of contributions from other components. In Bulgarian, besides the common Slavic component (i.e. the yellow one), others are detectable (though, as in Rumanian, not necessarily those prevailing in its closest neighbors), so that here Bulgarian is hardly identified as a member of the Slavic family. However, at least the proportion of the other components in its structure does not allow for more salient alternative hypotheses of grouping.

The expected 'forced' aggregation concerns Germanic and Celtic, in which the same blue component dominates, so that they cannot be plausibly attributed to separate intermediate ancestors. Actually, the closeness of Germanic and Celtic languages was overstated in the networks as well.

The Indic varieties are characterized by the same orange component that appears conspicuously in the bar of Farsi (and rather less saliently also in some other bars). This result is more in line with the topology of the networks. Both

experiments suggest some Indo-Iranian affinity, not equally recognized in any of the trees, i.e. when a stricter genealogical model is assumed. One should conclude that, within the sample, the strongest ancestral component is shared by Farsi with the two Indic languages, although this does not necessarily speak for a genealogical relationship, at least from the evidence of modern languages. In any case, the graph of *Structure* confirms that it was not interference from other Indo-European components of our sample that has syntactically detached Farsi from Indo-Aryan. This matches the particular length of the Farsi branch exhibited by the networks.

To summarize, the outcomes of this first implementation of *Structure* on parametric data do not contradict any of the hypotheses which emerged from the corresponding phylogenetic experiments, but also in some cases (the composite structure of Bulgarian and Rumanian) directly suggest plausible effects of admixture.

4.2 Preliminary character-based phylogenies

Once a parameter subset, relatively informative in spite of its limits (low number of variable sites, high amount of missing data, non-independence of many characters), was selected for use with *Structure*, it became tempting to test it with a character-based algorithm of a strictly phylogenetic type. Here, the present lack of directionality and weighting information is an even more serious bias. Indeed we chose a parsimony (the 'maximum parsimony method' was first discussed in Edwards & Cavalli-Sforza 1963, cf. also Felsenstein 2004) program, PAUP* (Swofford 2003), that does not require any explicit model of character evolution. Parsimony algorithms search for best trees assuming that the lowest amount of change occurred between the nodes and a common ancestor; their output is a series of trees tied at their best parsimony score (i.e. which share the minimum number of changes).[13] With lexical data, parsimony-based methods have been claimed to yield the best performances (Barbançon et al. to appear).

As the trees produced are unrooted, a certainly remote outlier language, Wolof (from the West Atlantic family of Niger-Congo), was included in order to root them. After an analysis through a branch-and-bound search, the program was able to produce 59 different trees, which require a total of 50 changes.

Then, a consensus tree (Figure 16) was built through the 'extended majority rule'.[14] Despite the noted limitations, it is consistent with *Structure*. Farsi is, once more, the outlier of Indo-European. Most of the genealogical subgroupings are correctly recognized, and display high consensus values: the two Celtic languages are always grouped together, like the two Indic ones and the compact group consisting of all Slavic languages but Bulgarian; most of the trees (55 out of 59) acknowledge the Greek group and its internal articulation, as well as Romance unity (53 out of 59) including Rumanian. As for Bulgarian, the composite nature emerged from

