New Advances in
Phylolinguistics
11-12 June 2025

Book of Abstracts (sorted according to the workshop schedule)

Beyond cognacy: Building the lexibank world tree

Gerhard Jäger – Universität Tübingen

State-of-the-art approaches to language phylogenetics rely on expert-annotated cognate sets, which are labor-intensive to construct and restricted to well-studied language families. This limits the scope of phylogenetic inference, particularly when addressing large-scale questions about global language evolution, contact, and typological diversity. In this talk, I present a scalable, fully automated method for inferring language phylogenies directly from lexical data—without relying on cognate annotations.

The method leverages the Lexibank database, comprising aligned wordlists from thousands of languages, and builds on recent advances in character extraction using multiple sequence alignment (MSA).

In the pilot phase, I present a “small world tree” inferred from 928 languages and 110 concepts, spanning diverse families and regions. Evaluation against the Glottolog classification, typological data from Grambank, and measures of phylogenetic signal show that this MSA-based method outperforms both traditional cognate-based inference and automatic cognate clustering pipelines. The results demonstrate that the approach captures genealogical signal within families and enables inference across macro-level relationships—without predefined cognate sets.

Building on these results, I extend the method to generate a “large world tree” comprising approximately 4,400 languages, offering new possibilities for modeling language evolution at a global scale.

Abstract: Toward practical handling of lexical borrowing in computational phylogenetics

Tim Vaughan, ETH Zürich

The borrowing of linguistic features of one language by another is a well-known occurrence in the evolution of language. Just as for analogous non-vertical inheritance processes in biological evolution such as horizontal gene transfer and recombination, borrowing results in the production of non-treelike ancestries among known languages.

Such ancestries pose significant challenges for computational phylogenetics reliant on lexical data, since a) the space of possible phylogenetic networks is infinitely larger than the already large space of phylogenetic trees, and b) the corresponding phylogenetic likelihoods induced by lexical borrowing tend to be highly multi-modal. As a result, while computational methods for phylogenetic inference in the face of lexical borrowing already exist (Neureiter, 2022), they tend to be at best computationally demanding and at worst impractical.

In this talk I will briefly outline two recent directions we have been exploring in order to improve the efficiency and overall practicality of phylogenetic inference in the face of lexical borrowing.

Evolutionary dynamics of syllable complexity

Laura C. Dees

Institute for the Interdisciplinary Study of Language Evolution,
University of Zurich

Linguistic evolution has given rise to variation in the degree of complexity that syllable structure permits. However, linguistic theory has since long proposed that the consonant-vowel (CV) syllable, or at least an asymmetry between strong onsets and weak codas, is a universally preferred (or even optimal) structure of language [1–7]. This claim is partially grounded in findings from acquisition [8], incremental processing [9–11] and the biomechanics of the articulation apparatus [12, 13]. Yet, large-scale surveys do not readily support a universal preference for CV syllables [14, 15] and it remains an open question whether linguistic evolution indeed favors the predicted onset-coda asymmetry.

The current study aims at shedding new light on this question by means of a quantitative phylogenetic approach. Modeling change in syllable structure according to a continuous-time Markov process, we estimate transition rates for maximal onset and coda sizes (in number of consonants). Our manually collected data come from reference grammars of 401 languages spread out over 5 families (Austronesian, Indo-European, Pama-Nyungan, Sino-Tibetan and Uto-Aztecan). Two competing models are compared: one that fits separate CTM processes on individual families, and one with a shared CTM process based on all attested states in each of the 5 families. Comparison is achieved by means of Bayes Factors based on the marginal likelihoods.

Results show that the lineage-specific model clearly outperforms the global model in both onset and coda (both Bayes Factors > 100 [16]). Through further exploration of lineage-specific stationary probabilities (informative of long-term preferences [17, 18]), gain and loss rates and speed of change we characterize the evolutionary dynamics that variably contribute to diversification in this domain of language. Our results imply that lineage-specific linguistic and/or socio-geographic contingencies impact syllable complexity more than any universal cognitive biases or biomechanical constraints on the articulators do – counter to the widespread belief that the CV template is a strong constraint on optimal syllable sizes in grammars.

