C-LESTE

Cross-Linguistic Explorations in Space, Time and Evolution

Programme & Abstracts

Content

Programme

Day 1 - Friday, 14.06.2024

Session 1 (Chair: Miri Mertner)

Place: Brechtbau, Room 0.27 (Wilhelmstraße 50, 72074 Tübingen)

- **9:00**: *Further on beyond cognacy*

  Gerhard Jäger (University of Tübingen)

- **9:30**: *Evaluating morphosyntax in an evolutionary perspective*

  Elena Anagnostopoulou (University of Crete and IMS FORTH)

- **10:15**: *The coevolution of sound patterns and the lexicon*

  Chundra Cathcart (University of Zurich)

- **11:00**: *coffee break*

- **11:30**: *The evolutionary dynamics of noun categorization*

  Gerd Carling (Goethe-University Frankfurt)

- **12:15**: *Areal contacts based on linguistic similarity*

  Outi Vesakoski (University of Turku)

- **1:00**: *lunch (1 hour 30 mins)*

Session 2 (Chair: Gerhard Jäger)

Place: Oberschulamt, Room 0.01 (Keplerstraße 2, 72074 Tübingen)

- **2:30**: *Recent Advances in General Phylogenetic Inference and Energy-efficient Computing*

  Alexandros Stamatakis (ICS FORTH)

- **3:15**: *From Phylogenetics into Historical Linguistics - Some Insights*

  Luise Häuser (Heidelberg Institute for Theoretical Studies)

- **4:00**: *coffee break (30 mins)*

- **4:30**: *The Dialect Chain Tree*

  Erik Elgh and Harald Hammarström (Uppsala University)

- **5:15**: *"Dollo" after all? Puzzling behaviour of models of vocabulary evolution can be explained by effects of ascertainment correction*

  Philipp Rönchen (Uppsala University)

- **6:00**: Conference closes for the day

- ***7:00**: Conference dinner at Ludwigs*

Day 2 - Saturday, 15.06.2024

Session 1 (Chair: Miri Mertner)

Place: Oberschulamt, Room 0.01 (Keplerstraße 2, 72074 Tübingen)

- **9:00**: *Resurrecting the Goblin: New Experiments With Bouchard-Côté's Reconstruction System*

  Johannes Dellert (University of Tübingen)
- **9:45**: *Computer-Assisted Language Comparison with EDICTOR 3*

  Johann-Mattis List (University of Passau)
- **10:30**: *coffee break*
- **11:00**: *Rates and/or stationary distributions? An exploration using tone and colexification*

  Dan Dediu and Thomas Brochhagen (University of Barcelona; Pompeu Fabra University)
- **11:45**: *The Uralic languages: trees and forests*

  Outi Vesakoski and Michael Dunn (University of Turku; Uppsala University)
- **12:30**: *lunch (1 hour 15 mins)*

Session 2 (Chair: Gerhard Jäger)

Place: Oberschulamt, Room 0.01 (Keplerstraße 2, 72074 Tübingen)

- **1:45**: *Why modeling space is hard: unidirectional contact, expansion events and other stuff*

  Matías Guzmán Naranjo (University of Freiburg)

- **2:30**: *Spatial models of linguistic diversity in Africa*

  Miri Mertner (University of Tübingen)

- **3:15**: *A new approach to the diversification of ancient Greek*

  David Goldstein (University of California)

- ***4:00**: break (15 mins)*

- **4:15**: *Spotting implicational relations among typological features through coupled evolution*

  Søren Wichmann (Kiel University)

- **5:00**: *Navigating linguistic trait space: Analogical modelling of inflectional evolution*

  Erich Round (Surrey Morphology Group)

- **5:45**: Closing remarks (Gerhard Jäger)

- **6:00**: Conference closes for the day

- **7:30***: Dinner at Freistil*

Abstracts

Further on beyond cognacy

Gerhard Jäger (University of Tübingen)

TBA

Evaluating morphosyntax in an evolutionary perspective

Elena Anagnostopoulou (University of Crete and IMS FORTH)

TBA

The coevolution of sound patterns and the lexicon

Chundra Cathcart (University of Zurich)

Formal properties of the vocabularies of the world's languages are argued to be communicatively optimal, aiding in acquisition, processing, and production. With exceptions, little research has focused on the specific mechanisms of change affecting word forms which bring about communicatively beneficial patterns. In this talk, I present results elucidating the evolutionary dynamics involved in two phenomena of this sort, identical consonant avoidance and sound symbolism.