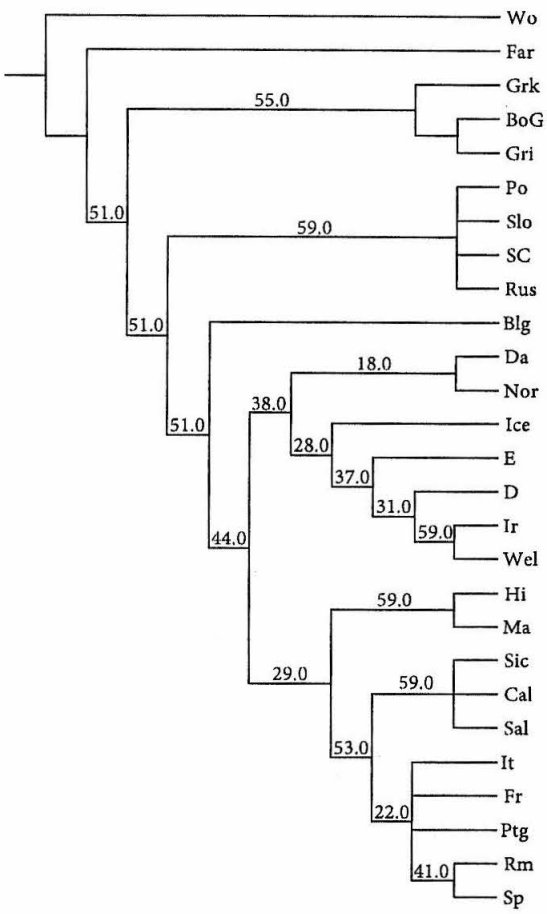


Figure 16. PAUP* consensus tree, from the reduced parameter grid (37 parameters).

Structure seems to affect its position in the tree: it is not included in the Slavic cluster, although it shows up close to it. Finally, the program expectedly fails to acknowledge the Germanic cluster; the overstated closeness between Celtic and West Germanic, which biases the reduced dataset also with *Structure*, emerges here as well; however, the cluster including Germanic and Celtic displays consensus values consistently lower than the others, thus resulting as the weakest of the whole tree.

4.3 Some remarks on language contact

In general, all algorithms and datasets agree in presenting obvious points of taxonomic failure only in a modest number of cases; rather interestingly, they are all

cases where geography and known history independently suggest the possibility of areal contact: the questionable positions of Farsi and of Bulgarian in the full dataset with 56 parameters and the overstated Celtic-West Germanic relation, in the reduced one.

Even if such plausible effects of horizontal transmission may help explain some of these perturbations of genealogical expectations, on the other side they are in any case limited enough and relatively controllable. First, programs such as *Structure* turned out to be able to warn us in any case about the possible mixed syntactic character of some Balkan languages, such as Rumanian and Bulgarian, yet interference is not so strong as to pull them away from Romance and Slavic, respectively (except for Bulgarian in a subset of the experiments with the reduced dataset).

Second, in other cases of obvious contact, like that of unbalanced interference affecting Southern Italian Greek, the expected genealogy is not disrupted by such effects under any program used.[15]

Third, secondary convergence may have contributed to the attraction between Celtic and West Germanic so visible in experiments with the reduced dataset, but is far from having caused it. Actually, one of the advantages of the parametric approach is that it allows one to define precisely a notion of **potentially borrowed characters** as follows: potentially borrowed characters are those parameter values shared by a language supposedly targeted by interference (e.g. Grico) with its neighbors (e.g. Italo-Romance), but crucially not with its closest genealogical relatives (e.g. homeland Greek). This provides a rough tool for measuring the plausibility of contact-based explanations. In this sense, the maximal set of characters potentially borrowed by Grico from any one of the three Romance dialects of Southern Italy in our sample is four out of 38 comparable parameter values (i.e. 10.5%); it reduces to 2/37 in Bovese (5.4%). Recall that all methods have been still able to connect Grico, Bovese, and homeland Greek to a common ancestry.

In the same vein, the perturbation of vertical expectations for Rumanian and Bulgarian can be correctly foreseen to be only marginally detectable, though possibly overstated by the combined contact sources of the area: Bulgarian may have maximally borrowed from Greek two parameter values (p7, DGP and p8, DGR) out of 40 (5%), and four out of 39 (less than 10.3%) from Rumanian (p7, DGP, p8, DGR, p12, DCN, and p32, GFO). This is expectedly not enough to remove it steadily from Slavic (only in PAUP* does it occur slightly out of its family, even if it is not attracted by any language in particular), though it may reinforce its surprising peripheral position in the family, because p7 (DGP) and p8 (DGR) are very consequential in terms of the neutralizing force their '–' value exerts on other parameters.[16] Rumanian, in turn, may have maximally borrowed three parameter values out of 40 (7.5%) from Greek (p29, GFR, p35, APO, and p43, TSL) and, indeed, shares one relevant (p12, DCN) out of 39 with Bulgarian (about 2.5%). With

such values, it is not surprising that only the reduced dataset displays questionable taxonomic choices.