In the remainder of the talk, we explore the evolution of syllable complexity further by reconsidering the supposed correlation with consonant inventory size [2, 7, 19–21] through the implementation of a phylogenetic correlation model using Brownian motion. Taken together, our results testify to the fact that quantitative phylogenetic methods allow for a more refined perspective on typological distributions and language evolution broadly.

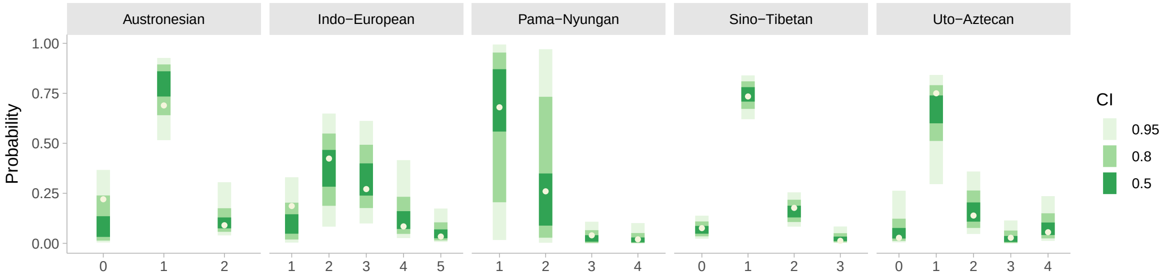

Figure 1: Family-level stationary probability distributions for each of the ascertained maximal coda sizes. Dots represent observed probabilities.

Model misspecification and overestimation of phylogenetic root age in linguistic and biological datasets

Benedict King, Aymeric Hermann, Mary Walworth, Simon Greenhill, Russell Gray

A long-standing debate in evolutionary biology has surrounded molecular clock estimates of clade ages, which are often much older than expected based on the fossil record, a phenomenon known as 'deep root attraction' (DRA). Here we demonstrate a parallel scenario for the linguistic phylogeny of the Oceanic language family. Our analysis, using node calibrations based on archaeological dates, estimates the age of Oceanic to be several centuries older than the archaeological data would suggest. We expanded our study to include the Indo-European languages, another language family with vigorous debate about the age of the root node, and Rhamnaceae, a flowering plant family. All three datasets show a similar pattern of a flat uninformative likelihood profile across different root ages, with the dominance of trees with older roots in the posterior only explainable by the density of old trees in tree space. In addition, posterior predictive simulations reveal some model inadequacy for all datasets. Simulation studies show that our phylogenetic model set-up correctly reconstructs the root age of young trees when data are simulated on trees drawn from the prior predictive distribution. However, simulations using young trees drawn from the posterior distribution, which we have shown to be misspecified, show a significant bias. Here the root age is consistently overestimated and increases as the amount of simulated data analysed increases. It follows that phylogenetic clock models can be biased about the age of the root when the tree model is misspecified, and that surprisingly old estimates of clade age in the literature should be treated with skepticism unless the posterior trees have been shown to overlap with those in the prior and posterior predictive distributions.

Building a model for a phylolinguistic inference of grammaticalization

Gerd Carling, Nour Efrat-Kowalsky (Goethe University, Frankfurt am Main)

Grammaticalization is a well-known principle of language change. The process, first observed by Meillet (Meillet 1912), involves a change in a specific direction: content words of open classes, such as nouns and verbs, gradually lose their semantic content (semantic bleaching) and become function words of closed classes. In the procedure, the form of the words often become weakened (phonological erosion) (McMahon 1994). Due to the importance of grammaticalization in language change, it is considered vital to the reconstruction of principles of language evolution (Heine and Kuteva 2011). However, to set up a quantified model, testing the evolutionary principles of grammaticalization is a challenge, in particular due to the patterns of behaviour of grammaticalization: a grammaticalized item may continue to exist in a language, side by side with its grammaticalized form. In the current presentation, we start by analysing a database of grammaticalization, digitized from the Dictionary of Grammaticalization (Heine and Kuteva 2004). Evidently, nouns (27%) and verbs (34%) are the most common starting points for grammaticalization. In nouns, some semantic classes are overrepresented, most commonly BODY PARTS (42%), PERSON (17%) and PHYSICAL WORLD (17%). Verbal meanings are also restricted. We will focus first on the evolutionary behaviour of the source items, including their substitution rates, propensity for colexification, and borrowability. Further on, we will present a proposal for a model how the procedure of change from content word to function word can be analysed by a phylolinguistic model.

