Mathematical Modeling of Grammatical Diversity supports the Historical Reality of Formal Syntax

Giuseppe Longobardi
University of York / Trieste

Andrea Ceolin
University of Pennsylvania

Luca Bortolussi
Università di Trieste

Cristina Guardiano
Università di Modena e Reggio Emilia

Monica Alexandrina Irimia
Università di Modena e Reggio Emilia

Dimitris Michelioudakis
University of York

Nina Radkevich
University of York

Andrea Sgarro
Università di Trieste

***Abstract*—Recent studies have taken advantage of computational techniques to investigate the evolution of Indo-European languages [1-3]. However, these methods are not able to overcome the time constraints on lexical evolution, which limit a broader application of the Classical Comparative Method, and therefore cannot be used above the family level. For this reason, evidence from cross-family relationships must come from other domains (e.g. phonetics, [4, 5]). Reference [6] shows that another domain, syntax, is a potential source for cross-family comparison. In this paper, we evaluate the method proposed in [6], the PCM, and argue through a random generation of possible grammars that syntactic distances can be useful to detect signals of historical relatedness above the Indo-European level, within some confidence probabilistic intervals.**

Keywords—Parametric Comparison Method, Principles & Parameters, PCM, Computational Cladistics

I. INTRODUCTION

The Parametric Comparison Method (PCM, [6]) uses syntactic parameters [7-9] to study relationships among languages. This method has already been successfully applied to the study of Indo-European (IE) languages [10]. Syntactic parameters in the PCM are encoded as discrete binary values ('+' or '-') which characterize aspects of the syntax of all natural languages. For this paper, we have coded 75 parameters for 40 languages which belong to different linguistic families (Indo-European, Finno-Ugric, Semitic, Altaic, Sinitic and some isolated languages of Africa and South-America).

One of the problems in coding linguistic data is that the assumption of character independence is often violated, a fact with relevant consequences, if overlooked, as it introduces pervasive redundancy in calculating taxonomic information. An advantage of the PCM is that the method, based on a saliently hypothetico-deductive linguistic theory, allows for the coding of parametric implications, i.e. for making interdependencies explicit in the system: thus, some parameters are automatically assigned state '0' (undefined) when their value is predictable, excluding interaction between character values. Then, we calculate Jaccard distances (defined to range between 0 and 1) between all the parametric strings, excluding from them all the correspondences involving the '0' (or undefined) states.

A question facing all computational historical methods is whether their conclusions about language relatedness are secure against chance similarities between languages [11]. Reference [12] attempted to answer this question by using a randomly simulated distribution of parametric distances between languages to perform statistical tests on the hypothesis that the distances observed in the real world are unlikely to arise by chance, and thus to motivate judgments of relatedness based on syntax. In the section Materials and Methods, we propose a refinement of the algorithm of [12] that allows us to randomly generate 5000 possible languages and calculate their distances. In the section Results, the sample is compared with distances of real languages. In Discussion, we analyze the results of the comparison, which support one hypothesis of a super-family previously proposed in the literature (the Finno-Ugric/Altaic family).

This work was supported by ERC Adv. Grant 295736 LANGELIN.

II. Materials and Methods

To refute the null hypothesis that syntactic differences do not encode useful historical information [13] we need to generate a population of random parameter strings and compare their distances with those between parameterizations of the languages in our sample. Crucially, due to the implications between parameters, we cannot simply independently sample from the set {+, -, 0} at each locus.

Reference [12] presented a sampling algorithm to calculate the number of admissible strings of parameters -- those which respect a set of implications. They sampled strings from a *uniform* distribution over the population of languages which are allowed by the PCM's parameterizations. However, since the PCM incorporates parametric implications, some (unimplied) parameters are instantiated by all possible languages, whereas some others (heavily implied) are instantiated by only a small subset. Sampling from a uniform distribution over languages transmutes this asymmetry into uneven distributions of parameter values in the output. Instead of each parameter taking on values '+' and '–' with equal probability across simulated languages, some are more likely to be '+' and some are more likely to be '–' (for no justifiable reason).

Here is a linguistic example. Languages which have articles trigger a subset of parameters which are variable, and so the subset of languages with articles will be big. Languages without articles do not trigger further parameters in that parametric subset, and so the subset of languages without articles will contain just one possible language. Sampling from a uniform distribution of languages has the consequence of generating many more languages with articles than languages without articles, something which is against our linguistic intuitions.

