Computational Phylogenetics
&
Language (Pre)history

Achievements, Challenges and Prospects

Student Cultural Center “XENIA”

23-25 May 2023

Abstract Booklet

ModelGloss

MODELLING GLOSSOGENY

Computational Phylogenetics and Language (Pre)history: Achievements, Challenges and Prospects

University of Crete

23rd-25th May 2023

Presentations

Approaching the language faculty with phylogenetic methods

Balthasar Bickel (University of Zurich)

One of the most specific features of the language faculty is its temporal dynamic. Characterizing what has evolved in the hominin lineage therefore requires capturing this dynamic at the species level. While phylogenetic modeling seems ideally suited for this task, two challenges have been noted. First, the current distribution of language data is heavily skewed by mass extinction of languages during post-Neolithic population history, so our models might not generalize to the species level. Second, the complex, multivariate nature of language makes it hard to estimate state transitions, so our models might miss key properties of the language faculty. While the first challenge can arguably be best addressed by convergence with non-linguistic evidence, I here mainly focus on the second challenge. I will review recent work in my group on models capturing the evolutionary dynamics of various splits and dependencies in language.

Adapting methods from evolutionary biology to explore language evolution.

Lindell Bromham (Australian National University)

The past decade has seen a flowering of collaboration between evolutionary biologists and linguists. Useful analytical tools from evolutionary biology have been modified to provide new ways of asking interesting questions about language change, including: What influences individual?s mix of language variants? How do the frequencies of language variants change over time? Do smaller or larger populations have faster rates of language change? What factors generate global patterns of language diversity? What are the drivers of language loss? I will illustrate these interdisciplinary approaches with case studies at local, regional and global scales of language diversity and change. Studies of language diversity are given a degree of urgency by the current crisis of language loss: using methods adapted from macroevolution and macroecology, our research suggests that global rates of language loss could triple within forty years, equivalent to at least one language lost per month.

The added value of comparative phylogenetic methods as an instrument for language reconstruction: a look at gender systems.

Gerd Carling (Goethe University Frankfurt am Main)

The used of phylogenetic comparative methods as an instrument for language reconstruction is being more and more established in the scholarly community. The method has evident advantages in favour of more traditional models, such as the comparative method or diachronic typology. However, there are also shortcomings. The most obvious advantage of comparative phylogenetic reconstruction is the possibility to make reconstructions based on large amounts of data, gauging the probability of a reconstruction by Bayesian methods, considering the branches and nodes of an underlying family tree structure. Probably the most striking disadvantage is the inability of the model to reconstruct any material that is not given in the input data, something that can be done by applying the comparative method and diachronic typology. In the lecture, I will look specifically at the linguistic feature gender at two different levels: family and worldwide level. I will discuss the results of reconstructing gender by a phylogenetic comparative model in the two families of Indo-European and Arawak (which are gendered families), and compare the results with reconstructions achieved by the comparative method and diachronic typology. At least in Indo-European, the results differ between a model using the comparative method and diachronic typology, whereas this is likely not the situation in Arawak (the analysis of Arawak is work in progress), where the gender systems are more evident. The results from these two families will then be compared to results of a phylogenetic reconstruction at global level, involving a database of gender based on 3,700 languages. The studies indicate that factors such as frequency and economy, as well as grammatical hierarchies, impact the process of grammar change. This in turn has is relevant for the reliability of a reconstruction, since it enables us to gauge a reconstruction by considering how likely a specific change it is to appear.

References

Allassonnière-Tang, M., et al. (2021). Expansion by migration and diffusion by contact is a source to the global diversity of linguistic nominal categorization systems. *Nature Humanities & Social Science - Communications* 8:331.

Carling, G. and C. Cathcart (2021). Reconstructing the evolution of Indo-European grammar. *Language* 97(3): 561-598.

Luraghi, S. (2011). The origin of the Proto-Indo-European gender system: Typological considerations. *Folia Linguistica* 45(2): 435-464.

Luraghi, S. (2014). Gender and word formation: the PIE gender system in a cross-linguistic perspective. In: *Studies on the Collective and Feminine in Indo-European from a Diachronic and Typological Perspective*, Brill: 199-231.

Matasović, R. (2004). *Gender in Indo-European*. Heidelberg, Winter.

More Weight!