Heine, Bernd, and Tania Kuteva. 2004. World Lexicon of Grammaticalization. Cambridge: Cambridge University Press.

Heine, Bernd, and Tania Kuteva. 2011. "Grammaticalization theory as a tool for reconstructing language evolution." In Oxford Handbook of Language Evolution, edited by Maggie Tallerman Kathleen R. Gibson, 512-527. Oxford: Oxford University Press.

McMahon, April M. S. 1994. Understanding language change. Cambridge: Cambridge Univ. Press.

Meillet, Antoine. 1912. "L'évolution des formes grammaticales." Scientio (Reivsta de scienza) XII(XXVI, 6).

The forest or the supertree: does the world tree make the same predictions about linguistic change as family-specific phylogenies?

Sarah Babinski[1], Alexandru Craevschi[1], Sirui Cheng[2], and Chundra Cathcart[1]

[1]Institute for the Interdisciplinary Study of Language Evolution, University of Zurich
[2]Corpus Christi College, University of Oxford

The global phylogeny inferred by Bouckaert et al. (2022) has opened up the possibility of estimating rates of change across a large number of the world's languages, including isolates and small families. Recent work has begun to use this "world tree" to study change in small families (Jódar-Sánchez & Allassonnière-Tang, 2024), to make claims about the connection between language and societal structure (Chen et al., 2024), and to infer ancestral features of the language of early humans (Her et al., 2024). This GLOBAL procedure departs from the LOCAL procedure proposed by Jäger and Wahle (2021), which infers rates based their joint likelihood for families, smaller taxonomic groups, and isolates (represented as phylogenetically independent elements). However, little work has investigated whether these approaches yield similar results. For example, it is possible that rates inferred using the global phylogeny are fundamentally skewed due to the long branches near the root when compared to shallower, lineage-specific trees. This study puts the global and local approaches to the test. We estimate rates of change for features in Grambank (Skirgård et al., 2023) under the global and local model settings, assessing the extent to which estimates agree. We subsequently estimate rates of change for the same features under a covarion model (which assumes regimes of slow and normal change), comparing the posterior support for rate variation under both model settings (under the rationale that the long branches of the global tree may be more prone to the detection of regimes of slow change). Finally, we select pairs of features (both those for which there are proposed associations as well as those without) in order to investigate whether there are differences in the inferred dynamics of correlated evolution across the two model settings. We evaluate the extent to which model settings agree, propose explanations for points of disagreement between them, and offer some recommendations for cross-family rate inference.

References

Bouckaert, R., Redding, D., Sheehan, O., Kyritsis, T., Gray, R., Jones, K. E., & Atkinson, Q. (2022, July). *Global language diversification is linked to socio-ecology and threat status.* Retrieved 2024-09-24, from `https://osf.io/f8tr6` doi: 10.31235/osf.io/f8tr6

Chen, S., Gil, D., Gaponov, S., Reifegerste, J., Yuditha, T., Tatarinova, T., … Benítez-Burraco, A. (2024, April). Linguistic correlates of societal variation: A quantitative analysis. *PLOS ONE*, *19*(4), e0300838. Retrieved 2025-05-16, from `https://dx.plos.org/10.1371/journal.pone.0300838` doi: 10.1371/journal.pone.0300838

Her, O.-S., Liang, Y.-P., Chan, E., Hsu, H.-H., Hsu, A. C.-P., & Allassonnière-Tang, M. (2024, February). Early humans out of Africa had only base-initial numerals. *Humanities and Social Sciences Communications*, *11*(1), 254. Retrieved 2025-05-16, from `https://www.nature.com/articles/s41599-023-02506-z` doi: 10.1057/s41599-023-02506-z