As noted above, no convergence of equal size and solidity can be identified in our parametric database for Celtic and West Germanic. This fact, compared to the Southern Italian and Balkan situation, suggests that some small amount of homoplasy rather than simple interference is likely to bias the groupings of Germanic and Celtic within the reduced dataset, which, precisely because it is smaller in size, is more exposed to such chance factors.

Pending further research, it appears that neither horizontal transmission of syntax nor homoplasy, for that matter, are such factors to severely impair the effectiveness of PCM, when applied to a robust number of parameters.

5. Summary

The application of various quantitative tools to parametric data already provides support to traditionally well-substantiated genealogical groupings, and confirms instability for taxonomies that previous comparative methods have equally challenged. It must be remarked that, presently, no available phylogenetic method seems capable of handling the implicational structure that characterizes parametric data; the application of those already existing inevitably implies some loss of important information. Therefore, in order for parameter data to become even more informative for phylogenetic purposes, it would be necessary to pursue the implementation of dedicated algorithms and methods.

In particular, the use of character-based algorithms is for now weakened by the absence of reliable models of markedness and parameter change/evolution. However, it is important that parametric comparison already yields acceptable genealogical outcomes in advance of, and unbiased by, any such hypotheses and of the inevitable idealizations and manipulations associated with them.

On the whole, only two points in our taxonomic results obtained from 56 parameters hint at some divergence from genealogies of traditional scholarship, which have been crucially assisted, unlike ours, by evidence from ancient varieties: the outlier position of Bulgarian within Slavic and that of Farsi. In the first case, simply importing grammaticalized definiteness from Greek and/or Rumanian may have overstated the position. In the second case, the intervention of minor areal interference from outside Indo-European is a plausible hypothesis (somewhat detectable through the further experiments); in addition, the problem partly finds parallelism in the outcome of vocabulary comparisons. Further local issues, such as the excessive interconnection of Celtic and West Germanic, seem to only occur as a byproduct of the reduced dataset.

In our view, once the rather moderate effects of horizontal transmission are properly factored out, the evidence of the refined experiments presented here appears to substantially back Longobardi & Guardiano's (2009) claim that parametric syntax encodes a genealogical signal of strength comparable to the lexical one.

Notes

* For useful suggestions and discussions we are especially indebted to F. Bernasconi, R. Lazzeroni, N. McCoy, C. Melchert, I. Roberts, D. Willis and two anonymous referees. The elaboration of this article has benefited from the ERC Advanced Grant n. 295733 (2012–2017) LANGELIN, awarded to G. Longobardi. This article is dedicated to the memory of Calvert Watkins (1933–2013), who kindly discussed some preliminary ideas pursued here with two of us.

1. Two main versions of Swadesh's list of supposedly 'basic' (i.e. universal and culturally independent) meanings are available, with 100 and 200 items, respectively. For a detailed discussion of the impact of the list length in refining relatedness hypotheses, see Embleton (1986).

2. Bootstrapping values have turned out slightly lower than one might wish more generally (eight nodes exceed the threshold of 500 in UPGMA, seven in *Kitsch*). However this is certainly influenced by the relatively small number of relevant taxonomic characters. Furthermore, the bootstrapping procedure is not completely appropriate for treating parametric data, again because of the pervasive implicational structure, which is not taken into account in the perturbation of the matrix.

3. For the relations between Gallo-Romance and other subgroups, cf. at least Tagliavini (1972), Harris & Vincent (1988), Renzi & Salvi (1994), Lindenbauer et al. (1995), and Holtus et al. (1988–2005).

4. Divergence is one of the typical effects of the contact relations across the Balkan languages pointed out in Hock & Joseph (1996), among others.

5. Exceptions: Farsi, Hindi and Marathi are –p7 (DGP); Rumanian, Bovese Greek, Grico and Greek are –p29 (GFR).

6. For proposals regarding the properties associated with the absence of grammaticalized definiteness in the Slavic languages, cf. Bošković (2008a, 2008b, 2010). Even if certain interesting hypotheses of such works turned out to be correct (e.g., if no language without grammaticalized definite articles could have prepositional genitives), it would not be the two parameters p7 (DGP) and p29 (GFR) that would be implicationally related in the current formulation of Figure 1, so that this would not reduce the distance between Bulgarian and other Slavic languages.