In order to avoid sampling biases, we decided to fix all parameters at a 50:50 ratio of '+' to '–' in simulated languages (0s are not counted), in the wake of our corresponding idealization, in calculating real language distances, that parameter values are equiprobable.

The equiprobability assumption will prevent us here from running a simulation which is too dependent on the sample of languages in our database, and so inconsistent. In the absence of specific evidence about how 'third factor' [14] computational pressures might bias the frequency of certain parameters, this seems to us the only sensible way to proceed, given that syntacticians are far from reaching an agreement on any 'markedness' hypothesis (like the one proposed in [15]).

Note that the equiprobability assumption also holds for the calculation of the Jaccard distances among the real languages (which, so far, has been supported by its obtaining plausible phylogenetic results in [6] and [10]).

Of course, once efficient generating and sampling algorithms for possible languages are provided, they can be adjusted to other assumptions about distances and probabilities. E.g. a conceivable alternative to the sampling used here is generating parameters with a probability *p* for either value, calculated empirically from the variation observed in the sample itself. Let us call this methodology *empirical* sampling. Obviously, though promising, even this method runs into another way of biasing the results towards the observed sample, so that corrections too long to discuss here need to be provided.

Thus, we encoded in our algorithm all the rules behind parametric interdependencies, so that all the generated languages are compatible with the rules that constrain the real ones. Parameters are hierarchically ordered, with the independent ones at the top of the hierarchy, thus '0' values can be automatically assigned when a particular combination predicts the values of other parameters which are lower in the hierarchy.

III. Results

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

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

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

A high proportion of pairs below this threshold indicates closeness. We might expect that a high proportion of pairs is represented by pairs within the IndoEuropean family, which is indeed the case.

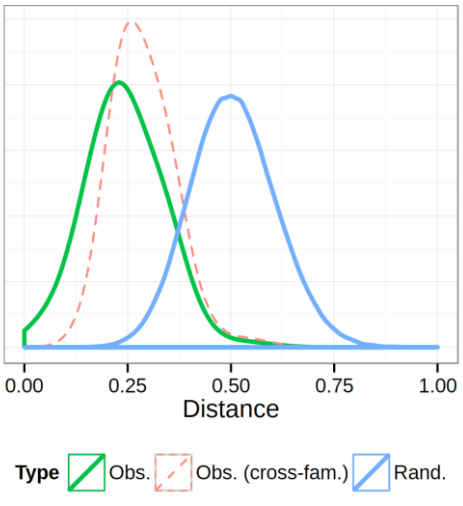

Fig. 1. Density plot comparing distances from real languages with distances from the languages generated through the sampling algorithm.

TABLE I.

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

The majority of the missing pairs include an Iranian language (Farsi or Pashto), showing that this sub-family is the one which exhibits the highest number of distances with other IE languages. However, it is remarkable that all the pairs between Finno/Ugric (Finnish, Hungarian and Estonian) and Altaic (Turkish and Buryat) languages are below the threshold.

While evidence for a Eurasiatic or Nostratic hypothesis is weak, the data suggest the possibility of a (primary or secondary) Ural/Altaic cluster which is undetectable through lexical comparison. This finding requires further investigation.

On the other hand, there are no pairs at all below this threshold which involve the Sinitic, South American, or African languages in our sample.

These results confirm that syntactic parameters provide a novel approach to the study of the prehistory of human languages: its results agrees with the outcomes of previous lexicon-based studies or other independently known historical variables and suggests the possibility of aiming toward a greater time depth, given that syntactic parameters are part of a universal faculty of language.

IV. Discussion

The results presented in Fig.1 and Table I, not being attributable to sheer chance, call for an explanation. Obviously, there are many possible interpretations for why the variability in the syntax of the languages of our sample is more constrained than one would expect from a random generation. Third factor principles in the sense of [14] and geographical influence are both alternatives to the strict phylogenetic hypothesis.

However, if we try to explain the results just in terms of third factors, this would not be enough to justify the presence of some historically plausible aggregations in the left tail of the real languages distribution: in fact, third factor principles should apply at a universal, not a particular scale. Changing the way in which the random languages are generated does not change the fact that some groups exhibit linguistic distances which are lower than others (though it might reduce the difference between the two distributions).

Therefore, if one wants to question the validity of the method and in particular the cross-family similarities, the only plausible argument is the geographical one.