James Clackson (University of Cambridge)

Many researchers in computational phylogenetics limit themselves to lexical data. The lexicon is well-adapted to large-scale, computational analysis because of the size of the lexical corpus, the ready availability of word-lists for many languages, and the comparative ease with which it is possible to make judgements about cognacy. As has long been recognised, however, long periods of language contact may have led to massive lexical borrowing in prehistory, with the consequence that lexical comparison may not give an accurate picture of phylogeny. The use of syntactic features for constructing phylogenies (see, for example, Guardiano et al. 2020) also produces results which run counter to those arrived at by non-computational methods for similar reasons. Some computational phylogenetic work has attempted to combine lexical, morphological and phonological data (such as Ringe et al. 2002) but these the face a problem of how to weight the importance of shared sound-changes or morphological agreements: should a trivial sound change such as palatalisation of velars before front vowels count less than a rare and unusual change? does a shared innovation of a morphological feature count the same as a shared lexical item? In this paper I shall discuss the issue of weighting in computational phylogenetics and propose some ways in which we can adjust current methods to add more weight.

References

Guardiano, C., Longobardi, G., Cordoni, G. & P. Crisma (2020) Formal syntax as a phylogenetic method. *The Handbook of Historical Linguistics* 2: 145-182

Ringe, D., Warnow, T. & A. Taylor. (2002). Indo-European and Computational Cladistics. *TPS*, 100, 59-129.

What do we need phylogenies for, if they're not the end goal? (Weak) biases and selection in language evolution.

Dan Dediu (ICREA & University of Barcelona)

Most of us need some kind of representation(s) of (aspects of) language history in our work. Ideally, these representations should be quantitative, computer-readable, come with an estimation of error/posterior distribution, and have transparent assumptions that arguably fit what we believe really happens/happened. In my own little corner, I need to use such representations to check for (or detect) various kinds of adaptive (i.e., non-neutral) processes affecting aspects of language, in particular, potentially very weak forces due to the environment (writ large). To make things clearer I will present a few cases of such forces and the methods we use to identify and study them, and I will argue that such forces have a fundamentally temporal dimension ranging across several scales (from within individuals, to between individuals, and across communities and generations). While phylogenetic methods should be appropriate for capturing and studying aspects of this dynamics, it is currently unclear to me how precisely we should use them without "throwing the bay with the bathwater", in the sense that we need to control for the influence of history on our inferences while preserving our capacity to detect (very) weak and complex processes of language evolution.

The evolutionary behavior of words, languages, and texts

Michael Dunn (Uppsala University)

Language is our best and most tractable example of a cultural evolutionary system, and computational evolutionary modelling lets us infer a great deal about historical processes from language. But the kinds of data that we model can be quite diverse. They require different kinds of models, and offer different possible inferences, in many cases tracking different aspects of the historical process, or even separate histories. 'Typical' Bayesian Phylogenetic Inference models the history of Swadesh list exponents?terms which are not only cognate, but which have a meaning equivalent to a particular Swadesh list item. This is done because modelling true cognates (i.e. allowing for possible semantic drift) would require access to the complete lexicon of all the languages under analysis, which would be impractical or impossible. Abstract structural features are also often subject of Bayesian Phylogenetic Inference, most often as part of a comparative method analysis investigating the evolutionary behaviour of the feature itself. Such features differ markedly from lexical cognates, in particular because chance homoplasy is possible and expected. As such a different class of models is appropriate for such features. In this paper I will report some research investigating the evolutionary behaviour of different kinds of subsystems within linguistics and stemmatology, and present notes towards a typology of the evolutionary behaviour of linguistic and language systems, with some suggestions towards how they can and should be analysed.

From phonology to phylogeny: Toward event-based modeling in historical linguistics

David Goldstein (University of California, Los Angeles)

Linguistic phylogenies are typically inferred on the basis of lexical cognate relationships (e.g., Bouckaert et al. 2012, Chang et al. 2015, Sagart et al. 2019). Despite the predominance of this practice, it suffers from well-known drawbacks. First, it disregards the phylogenetic signal that exists in the form of the words themselves. Second, it limits the modeling possibilities since it relies on an arbitrary coding of the data. In this talk, I introduce a novel framework for linguistic phylogenetics that overcomes both of these shortcomings. The heart of this framework is the TKF91 model (Thorne et al. 1991), which allows phylogenetic inference to be carried out directly from word-forms. This model not only opens up a new horizon in the study of linguistic phylogenetics, but allows historical linguists to investigate questions of sound change that were previously out of reach.