7. In their **subgrouping method**, 2.5 is a safety coefficient called **critical difference** (c.f. Dyen et al. 1992: 26).

8. By itself, this kind of evidence does not help one choose between the hypothesis of secondary contact as disrupting a primeval genealogical relatedness or as overstating an original looser kinship of the two groups. The hypothesis that many isoglosses shared by ancient Indic and

Iranian literary varieties derive in fact from cultural sharing in contiguous areas was suggested by Lazzeroni (1968, 1998), after a detailed exploration of their exact chronology.

9. Some statistical procedures were implemented in Bortolussi et al. (2011) on a corpus of roughly 10,000 fictitious languages, randomly generated, though according to the constraints imposed by the implicational structure of the parameter list. The experiment revealed that the highest value attended for syntactic distances cannot exceed 0.75 rather than 1 (cf. Guardiano and Longobardi's 2005: 161–2, Anti-Babelic Principle for precisely an expectation of this kind). Distance 1 must instead be attended for the vocabulary, given that two languages with no recognizable common etymology at all do precisely attain this value.

10. Both trees have been computed from a distance-matrix using *Kitsch*. A character matrix of the lexical data is not available; therefore, bootstrapping procedures cannot be implemented here.

11. For details on the selection cf. the support material in the Appendix.

12. Colonna et al. (2009) show that genetic structures are clearly detectable even within individuals from the same geographic domain (Europe) by using 239 STRs (Short Tandem Repeat) loci, while comparisons become uninformative when the number of loci is reduced to 36. Based on three continental populations (Afro-American, Asian and European), Turakulov & Easteal (2003) conclude that for the correct assignment of an individual to the continent of origin with a mean accuracy of at least 90%, a minimum of 65–100 Single Nucleotide Polymorphisms (SNPs) are needed, and the same holds for other genetic markers as STRs.

13. An anonymous referee correctly notices that phylogenetic classification is traditionally guided by the principle of shared innovations. This principle cannot be implemented in a distance-based method but is automatically embodied in a parsimony system, which is designed precisely to minimize homoplasy.

14. Consensus trees built through an extended majority rule retains **all** best nodes, even if they score lower than 50%. On the contrary, a majority rule retain **only** branchings with a consensus score higher than 50%.

15. The Greek varieties are associated with 'minority' communities and, traditionally, have not had any systematic exposure to their homeland variety (cf. the sources reviewed in Guardiano & Stavrou to appear).

16. Cf. Section 3.4 above. For the impact of the value '–' of p7 (DGP) and p8 (DGR) on the overall implicational structure of Table A, cf. also the support material in the Appendix.

References

Alexiadou, Artemis, Liliane Haegeman & Melita Stavrou. 2007. *Noun Phrase in the Generative Perspective*. Berlin: Mouton de Gruyter.

Baker, Mark C. 1996. *The Polysynthesis Parameter*. Oxford: Oxford University Press.

Baker, Mark C. 2001. *The Atoms of Language*. New York: Basic Books.

Barbançon, François G., Steven N. Evans, Luay Nakhleh, Donald Ringe & Tandy Warnow. To appear. An Experimental Study Comparing Linguistic Phylogenetic Reconstruction Methods. *Diachronica* 30:2.

Biberauer, Theresa, ed. 2008. *The Limits of Syntactic Variation*. Amsterdam: John Benjamins.

Bortolussi, Luca, Giuseppe Longobardi, Cristina Guardiano & Andrea Sgarro. 2011. How Many Possible Languages Are There? *Biology, Computation and Liguistics* ed. by Gemma Bel-Enguix, Verónica Dahl & Maria Dolores Jiménez-López, 168–179. Amsterdam: IOS Press.

Bošković, Željko. 2008a. What Will You Have, DP or NP? *Proceedings of the North East Linguistic Society 37* ed. by Emily Elfner & Martin Walkow, 101–114. Amherst: University of Massachusetts Press.

Bošković, Željko. 2008b. The NP/DP Analysis and Slovenian. *Proceedings of the Novi Sad Generative Syntax Workshop* ed. by Ljiljana Subotić, 53–73. Novi Sad: Filozofski fakultet u Novom Sadu.