Jäger, G., & Wahle, J. (2021, July). Phylogenetic Typology. *Frontiers in Psychology*, *12*, 682132. Retrieved 2024-09-18, from `https://www.frontiersin.org/articles/10.3389/fpsyg.2021.682132/full` doi: 10.3389/fpsyg.2021.682132

Jódar-Sánchez, J. A., & Allassonnière-Tang, M. (2024, September). The evolutionary dynamics of grammatical gender in Torricelli languages. *STUF - Language Typology and Universals*, *77*(3), 353–369. Retrieved 2025-05-16, from `https://www.degruyter.com/document/doi/10.1515/stuf-2024-2010/html` doi: 10.1515/stuf-2024-2010

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., … Gray, R. D. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, *9*(16). doi: 10.1126/sciadv.adg6175

Mosaic versus concerted evolution in grammatical systems

Angela Chira

Mosaic evolution describes how different modules of a system can show decoupled and diverse evolutionary dynamics. By breaking the integration of systems, mosaic evolution leads to remarkable diversification of forms at macro-evolutionary scales. Language shows great potential for mosaicism, given its multidimensional structure and the potential for its dimensions to change under different selection pressures. However, mosaicism in language is poorly understood.

Here, we investigate this issue with a focus on grammar. We use Bayesian phylogenetic models to map rates of evolution across four major grammatical dimensions: word order, locus of marking, fusion, and informativity. Our results reveal frequent decoupling of evolutionary trajectories across these domains. Specifically, deviations from background rates of change are often weakly correlated across grammatical dimensions. The magnitudes of exceptional change episodes also frequently vary between dimensions. These findings show that mosaic evolution is common and likely played a sizable role in the evolution of grammatical disparity.

Punctuated Evolution, and the Social Dynamics of Language Diversification

Thomas Paul Roberts, Paul Widmer, John Mansfield

ISLE, University of Zurich

Language contact is primarily assumed to lead to convergence across features and domains. More recently however, linguistic divergence has been observed as a frequent consequence of contact, potentially driven by our innate tendency to exaggerate social and cultural differences to solidify group identity (termed 'schismogenesis' after Bateson 1935). While fieldwork studies provide examples, there have been few quantitative studies providing a more explicit, systematic cross-linguistic account of schismogenesis.

Greenhill et al. 2017 proposed that the punctuated bursts of change previously observed in the phylogenies of multiple language families (Atkinson et al. 2008) could be a result of schismogenesis during language diversification. This would imply that while a group is splitting, the selection of contrasting linguistic variants (e.g. lexemes) to mark group identity produces a sudden accelerated 'burst' of change, differing from the usual mechanisms in gradual evolution.

To test this hypothesis, we investigate whether social dynamics between (sub-)groups at the split event correlate with the level of inferred punctuated change. We use recent phylogenetic models (Douglas et al. 2024) to quantify independent amounts of punctuated evolution in the lexicon at language splits. We use social predictors, including levels of contact between (sub-)groups and the spread of groups over geographic areas, to test for systematic relationships between punctuated change and the dynamics of the contact situation at the split.

In this workshop, we present preliminary findings suggesting a statistically significant relationship between level of social contact and punctuated change.

Atkinson, Quentin D. et al. (2008). "Languages Evolve in Punctuational Bursts". In: *Science* 319.5863 (Feb. 2008), pp. 588–588. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.1149683`.

Bateson, Gregory (1935). "Culture Contact and Schismogenesis". In: *Man* 35 (Dec. 1935), p. 178. ISSN: 00251496. DOI: `10.2307/2789408`. JSTOR: `2789408`.

Douglas, Jordan et al. (2024). *Evolution Is Coupled with Branching across Many Granularities of Life*. Sept. 2024. DOI: `10.1101/2024.09.08.611933`.

Greenhill, Simon J. et al. (2017). "Evolutionary Dynamics of Language Systems". In: *Proceedings of the National Academy of Sciences* 114.42 (Oct. 2017). ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.1700388114`.