The evolutionary dynamics of noun categorization

Gerd Carling (Goethe-University Frankfurt)

Collaborators: Noor Efrat-Kowalsky, Marc Allassonière Tang, Lev Michael, Filip Larsson, and Niklas Erben Johansson

In the presentation, I will describe ongoing research of reconstructing the evolutionary dynamics of noun categorization, both at a typological and a lexical level. Noun categorization is found in about half of the world's languages, and of these, about 20% have an (limited or extended) gender system, 7,5% a noun class system, and about 26% a classifier system (Allassonnière-Tang et al. 2021). In gender an noun class languages, lexemes are (partly or completely) assigned to a category, such as masculine, feminine, or neuter (Corbett 1991, 2013), something that the current study investigates from a statistical and evolutionary perspective. For the purpose, we have compiled a typological data set with (nominal and pronominal) gender, noun class and classifier data from 3240 languages. This data set is completed by a more fine-grained data set of typological data, and a set of gender-coded lexical data, from the Indo-European and Arawakan language families. We can demonstrate that the semantic core of gender assignment, in particular the connection between gender and animacy, has an impact on the evolutionary rates of gender change, both in the lexicon and in the grammar. We run parallel evolutionary tests on typological and lexical data, using a phylogenetic reconstructing model (MCMC, RJHP, BayesTraits, based on a worldwide tree (Bouckaert 2022)). The model defines the transition rates of gender systems as well as classes of concepts (together with their gender), and reconstructs the most likely system at the roots of families by a comparative phylogenetic model (Jäger 2019; Carling and Cathcart 2021). Results indicate an all-over strong preference for sexus-based systems, as well as semantic influence on both assignment principles as well as transition rates. The tendency is obvious in the world-wide data, in spite of a variation in individual families.

Allassonnière-Tang, Marc, Olof Lundgren, Maja Robbers, Sandra Cronhamn, Filip Larsson, One-Soon Her, Harald Hammarström, and Gerd Carling. 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.

Bouckaert, Remco, David Redding, Oliver Sheehan, Russell Gray, Kate E Jones, Quentin Atkinson. 2022. Global language diversification is linked to socio-ecology and threat status. SocArxiv Papers.

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

Corbett, Greville G. 1991. *Gender*, *Cambridge textbooks in linguistics, 99-0104661-0*. Cambridge: Cambridge Univ. Press.

Corbett, Greville G. 2013. "Gender typology." In *The Expression of Gender*, edited by Greville G. Corbett, 87-130. Berlin - New York: Mouton de Gruyter.

Jäger, Gerhard. 2019. "Computational historical linguistics." *Theoretical Linguistics* 45 (3/4):151-182. doi: 10.1515/tl-2019-0011.

Areal contacts based on linguistic similarity

Outi Vesakoski (University of Turku)

TBA

Recent Advances in General Phylogenetic Inference and Energy-efficient Computing

Alexandros Stamatakis (ICS FORTH)

In this talk I will initially discuss some recent advances in general (i.e., not only language-specific) phylogenetic inference.
First, I will introduce a machine-learning based method that can predict the phylogenetic inference difficulty for a given dataset.
Then, I will present a novel machine-learning based method for the rapid, yet accurate prediction of bootstrap support values.
Further, I will show how difficulty prediction can be deployed to unravel potential biases in the design of computational experiments in phylogenetics.
In the second part of the talk, I will present a novel method called ecofreq that can help to reduce the CO2 footprint and cost of scientific computing.

From Phylogenetics into Historical Linguistics - Some Insights

Luise Häuser (Heidelberg Institute for Theoretical Studies)

Approaching the intersection of phylogenetics and historical linguistics from the perspective of a computer scientist leads me to some interesting insights as well as many open questions.
The work focuses on finding suitable ways to represent the input data and model the underlying evolutionary processes. Using cognate data as input involves handling synonyms, i.e. multiple words that describe the same concept in a language.
Binary character matrices, which are used as input for computational methods, do allow for representing the entire dataset including all synonyms.
I address the question of whether one should include all synonyms or whether it is better to select the synonyms a priori.
Maximum likelihood tree inferences with the widely used RAxML-NG tool yield plausible trees when all synonyms are used as input.
Furthermore, the a priori selection of synonyms can lead to topologically substantially different trees, so that I advise against doing so.
To represent cognate data including all synonyms, I introduce two types of character matrices that go beyond the standard binary matrices:
probabilistic binary and probabilistic multivalued character matrices.
However, it is data-dependent for which character matrix type the derived RAxML-NG tree is topologically closest to the gold standard. Further interesting findings result from the investigation of rate heterogeneity for binary character matrices that represent cognate data.
Rate heterogeneity describes the phenomenon that different columns of character matrices evolve at different rates.
For some cognate data sets, almost no rate heterogeneity occurs, which is very unusual, especially when compared to molecular data.
Various analyses show, among other things, that the low rate heterogeneity is associated with poor data quality.
What other reasons there might be, however, remains an open question. In the end, I come to a larger open question I keep asking myself:
Is the data available enough for appropriate phylogenetic inferences with computational methods or are we to reach out for sources where we can acquire more data from?