Parametric resetting due to contact is rare [16] but not impossible. Many suggestions have been made of contact-mediated parametric resetting: in particular, syntactic changes in Old English have been attributed to Celtic and Scandinavian influence, and some uniformity in the syntax of Balkan languages has been attributed to the Balkansprachbund. However, these claims have never been supported by uncontroversial evidence. In particular, studies like [17, 18] have shown that many cases of syntactic change are independent of contact with another language (including many cases which have been traditionally attributed to contact), and in cases of obvious historical and sociolinguistic pressure the grammar tends to be conservative [19].

Recently, another kind of evidence against historically mediated change of abstract linguistic system is coming from sociolinguistics [20] and from studies that correlate linguistic, genetic and geographical variation. Reference [21] showed that in Europe the correlation between syntax and geography is null if we control for genetic similarity.

These arguments weaken the claim that secondary contact between population speaking different languages is the main explanation for why the syntactic variation found in the world languages is historically constrained.

V. Conclusion

We provided an algorithm for generating random languages and modeling the space of variation taking into account implications among parameters, one of the distinctive contributions to the deductive structure of modern linguistic theory.

The Principles-and-Parameters approach to syntax proved particularly amenable to quantitative analysis, and retrieved a high level of correct historical information as judged against independently known taxonomies. When a large number of formal parameters is compared, syntactic distances produce plausible results.

More specifically, our results provided some evidence for a convergence between the Finno-Ugric and Altaic languages, which must be further investigated. Hopefully, increasing the number of languages will help us explain which are the historical processes that lead to these results and to what extent syntactic borrowing influences the classification of these families.

Acknowledgment

We are especially grateful to Aaron Ecay for the modeling of the experiments, to Bruna Franchetto and Filomena Sandalo for making the analysis of South American languages possible, to Peter Sells for helpful feedback on these ideas, and to all our native speaker consultants.

REFERENCES

[1] I. Dyen, J. Kruskal, P. Black. "An Indo-European classification: a lexicostatistical experiment", Transactions of the Philosophical Society, 82 (5), 1992.

[2] D. Ringe, T. Warnow, A. Taylor. "Indo-European and computational cladistics", Transactions of the Philological Society, 100 (1), 59-129, 2002.

[3] R. Bouckaert et al., "Mapping the origins and expansions of the Indo-European language family", Science, 337(6097), 957-960, 2012.

[4] B. Kessler, "Phonetic comparison algorithm", Transactions of the Philological Society, 103 (2), 243-260, 2005.

[5] G. Jaeger, "Support for linguistic macrofamilies from weighted sequence alignment", PNAS, USA, 112(41): 12752, 12757, 2015.

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

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

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

[9] I. Roberts, "On the nature of syntactic parameters: a program for research", Parameter Theory and Language Change, eds C. Galves, S. Cyrino, R. Lopes, F. Sandalo, J. Avelar. Oxford University Press, Oxford, pp. 319-334, 2012.

[10] G. Longobardi, C. Guardiano, G. Silvestri, A. Boattini, A. Ceolin, "Toward a syntactic phylogeny of modern Indo-European languages", Journal of Historical Linguistics 3(1):122-152, 2013.

[11] J. Nichols, "The comparative method as heuristics". The comparative method reviewed, eds M. Durie, M. Ross, Oxford University Press, Oxford, pp.39-71, 1996.

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

[13] F. J. Newmeyer. "Possible and probable languages. A generative perspective on linguistic typology", Oxford University Press, Oxford, 2005.

[14] N. Chomsky, "Three factors in language design", Linguistic Inquiry 36:1-22, 2005.

[15] I. Roberts, Diachronic Syntax. Oxford University Press, 2007.

[16] S. G. Thomason, T. Kaufman, "Language Contact, Creolization and Genetic Linguistics, University of California Press, Berkeley, 1988.

[17] C. Gianollo, "Native Syntax and translation effects: adnominal arguments in the Greek and Latin New Testament", Indo-European syntax and pragmatics: contrastive approaches, ed. Eirik Welo, Oslo Studies in Language 3(3), 75-101, 2011.

[18] G. Longobardi, "Convergence in Parametric Phylogenies: Homoplasy or Principled Explanation?", Parameter Theory and Language Change, eds C. Galves, S. Cyrino, R. Lopes, F. Sandalo, J. Avelar. Oxford University Press, Oxford, pp. 304-319, 2012.

[19] C. Guardiano et al., "South by South East", L'Italia dialettale, 2015.

[20] W. Labov, "Transmission and Diffusion", Language 83(2), 344-387, 2007.

[21] G. Longobardi et al., "Across Language Families: Genome diversity mirrors linguistic variation within Europe", American Journal of Physical Anthropology, 157(4):630-640, 2015.