References

Bouckaert, Remco R., Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson (2012). Mapping the origins and expansion of the Indo-European language family. *Science* 337.6097 (Aug. 2012), 957?960. doi: 10.1126/science.1219669.

Chang, Will, Chundra Aroor Cathcart, David P. Hall, and Andrew J. Garrett (2015). Ancestry- constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91.1 (Mar. 2015), 194?244. doi: 10.1353/lan.2015.0005.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List (2019). Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences of the United States of America.* doi: 10.1073/pnas.1817972116.

Thorne, Jeffrey L., Hirohisa Kishino, and Joseph Felsenstein (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* 33.2 (Aug. 1991), 114?124. doi: 10.1007/BF02193625.

Quantifying the Evolutionary Dynamics of Language Systems

Simon Greenhill (University of Auckland & Max Planck Institute for Evolutionary Anthropology)

Language is an evolving system that is comprised of characteristics that are passed onto new speakers and daughter languages, some of these traits persist while others do not. Recent years have seen an influx studies applying phylogenetic methods to languages. However, we have only just scratched the surface of using these tools to gain valuable insights into the processes that shape languages over thousands of years. As the evolutionary biologist G.G. Simpson noted, one of the most fundamental evolutionary dynamics is the rate of change: Rates are the signatures of selection such that quantifying the tempo of change allows us to identify the processes driving this change.

In this talk I will present a series of case studies that investigate the rates of language evolution and use these rates to provide insight into the factors shaping diversity. I will begin by quantifying the similarities and differences between rates of lexical and grammatical evolution to show that most grammatical features actually change faster than items of basic vocabulary, but that there is a core that are highly stable. Strikingly, the slowly evolving grammatical features tend to be those that are more covert and less available to sociolinguistic reflection by speakers. Further, the lexicon shows more changes linked to language diversification events than the grammar, while the grammar shows higher rates of conflicting signal ('homoplasy'). Our results suggest that different subsystems of language have differing dynamics driven by different causal factors. I will then move on to present some newer work discussing what factors might drive these rates by applying a Bayesian model to quantify spatial and phylogenetic signal on a global sample of grammatical data ('grambank') and numeral systems ('numeralbank').

Generative syntax and language (pre)history

Cristina Guardiano (Università di Modena e Reggio Emilia)

The development of quantitative phylogenetics has prompted an enormous progress in historical linguistics, thanks to the introduction of sophisticated quantitative methods and computational techniques which allow accurate and objective reconstructions. Such methods have largely (though not exclusively) been implemented on the taxonomic characters traditionally used as key evidence for historical relatedness (word etymologies based on crosslinguistic sound regularities). The latter are very improbable phenomena, hence inevitably rare, which constitute compelling proof of historical relatedness, but are impossible to retrieve across long-separated languages; hence, they cannot be used as heuristics when it comes to exploring long-range (pre)historical relations.

Carrying out the intuition that the problem of deep-time investigation at a cross-family level can only be solved by means of a radical shift in the taxonomic characters used for comparison, the Parametric Comparison Method (PCM) has suggested that the abstract cognitive structures discovered by generative syntax are apt for this purpose thanks to their *abstract, discrete,* and *universal* nature. Using the toolkits provided by 20th century theoretical linguistics and quantitative phylogenetics, the PCM has been pursuing a "phylogenetics of grammars" relying on computational representations of syntactic distances and statistical procedures. The PCM has shown that syntactic parameters in fact retain a historical signal that not only matches the results of classical etymological classifications but also suggests statistically significant cross-family aggregations, proving that the method is effective in addressing issues of long-distance language relations, and can thus contribute to a deep investigation of the human past. These results not only confute, after two hundred years, the bias against the use of syntactic information for phylogenetic reconstruction, but also indicate that generative syntax can play an explanatory role as a historical cognitive science, over and above its success as a theory of language knowledge and language acquisition, prompting in language phylogenetics a *qualitative* revolution able to complement the *quantitative* one.

General and Language-specific Aspects of Phylogenetic Inference

Luise Häuser (Karlsruhe Institute for Technology) &
Alexandros Stamatakis (Foundation for Research and Technology & Heidelberg Institute for Theoretical Studies & Karlsruhe Institute for Technology)

General Part (Alexandros Stamatakis)

In the first part of this talk we introduce the concept of phylogenetic difficulty, that essentially quantifies the signal strength of a given alignment. We will outline how difficulty can be predicted via our Pythia machine learning tool **prior** to conducting a phylogenetic inference under Maximum Likelihood. We then use this predicted difficulty to accelerate phylogenetic inferences with RAxML-NG by a factor of 3.