The Dialect Chain Tree

Erik Elgh and Harald Hammarström (Uppsala University)

The Dialect Chain Tree (DCT) is a new model that systematically combines language family trees and waves during the early stages of linguistic divergence. Here, we present the rationale behind the model as well as its basic functions and assumptions. Furthermore, we elaborate on a framework where Maximum Parsimony can be used to discriminate between different DCTs.

“Dollo” after all? Puzzling behaviour of models of vocabulary evolution can be explained by effects of ascertainment correction

Philipp Rönchen (Uppsala University) & Tilo Wiklund (UAB Sensmetry)

Phylolinguistic methods make inferences about the divergence of protolanguages by evaluating the patterns of inheritance in the vocabulary of modern languages, given the specification of a model of vocabulary evolution. There are essentially two types of models of vocabulary evolution that are frequently used today: one is the Stochastic Dollo model (Nicholls and Gray, 2006), which assumes that each group of words which are cognate goes back to one proto-form that appeared at some point in the language tree. Another way to say this is that the every “cognate class” appears only once in the language tree, and after that it only spreads by inheritance.
Models of the second type are two-state continuous time Markov chains (“CTMC” model), for which the same cognate class can appear arbitrarily often in different parts of the tree. The commonly used “Covarion” model (Tuffley and Steel, 1998) is a slight modification of the basic CTMC model. An issue with CTMC type models is that the total number of cognate classes in the language tree cannot be derived from the observed number of cognate classes in the data. To avoid biasing inferences due to unobserved cognate classes, an ascertainment correction (Felsenstein 1992, Bouckaert et al. 2018) is used.
In this talk we show that the ascertainment correction of the CTMC model has peculiar effects that greatly change the evolutionary dynamics of the model. It turns out that both for real and simulated data sets, the CTMC model with ascertainment correction becomes similar to a “Dollo death model”, which is a simplified version of the Stochastic Dollo model. This has consequences for how the realisticness and appropriateness of different models of vocabulary evolution should be evaluated.

Resurrecting the Goblin: New Experiments With Bouchard-Côté's Reconstruction System

Johannes Dellert (University of Tübingen)

TBA

Computer-Assisted Language Comparison with EDICTOR 3

Johann-Mattis List (University of Passau)

Computer-assisted approaches to historical and typological language comparison have made great progress over the past two decades. Specifically for the classical tasks of historical language comparison, many computational methods have been published that mimic certain steps of the traditional workflow of the comparative method. In contrast to the diversity of new computational methods, there is only a limited number of interactive tools that help scholars to curate and refine their data both prior and after the application of computational methods. One of the few publicly available interfaces, the EDICTOR (https://edictor.org), an interactive tool for computer-assisted language comparison, has been around for some time, allowing scholars to annotate and align cognate sets in various ways. With EDICTOR 3, the original tool is enhanced in various ways, offering various new features, including simplified automated methods for cognate detection, phonetic alignment, and correspondence pattern inference, to facilitate computer-assisted workflows.

Rates and/or stationary distributions? An exploration using tone and colexification

Dan Dediu and Thomas Brochhagen (University of Barcelona; Pompeu Fabra University)

There are various reasons to be interested in aspects of the diachronic processes that shape linguistic diversity, including questions concerning any cognitive, environmental or communicative biases that might affect them. In particular, one can look at the rate with which a given linguistic character changes on a phylogeny, with the hope that differences between characters, families or areas might point to substantive linguistic insights, such as differences in "stability" between structural features (Greenhill et al., 2017; Dediu, 2011). However, a complementary way of looking at things is to instead ask what information might give us the stationary distribution of a given feature in a given family in terms of, for example, long-term biases affecting the feature's dynamics. However, stationary distributions come with their own issues, and we explore here, starting from an informal suggestion by Balthasar Bickel and an actual implementation by Chundra Cathart (2023), how we might estimate and use them for two rather different cases. First, we look at colexification patterns where, besides large amounts of missing data, we usually deal with very low frequency phenomena. Second, we look at tone and, in particular, the detection of extra-linguistic biases affecting it. Our presentation here is a snapshot of a very early exploration, where we are still developing, debugging and optimising the methods and code, and where we are still a bit far from actual results. However, we hope to generate useful discussions concerning the advantages and disadvantages of looking at rates and/or the stationary distribution, as well as their actual implementation and interpretation.