Tracing the Evolution of Pano Languages in Parallel with Archaeological Changes in the Ucayali Basin

Frederic Blum, Pilar Valenzuela, Roberto Zariquiey

Panoan is one of the largest language families in Western Amazonia, with approximately thirty-three languages spoken by around 65,000 individuals (Fleck 2013). Despite its size and potential significance for understanding the regional history, many debates persist regarding its origin, expansion, and internal classification. Following Myers (1976), the first evidence for presence of Pano-speaking groups in the Ucayali basin is manifested through the arrival of the Pacacocha culture around 1500 BP. Pacacocha would then have been followed up by Cumancaya, for which Lathrap establishes a cultural continuity to the present-day Shipibo-Konibo. From the linguistic point-of-view, however, the results are not as clear. The hypotheses for the origin of the Pano language family range from a rather recent origin ~1000 years ago to origins at a larger time depth 3000-5000 years ago (Fleck 2013, Valenzuela & Guillaume 2017). However, a definitive dating of the time-depth of the Pano family remains undetermined. The genealogical relationship between Pano and Takanan languages further complicates the discussion by contributing different hypotheses about a possible homeland of Proto-Pano-Takanan. Scholars have proposed a Northern, Southern, and Central origin hypothesis, all with different implications for the analysis of the different archaeological traditions. By integrating the linguistic history of Pano with relevant archeological information, the study aims to uncover more insights into the complex history of the Ucayali basin.

To answer some of these questions, this study brings a fresh perspective by computing a Bayesian linguistic phylogeny. This method allows for the reconstruction of historical relationships between languages based on lexical data (Greenhill et al. 2016). For this, we have created a database with a basic vocabulary list of 181 concepts for 26 languages and annotated the data for cognacy using EDICTOR (List, Blum, van Dam 2025). In order to date the phylogeny in time, we have used anthropological sources as calibration points for a relaxed molecular clock. All results are reported with uncertainty intervals in order to reflect the statistical probability. We further verify the resulting phylogeny with relevant linguistic innovations shared by subgroups of languages. Combining both approaches, we present a new internal classification that is anchored in time.

The results suggest that the original diversification of the Pano language family occurred between 1000 and 1800 years ago. Without using the archaeological data as calibration points, those results support a possible link between the arrival of Pano languages in the Ucayali basin along the Pacacocha culture and provide evidence against a larger time depth. Further splits of the language family can be mapped to other changes in archaeological traditions, with the breakup of the largest group of Pano languages (Central-Southern) happening in parallel to the occupation and raids by Kukama-speaking groups. The proposed findings not only contribute to a better understanding of the linguistic and historical aspects of the Pano language family, but also open up the opportunity for further examination of its relationship with archaeological evidence and cultural transformations in the region and beyond.

References

Fleck, D. W. (2013). Panoan languages and linguistics. *Anthropological Papers of the American Museum of Natural History*, *99*. https://doi.org/http://dx.doi.org/10.5531/sp.anth.0099

Greenhill, S. J., Heggarty, P., & Gray, R. D. (2020). Bayesian Phylolinguistics. In B. S. V. Richard D. Janda Brian D. Joseph (Ed.), *The Handbook of Historical Linguistics* (pp. 226–253). Wiley. https://doi.org/10.1002/9781118732168.ch11

Lathrap, D. W. (1970). *The upper Amazon* (Vol. 70, p. 256). London: Thames and Hudson.

List, J.-M, Blum, F. and van Dam, K. P. (2025): EDICTOR 3: A Web-Based Tool for Computer-Assisted Language Comparison [Software Tool, Version 3.1]. MCL Chair at the University of Passau: Passau. URL: https://edictor.org/.

Myers, T. P. (1976). Isolation and ceramic change: A case from the Ucayali River, Peru. *World Archaeology*, *7*(3), 333–351. https://doi.org/10.1080/00438243.1976.9979645

Valenzuela, P., & Guillaume, A. (2017). Estudios sincrónicos y diacrónicos sobre lenguas Pano y Takana: una introducción. *Amerindia*, *39*(1), 1–49. https://hal.science/hal-01920483

Taiwan as a lingustic area, or how not to get catfished by Bayesian phylogeny results

Isaac Stead

Interpreting the results of a Bayesian phylolinguistic analysis is never as simple as just glancing at the consensus tree and Densitree. In this talk, I demonstrate that potentially alluring subgroupings in results can end up ghosting you, and conversely that an unpromising result can hide more than you thought about your research question. I will present the results of a Bayesian analysis of Austronesian primary-level subgrouping, using BEAST2 (Bouckaert et al. 2018) and data from the Austronesian Basic Vocabulary Database (Greenhill et al. 2008). I will show you how fraught the interpretation of subgroupings given by the model can be and that apparently solid subgrouping results from basic vocabulary can actually be the result of a tangle of contact processes.