Language Part (Luise Häuser)

We then present a database containing a plethora of linguistic MSAs and quantify how language data differs from biological morphological data taking into account the difficulty as predicted by Pythia and tree inference results with RAxML-NG. We will also outline how the selection of cognate candidates affects trees inferred with RAxML-NG and how candidate selection uncertainty can be systematically represented via probabilistic MSAs.

Beyond cognacy

Gerhard Jäger (University of Tübingen)

Most work in phylogenetic linguistics uses manually annotated characters as basis for phylogenetic inference. These characters can be grammatical and typological features or cognate classes. Since cognate-coded vocabulary items are more numerous and easier to obtain than more abstract features, cognacy characters can be considered state of the art.

There are three major drawbacks to this methodology:

- Manually annotating for cognacy requires the annotator to have some hypotheses about the phylogenetic structure and sound laws pertaining to the languages under investigation. This might lead to an unconscious bias.
- By definition of the term *language family*, only languages from the same family can have cognate words. This makes the open-ended search for deep relationships impossible in this framework, because any such relationship has to be identified *a priori* via cognate coding.
- Classical historical linguistics identifies family trees not just on the basis of cognates, but also uses sound changes. This valuable source of a phylogenetic signal is not used when relying only on cognacy characters.

Several authors have proposed machine learning techniques for automatic cognate detection (Jäger et al. 2017, List et al. 2017, among many others). This approach potentially addresses the first two issues, but not the third.

In this talk I will present a novel approach to infer phylogenetic characters from unannotated multilingual word lists. Words, i.e., strings of sound-class symbols, are embedded into a high-dimensional space using a deep network. This network is trained in such a way that the embeddings of cognate words are close together, and the distance between them is correlated with the amount of sound change and morphological separating them. From this high-dimensional continuous representation, discrete binary characters are extracted which can be fed into standard phylogenetic inference algorithms.

References

Jäger, G., J.M. List & P. Sofroniev (2017), Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *15th Conference of the European Chapter of the Association for Computational Linguistics. Proceedings of Conference, Volume 1: Long Papers*, ACL, 1204-1226.

List, J.M., S.J. Greenhill & R.D. Gray (2017) The Potential of Automatic Word Comparison for Historical Linguistics, *PLoS ONE*, 12.

Why appropriate measures and methods in typology and historical linguistics matter

Annemarie Verkerk (Universität des Saarlandes)

Historical linguistics and typology are traditionally non-quantitative disciplines within linguistics. Around the turn of the century, methods taken from evolutionary biology started being applied to answer questions from these two disciplines. Controversy arose. In this talk, I review some of the history associated with this quantitative turn since 2000 and highlight some emerging new approaches. Ultimately, I argue that appropriate measures and methods matter because they enable a more empirical approach to language change that is essential to the core questions asked by typologists and historical linguists.

Population genetic approaches as an alternative to phylogenetics in revealing linguistic history

Outi Vesakoski (University of Turku)

The phylogenetic - or better phylolinguistic - approach produces genealogies that reflect the vertical evolution of the family. This is indeed one part of the evolutionary history of languages families. Some studies have stepped beyond treelike evolution, and studied horizontal development of families or contacts between families by using population genetic admixture models. Model based clustering analyses (eg. STRUCTURE, ADMIXTURE, BAPS) are built for studying population genetic data where contact-driven convergence between populations operate alongside with diverging, isolating forces. We tested the technique with Uralic typological data (N=165 binary UT traits, uralic.clld.org) with STRUCTURE-like analyses (Norvik et al. 2022) and identified four linguistic areas or Sprachbünde. The new picture taking together both vertical and horizontal evolution hints at more complex linguistic evolutionary and contact dynamics within the study area than what tree model indicate. I will demonstrate the possibilities of this approach by first presenting some of our preliminary work on finding the contact areas within the North-Eurasian languages and second by presenting our studies on drivers of divergence and convergence of Finnish dialects measured with the new metrics the STUCTURE/BAPS provides.