Cathcart, Chundra (2023): Rate variation in language change: Toward distributional phylogenetic modeling. In: Karakostis, Fotios Alexandros; Jäger, Gerhard. Biocultural Evolution: An Agenda for Integrative Approaches. Tübingen: Kerns Verlag, 179-202.

Greenhill, S. J., Wu, C.-H., Hua, X., Dunn, M., Levinson, S. C., & Gray, R. D. (2017). Evolutionary dynamics of language systems. Proceedings of the National Academy of Sciences, 114(42), E8822–E8829. https://doi.org/10.1073/pnas.1700388114

Dediu, D. (2011). A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. Proc R Soc B, 278, 474–479. https://doi.org/10.1098/rspb.2010.1595

The Uralic languages: trees and forests

Outi Vesakoski and Michael Dunn (University of Turku; Uppsala University)

TBA

Why modeling space is hard: unidirectional contact, expansion events and other stuff

Matías Guzmán Naranjo (University of Freiburg)

In recent years there has been an increased interest in computational modeling of spatial phenomena in typology (Guzmán Naranjo & Becker 2022, Guzmán Naranjo & Mertner 2023, Hartmann 2022, Hartmann & Jäger 2023, Ranacher et al. 2021, Urban & Moran 2021). While the main focus of most work so far has been on direct language contact, there are two different types of spatial dynamics of interest to typologists and areal linguists: language expansion, and unidirectional contact effects. In this talk I present three new statistical techniques to model expansion and unidirectional contact effects. I illustrate these techniques with a case study on Polynesian phoneme inventory sizes, and show that there is strong evidence for unidirectional effects, but little evidence for expansion effects (contra Atkinson 2011). I also argue that we are still far from having a complete understanding of how to model all spatial dynamics that can affect language contact, and that more attention should be paid to these issues.

Spatial models of linguistic diversity in Africa

Miri Mertner (University of Tübingen)

TBA

A new approach to the diversification of ancient Greek

David Goldstein (University of California)

The diversification of the ancient Greek dialects remains one of the most recalcitrant problems in both Greek and Indo-European linguistics. Debates persist over a number of fundamental issues, including methodology, divergence times, and the topology of the dialects in the second millennium BCE. In this talk, I present the results of a Bayesian phylogenetic analysis of the dialects, which challenges received wisdom regarding the tree topology, the ages of four traditional dialect "groups," and the number of dialects spoken during the Mycenaean period.

Spotting implicational relations among typological features through coupled evolution

Søren Wichmann (Kiel University)

In 2011, Dunn et al. caused a stir by claiming that implicational relations among word-order features are not universal but lineage-specific. This was based on applying methods of identifying correlated evolution implemented in BayesTraits to four language families. It is, however, inherently doubtful that just four language families can provide sufficient evidence for a statistical universal and it is also not clear that a conservative cutoff like the log Bayes Factor (BF) of 5 used by the authors is an adequate criterion for identifying implicational tendencies in linguistic typology. In order to improve on our understanding on what to expect from tests of correlated evolution in typology I have (1) run simulation experiments and (2) tested for correlated evolution among all GramBank features across all language families (carrying out 1,683,436 runs of BayesTraits). The simulations reveal the adequacy with which a certain BF cutoff identifies an implicational relation of a certain strength. For instance, according to the simulation results, a BF cutoff of 5 will adequately identify a 0.8 probability that one trait in a language is coupled with some other trait; a BF cutoff of 2, however, will almost equally adequately identify a 0.7 probability of two traits being coupled. Our choice of a BF cutoff, then, depends on how strong we require implicational relations to be before we accept them as such. The empirical results for GramBank indicate that, across families, a merely positive median BF will recover most of the trivial implications among traits with a small false discovery rate; lowering the median BF cutoff will recover some less trivial implications, but only at the cost of a steeply growing false discovery rate. All in all, it is a subtle matter to interpret the results of phylogenetically-based tests for correlated evolution, and it is likely that it is best to complement such tests with other methods in order to spot implicational relations among typological features.

Navigating linguistic trait space: Analogical modelling of inflectional evolution

Erich Round (Surrey Morphology Group)

TBA