The primary-level subgrouping of the Austronesian language family has been and continues to be controversial, and many subgrouping hypotheses have been proposed, based on different types of evidence (e.g. Blust 1999; Ho & Yang 2000; Ross 2008; Sagart 2014; Chen 2017; Chen et al. 2023 among many others.) The results of a Bayesian analysis show that this controversy has obscured what is perhaps more interesting: that Taiwan has been operating as a sprachbund for around 5,000 years and that even within basic vocabulary there is overlooked but tantalising evidence of this unwritten history of contact and change.

Some useful techniques I'll demonstrate will include increasing the interpretability of results by doing ancestral state estimation directly in BEAST2, different methods of reasoning about the entire posterior sample instead of just the consensus tree, the importance of competing standards and types of evidence in historical linguistics, using the power of CLDF and Glottolog to do clustering and other kinds of analyses on the results, and the importance of qualitative data to understanding quantitative results.

Testing linguistic universals with synchronic spatiophylogenetic and diachronic models

Annemarie Verkerk, Olena Shcherbakova, Hannah J. Haynie, Hedvig Skirgård, Christoph Rzymski, Quentin D. Atkinson, Simon J. Greenhill & Russell D. Gray

The 7,500 extant human languages are remarkably varied but this diversity is not without limits. The nature and extent of these limits are debated. Part of this debate has focused on so-called universals – strong patterns in the co-occurrence of grammatical features. Here we use a comprehensive new global grammatical database – Grambank – to test 191 putative universals. We test these using Bayesian methods while controlling for spatial and phylogenetic non-independence. We find statistical support for 31% of the proposed linguistic universals. The majority of these concern word order and hierarchical universals: two types that have featured prominently in earlier work. We study the rates of change estimated in the phylogenetic models and find evidence of preferred change towards predicted feature configurations. This suggests that, despite the enormous combinatorial flexibility of language systems, shared cognitive and communicative pressures mean that languages repeatedly evolve toward the same preferred regions of design space.

Extrinsic drivers of language diversification

Ewan Ciuffi

Extrinsic factors such as ecological and demographical covariates are hypothesised to play a central role in the diversification of languages. This project investigates whether variables like population density and ecological factors can predict diversification rates across space and time in the Indo-European language family. Using a Bayesian phylodynamic framework, diversification rates are modelled as functions of predictors, providing a quantitative assessment of the impact of extra-linguistic factors on language evolution.

Interpolating the spatio-temporal distribution of languages using ancient genomic data

Clemens Schmid, Joscha Gretzinger, Stephan Schiffels

Mounting evidence points to a robust correlation between genetic and linguistic distance, especially among modern speakers of Indo-European languages (Barbieri et al. 2022). This may be the result of a shared mechanism of intergenerational vertical transmission for genes and languages, suggesting an intertwined evolutionary history. Building on this premise, we developed an experimental predictive model for diachronic, spatial language distribution in Europe, leveraging ancient genomic data as a proxy.

Using statistical dimensionality reduction, we first constructed two independent Euclidean spaces for genomic (genotypes in the AADR dataset, Mallick et al. 2024) and linguistic diversity (cognate-sets in the IE-CoR dataset, Heggarty et al. 2023). Linking these spaces then enabled a two-step interpolation from geo-temporal to genetic, and subsequently from genetic to linguistic space. The resulting model successfully reconstructs major historical linguistic expansions over the last 3000 years (e.g. Germanic, Slavic), but also faces limitations due to sparse ancient genomic sampling, incomplete data on extinct languages (e.g. Continental Celtic), and the restriction to only one language family.