Bošković, Željko. 2010. *On NP and Clauses*. Ms. University of Connecticut.

Bromham, Lindel & David Penny. 2003. The Modern Molecular Clock. *Nature Reviews Genetics* 4.216–224.

Chomsky, Noam. 1975. *Reflections on Language*. New York: Pantheon.

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

Colonna, Vincenza, Teresa Nutile & Ronald R. Ferrucci. 2009. Comparing Population Structure as Inferred from Genealogical Versus Genetic Information. *European Journal of Human Genetics* 17.1635–1641.

van Cort, Tracy. 2001. *Computational Evolutionary Linguistics: Tree-based Models of Language Change*. Pomona, CA: Harvey Mudd College PhD Dissertation.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science* 309:5743.2072–2075.

Dyen, Isidore, Joseph B. Kruskal & Paul Black. 1992. An Indoeuropean Classification: A Lexicostatistical Experiment. *Transactions of the American Philosophical Society* 82:5.1–132.

Edwards, Anthony W. F., & Luigi Luca Cavalli-Sforza. 1963. The Reconstruction of Evolution. *Annals of Human Genetics* 27.105–106.

Embleton, Sheila M. 1986. *Statistics in Historical Linguistics*. Bochum: Brockmeyer.

Felsenstein, Joseph. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Ghomeshi, Jila, Ileana Paul & Martina Wiltschko, eds. 2009. *Determiners: Universals and Variation*. Amsterdam: John Benjamins.

Gimbutas, Marija. 1973. The Beginning of the Bronze Age in Europe and the Indo-Europeans 3500/2500 B.C. *Journal of Indo-European Studies* 1.163–214.

Gray, Russel D. & Quentin D. Atkinson. 2003. Language Tree Divergences Support the Anatolian Theory of Indo-European Origin. *Nature* 426.435–439.

Guardiano, Cristina & Giuseppe Longobardi. 2005. Parametric Comparison and Language Taxonomy. *Grammaticalization and Parametric Variation* ed. by Montserrat Batllori, Maria-Lluïsa Hernanz, Carme Picallo & Francesc Roca, 149–174. Oxford: Oxford University Press.

Guardiano, Cristina & Melita Stavrou. To appear. Greek and Romance in Southern Italy: History and Contact in Nominal Structures. *L'Italia Dialettale*.

Harris, Martin & Nigel Vincent, eds. 1988. *The Romance Languages*. London: Routledge.

Heggarty, Paul. 2004. Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully be Applied to Language Data — And to Dating Language? *Phylogenetic Methods and the Prehistory of Languages* ed. by James Clackson, Peter Forster & Colin Renfrew, 183–194. Cambridge: McDonalds Institute for Archaeological Research.

Hickey, Raymond, ed. 2010. *The Handbook of Language Contact*. Oxford: Wiley-Blackwell.

Hock, Hans H. & Brian D. Joseph. 1996. *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics*. Berlin: Mouton De Gruyter.

Holtus, Günter, Michael Metzeltin & Christian Schmitt, eds. 1988–2005. *Lexikon der Romanistischen Linguistik (LRL)*. Tübingen: Niemeyer.

Huson, Daniel H. & David Bryant. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution* 23:2.254–267.

Joseph, Brian D. 1999. Romanian and the Balkans: Some Comparative Perspectives. *The Emergence of the Modern Language Sciences: Studies on the Transition from Historical-Comparative to Structural Linguistics in Honour of E.F.K. Koerner*, Vol. 2 ed. by Sheila Embleton, John E. Joseph & Hans-Josef Niederehe, 218–235. Amsterdam: John Benjamins.

Kayne, Richard. 2000. *Parameters and Universals*. New York: Oxford University Press.

Keenan, Edward L. & Denis Paperno, eds. 2012. *Handbook of Quantifiers in Natural Language*. (= *Studies in Linguistics and Philosophy*, 90.) Dordrecht: Springer.

Lazzeroni, Romano. 1968. Per una definizione dell'unita indo-iranica. *Studi e Saggi Linguistici*. Supplement to *L'Italia Dialettale* 8.131–159.

Lazzeroni, Romano. 1998. Sanskrit. *The Indo-European Languages* ed. by Anna Giacalone Ramat & Paolo Ramat. London: Routledge.