Typological features and diachrony: bigger data and better ideas

Søren Wichmann (Kiel University)

Largely only preceded by Nichols (1992), the time from the mid-2000?s onwards saw a boom in the interest of understanding the diachronic behavior of abstract typological features, i.e. features of language structure not tied to specific linguistic forms. Some expressed hope that such features might extend the temporal reach of the comparative method (Dunn et al. 2005), while other were skeptical of this idea, pointing to the areal sensitivity of abstract typological features and their non-tree-like behavior (Greenhill et al. 2010, Donohue et al. 2011). The differential stabilities of typological features were computed by different metrics in various works, including Wichmann and Holman (2009). It was investigated whether Bayesian phylogenetic correlations would support implicational universals of syntax (Dunn et al. 2011). There were many more interesting studies, including correlational studies of extra-linguistic factors influencing language structure (e.g., Lupyan and Dale 2010), which I will not be concerned with in this talk. While WALS, published in 2005, spawned much of this work, it has very recently (2023.04.19) been largely superseded by Grambank, allowing us to rerun and extend some of the more interesting analyses of the past. This is the aim of this talk. I will present stability calculations for the Grambank features focusing somewhat more on explanations than methods, discuss how typological features might productively be used in historical linguistic research even if they are not particularly tree-like in their behavior, and I will show how phylogenetic correlational studies can, in principle, be extended to phylogenies of all the world?s languages through the use of tree topologies from Glottolog supplied with branch lengths from ASJP. While most of the results will be preliminary I hope to at least point to some exciting avenues of research.

References

Donohue, Mark, Simon Musgrave, Bronwen Whitting, and Søren Wichmann. 2011. Typological feature analysis models linguistic geography. *Language* 87.2: 369-383.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473: 79?82.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309: 2072-2075.

Greenhill, Simon J., Quentin D. Atkinson, Andrew Meade, and Russell D. Gray. The shape and tempo of language evolution. *Proceedings of the Royal Society B: Biological Sciences* 277: 2443-2450.

Lupyan, Gary and Rick Dale. 2010. Language structure is partly determined by social structure. *PLOS ONE* 5(1): e8559. https://doi.org/10.1371/journal.pone.0008559

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

Wichmann, Søren and Eric W. Holman. 2009. *Temporal Stability of Linguistic Typological Features.* München: LINCOM Europa.

⋆ The large and well-known databases WALS, Grambank, Glottolog, ASJP are excluded from the references to save space.

Evaluating the phylogenetic signal of morphosyntax

The ModelGloss Team

Computational linguistic phylogenetics has so far relied heavily on cognate data, which have been extensively analyzed over the past decades and have produced phylogenies largely aligning with existing knowledge of language history. In contrast, the potential of morphosyntactic characters as a valuable source of data for phylogenetic analysis has been largely overlooked. Such characters can provide insights into aspects that cognate data cannot address, especially with respect to genealogical/historical relationships beyond individual language families. Notably, however, recent studies employing morphosyntactic characters have not reconstructed the phylogeny of Indo-European languages with accuracy and/or with significant statistical support. In this study, we explore the usefulness of the World Atlas of Language Structures (WALS) data for reconstructing phylogenies, with a specific focus on Indo-European languages. We constructed a table with 425 states of WALS features as (binary) taxonomic characters and 60 IE languages, providing our own values for >70% of the cells of the table. It turns out that WALS (or rather WALS-type) data often contain a strong phylogenetic signal, but fail to yield a purely historical tree of IE. We subjected these initial characters to extensive linguistic evaluation, which involved reformulation of many of them on the basis of theoretical, historical and typological reasoning, and constructed a new table of 530 characters for the same Indo-European languages and dialects. We used this table too to generate phylogenies. Although the resulting tree largely aligns with a cognate-based tree, consistent discrepancies are observed. We argue that these inconsistencies arise from the quantity and quality of the data employed. While cognate data comprise a few thousand entries, morphosyntactic data are counted in hundreds (at best). Moreover, the morphosyntactic data currently employed for phylogenetic analysis lack qualitative filtering and contain elements prone to horizontal transfer or homoplasy, which obscure the underlying phylogenetic signal. To address these issues, we propose three novel methods that leverage both linguistic expertise and computational approaches to evaluate morphosyntactic data, effectively distinguishing between vertically transmitted and horizontally transmitted or homoplastic data, or relativising such properties to specific scales of variation, families, language types or areas.

Acknowledgements

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: HFRI-FM17-44)