Barbieri, Chiara et al. (2022). “A global analysis of matches and mismatches between human genetic and linguistic histories”. In: Proceedings of the National Academy of Sciences 119.47. doi: 10.1073/pnas.2122084119.

Mallick, Swapan et al. (2024). “The Allen Ancient DNA Resource (AADR) - A curated compendium of ancient human genomes”. In: Scientific Data 11.1. doi:10.1038/s41597-024-03031-7.

Heggarty, Paul et al. (2023). “Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages”. In: Science 381.6656. doi:10.1126/science.abg0818.

Who Did What to Whom - and Why? Untangling typological correlations from the perspective of communicative efficiency and token-based typology

Natalia Levshina

Previous research in typology and psycholinguistics has revealed numerous pairwise correlations among the core properties of simple transitive clauses—such as the fixed order of agent (A) and patient (P), case marking patterns, verb position—and broader factors including speaker population size and cultural reliance on contextual cues. Although these correlations have been studied for decades, there remains no clear consensus on the strength of the evidence or the underlying causal mechanisms.

In this talk, I will present a comprehensive inventory of these correlations and critically assess the cross-linguistic and experimental evidence supporting them. Drawing on data from the Leipzig Corpora Collection and the Parallel Bible Corpus, I will apply causal inference techniques and hierarchical Bayesian regression models incorporating phylogenetic and spatial random effects. Based on this analysis, I will propose several hypotheses about the underlying relationships among these variables, framed within a communicative efficiency perspective, with particular attention to the optimization of processing costs.

Phylogenetic analysis with concerted changes

Takuya Takahashi (University of Zurich)

Concerted evolution is a specific process in biological evolution where parallel changes occur at multiple gene copies due to genetic exchange among related genes. Similarly, linguistic evolution often features simultaneous changes across multiple linguistic features. Examples include regular sound changes that affect all words containing a particular sound, and innovations in inflectional morphology that alter the conjugation patterns of multiple verbs within an inflection class. Standard phylogenetic models cannot treat such simultaneous changes as they typically assume that every feature, or sequence site, evolves independently of each other. In this talk, I will introduce a new Bayesian phylogenetic model designed to infer evolutionary trees from sequence alignments that include concerted changes. In this model, each sequence site represents a state (or property) of a word or cognate, and each language possesses its own set of possible states. The model accounts for three types of evolutionary change. The first type, *sporadic substitution*, involves a change in the state of a single site. The second, *concerted substitution*, occurs when all occurrences of one state are replaced with another across the sequence. The third type, *concerted merger*, involves replacing all instances of a state with another state that already exists in the language, thereby reducing the total number of distinct states. I will briefly discuss the potential application of this model to the pitch-accent features of Japanese dialects.

Semantics drives analogical change in Germanic strong verb paradigms: a phylogenetic study

Alexandru Craevschi, Sarah Babinski, and Chundra Cathcart

Institute for the Interdisciplinary Study of Language Evolution
University of Zurich

Abstract

A large body of research on morphological paradigms makes the prediction that irregular morphological patterns of allomorphy are more likely to emerge and persist when they serve to mark important functional distinctions. More specifically, it has been observed that in some Germanic languages in which narrative past tense is expressed by the past participle, there is a greater affinity for stem allomorphy shared by preterite forms and past participles to the exclusion of present forms (the so-called ABB pattern), as it serves to enhance marking of the binary semantic opposition between present and past. Using data from 107 cognate verbs attested across 14 archaic and contemporary Germanic languages and a novel hierarchical phylogenetic model, we show that there is a greater long-term preference for this alternation pattern in situations where narrative past tense has been extended to the past participle, confirming this hypothesis. We further elucidate the mechanisms underlying this association, demonstrating that this association holds because verbs with the ABB pattern are more likely to preserve it in situations where it marks an important binary semantic opposition; however, there is less evidence that the ABB pattern is extended to verbs with different patterns under the same circumstances. These results bear on debate as to whether the distribution of irregularity we observe cross-linguistically is due primarily to (1) the preservation of irregular patterns or (2) an active drive toward irregularization in certain contexts, and are more in line with the first hypothesis.

Keywords: phylogenetic comparative methods, morphological change, analogical change, paradigm leveling, Germanic