Lightfoot, David. 1979. *Principles of Diachronic Syntax*. Cambridge: Cambridge University Press.

Lightfoot, David. 1991. *How to Set Parameters*. Oxford: Blackwell.

Lindenbauer, Petrea, Michael Metzeltin & Margit Thir. 1995. *Die romanischen Sprachen: Eine einführende Übersicht*. Wilhelmsfeld: Egert.

Lohr, Marisa. 1998. *Methods for the Genetic Classification of Languages*. Cambridge: Cambridge University PhD dissertation.

Longobardi, Giuseppe. 1994. Reference and Proper Names. *Linguistic Inquiry* 25:4.609–665.

Longobardi, Giuseppe. 2001. How Comparative is Semantics? A Unified Parametric Theory of Bare Nouns and Proper Names. *Natural Language Semantics* 9.335–369.

Longobardi, Giuseppe. 2003. Methods in Parametric Linguistics and Cognitive History. *Linguistic Variation Yearbook* 3.101–138.

Longobardi, Giuseppe. 2005. A Minimalist Program for Parametric Linguistics? *Organizing Grammar: Linguistic Studies for Henk van Riemsdijk* ed. by Hans Broekhuis, Norbert Corver, Riny Huybregts, Ursula Kleinhenz & Jan Koster, 407–414. Berlin: Mouton de Gruyter.

Longobardi, Giuseppe. 2008. Reference to Individuals, Person, and the Variety of Mapping Parameters. *Essays on Nominal Determination* ed. by Alex Klinge & Henrik Høeg Müller, 189–211. Philadelphia: John Benjamins.

Longobardi, Giuseppe. 2012. Convergence in Parametric Phylogenies: Homoplasy or Principled Explanation? *Parameter Theory and Language Change* ed. by Charlotte Galves, Sonia Cyrino, Ruth Lopes, Filomena Sandalo & Juanito Avelar, 304–319. Oxford: Oxford University Press.

Longobardi, Giuseppe & Cristina Guardiano. 2009. Evidence for Syntax as a Signal of Historical Relatedness. *Lingua* 119:11.1679–1706.

McMahon, April. 2010. Computational Models and Language Contact. In Raymond Hickey, ed., 31–47. Oxford: Wiley-Blackwell.

McMahon, April, Paul Heggarty, Robert McMahon & Natalia Slaska, 2005. Swadesh Sublists and the Benefits of Borrowing: An Andean Case Study. *Quantitative Methods in Language Comparison* (= *Transactions of the Philological Society*, 103:2) ed. by April McMahon, 147–169. Oxford: Blackwell.

McMahon, April & Robert McMahon. 2003. Finding Families: Quantitative Methods in Language Classifying. *Transaction of the Philological Society* 101:1.7–55.

McMahon, April & Robert McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.

Meillet, Antoine. 1905. Avertissement. *Abrégé de grammaire comparée des langues indo-européennes d'après le Précis de grammaire comparée de K. Brugmann et B. Delbrück* ed. by Antoine Meillet & Robert Gauthiot, i–v. Paris: Klincksieck.

Meillet, Antoine. 1922. *Les dialectes indo-européens*. Paris: Champion.

Meillet, Antoine. 1924. *Le slave commun*. Paris: Champion.

Nakhleh, Luay, Tandy Warnow, Don Ringe & Steven N. Evans. 2005. A Comparison of Phylogenetic Reconstruction Methods on an IE Dataset. *Transactions of the Philological Society* 3:2.171–192.

Nelson-Sathi, Shijula, Johann Mattis List, Hans Geisler, Heiner Frangerau, Russel D. Gray, William Martin & Tal Dagan. 2010. Networks Uncover Hidden Lexical Borrowing in Indo-European Language Evolution. *Proceedings of the Royal Society, Biological sciences* 1713.1794–1803.

Nichols, Johanna. 1992. *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Noonan, Michael. 2010. Genetic Classification and Language Contact. In Raymond Hickey, ed., 48–65.

Pellegrini, Giovan Battista. 1977. *Carta dei dialetti d'Italia*. Pisa: Pacini.

Plank, Frans, ed. 2003. *Noun Phrase Structure in the Languages of Europe*. Berlin: Mouton de Gruyter.

Pritchard, Jonathan K., Matthew Stephens & Peter Donnelly. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155.945–959.

Reesink, Ger, Ruth Singer & Michael Dunn. 2009. Explaining the Linguistic Diversity of Sahul Using Population Models. *PLoS Biology* 7(11): e1000241.

Renfrew, Colin. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins*. London: Jonathan Cape.

Renzi, Lorenzo & Gianpaolo Salvi. 1994. *Nuova Introduzione alla filologia romanza*. Bologna: Il Mulino.

Ringe, Don, Tandy Warnow & Ann Taylor. 2002. Indo-European and Computational Cladistics. *Transactions of the Philological Society* 100:1.59–129.

Rigon, Gabriele. 2009. *A Quantitative Approach to the Study of Syntactic Evolution*. Pisa: Università di Pisa PhD dissertation.

Rigon, Gabriele. 2012. An Evolutionary Perspective on Diachronic Syntax. *Studi e Saggi Linguistici* 50:2.31–95.

Roberts, Ian. 1998. Review of *Historical Syntax in Cross-Linguistic Perspective* (= *Cambridge Studies in Linguistics*, 74) by Alice Harris & Lyle Campbell (Cambridge: Cambridge University Press, 1995). *Romance Philology* 51.363–370.

Roberts, Ian. To appear. The Mystery of the Overlooked Discipline: Modern Syntactic Theory and Cognitive Science. Available at http://ling.auf.net/lingbuzz/001611.

Schleicher, August. 1861–1862. *Compendium der vergleichenden Grammatik der indo-germanischen Sprachen: Kurzer Abriß einer Laut- und Formenlehre der indo-germanischen Ursprache*. Weimar: Böhlau.

Schmidt, Johannes. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar: Böhlau.

Sokal, Robert R. & Charles D. Michener. 1958. A Statistical Method for Evaluating Systematic Relationships. *University of Kansas Science Bulletin* 38.1409–1438.

Swadesh, Morris. 1950. Salish Internal Relationships. *International Journal of American Linguistics* 16.157–167.

Swofford, David L. 2003. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sunderland, MA: Sinauer Associates.

Szemerényi, Oswald. 1957. The Problem of Balto-Slav Unity. *Kratylos* 2.97–123.

Tagliavini, Carlo. 1972. *Le origini delle lingue neolatine. Introduzione alla filologia romanza*. Bologna: Pàtron.

Thomason, Sarah G. 2000. Linguistic Areas and Language History. *Languages in Contact* ed. by Dicky Gilbers, John Nerbonne & Jos Schaeken, 311–327. Amsterdam: Rodopi.

Thomason, Sarah G. 2001. *Language Contact: An Introduction*. Washington, DC: Georgetown University Press.

Thomason, Sarah G. 2004. *Rule Borrowing*. Ms., University of Michigan.

Thomason, Sarah G. 2010. Contact Explanations in Linguistics. In Raymond Hickey, ed., 31–47.

Thomason, Sarah G. & Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Tomić, Olga M. 2003. The Balkan Sprachbund Properties. *Topics in Balkan Syntax and Semantics* ed. by Olga M. Tomić, 1–58. Amsterdam: John Benjamins.

Turakulov, Rust & Simon Easteal. 2003. Number of SNPs Loci Needed to Detect Population Structure. *Human Heredity* 55.37–45.

Willis, David. 2008. Celtic and English Language Contact and Shift. *2nd Historical Sociolinguistics Network Summer School*. University of Bristol, 7–13 August 2008.

Appendix

Please see the appendix online at http://dx.doi.org/10.1075/jhl.3.1.07lon.additional

Corresponding authors' addresses

Giuseppe Longobardi
Department of Language and Linguistic Science
Vanbrugh College
University of York
Helsington, York YO10 5DD
United Kingdom

giuseppe.longobardi@york.ac.uk

Cristina Guardiano
Università di Modena e Reggio Emilia
Dipartimento di Comunicazione ed Economia
Viale A. Allegri 9
42121 Reggio Emilia
Italy

cristina.guardiano@unimore.it