University of Zurich[UZH]

University Research Priority Program
Language and Space

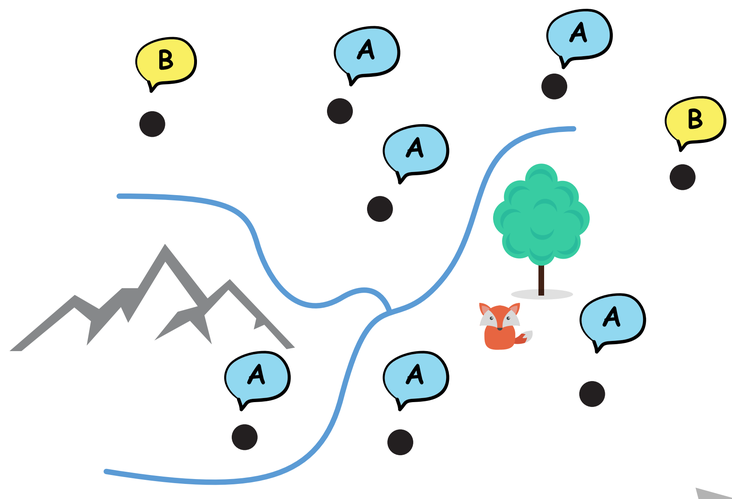
Finding areas of language contact in space

P. Ranacher[1], N. Neureiter[1], R. van Gijn [2]
[1]URPP Language and Space Lab, UZH, [2] FFG Areal Morphology, UZH

Introduction

How can we explain that languages share specific features?

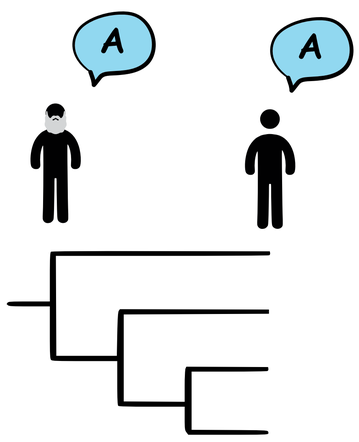
The answer to this seemingly innocuous question is not trivial. Language evolution is a complex process with little ground truth, which makes historical reconstruction difficult. However, we can put forward three simple explanations that might shed light on why particular languages happen to have particular features.



The dots ● in the image to the left represent languages in space.

The speech bubbles A B represent different variants of a linguistic feature. A feature is a structural (grammatical, phonological, lexical) property of a language.

INHERITANCE

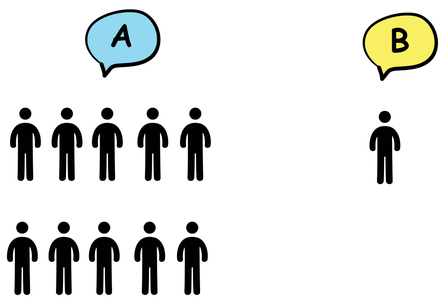


The feature was inherited within the language. Recently, several quantitative phylogenies have been published [1, 2, 3], which allow to estimate

P(inherited)

- the probability that a feature was inherited from related ancestral languages.

GLOBAL PREFERENCE

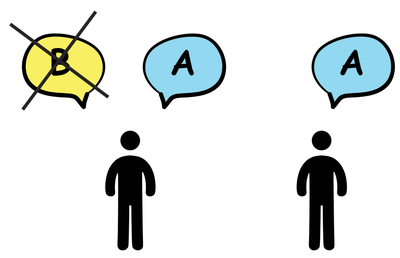


The feature is globally preferred over others. Large databases of language properties [4, 5] help us to estimate

P(global)

- the probability that a feature is present in a language due to global preference.

CONTACT



The feature was adopted from neighboring languages. Speakers interact and influence each other. (Thou shall not covet thy neighbor's features, one would assume. But not a thing of it!) Until recently, only few quantitative approaches have been put forward to reconstruct language contact in space [6]. In this study we randomly propose potential contact regions and estimate

P(contact)

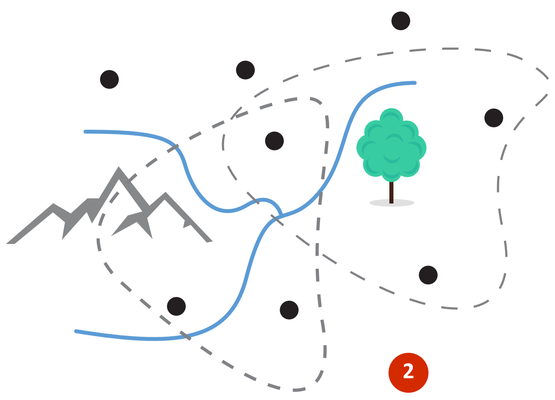
- the probability that features were passed on through contact.

Bayesian Model

We can model the above insights as a Bayesian mixture model. The probability to observe a language to have a particular feature is a weighted sum of the probability that the feature was inherited, globally preferred or adopted through contact.

$$P(A) = w_1 \cdot P(\text{inherited}) + w_2 \cdot P(\text{global}) + w_3 \cdot P(\text{contact})$$

The probability of inheritance and global preference are known, whereas the influence of contact and the respective weights w_1, w_2, w_3 are unknown.



1 Propose initial sample (contact area with weights) and evaluate likelihood.

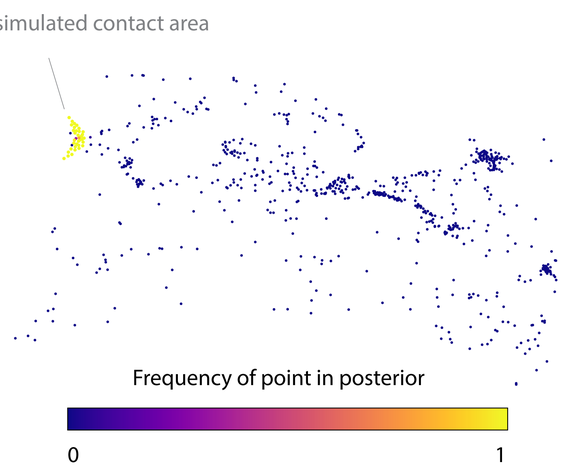
2 Propose new sample. Evaluate likelihood and compare to previous sample. Accept with metropolis hastings acceptance probability.

3 Repeat 2 many, many times.

We use this model as a likelihood function in a Markov Chain Monte Carlo (MCMC) algorithm. We randomly propose samples - contact areas in space and weights - and evaluate their likelihood (see left). Put simply the likelihood function assigns high values to samples that propose contact areas where many features are similar, while accounting for the influence of inheritance and global preference.

The MCMC generates a posterior distribution of contact areas in space.

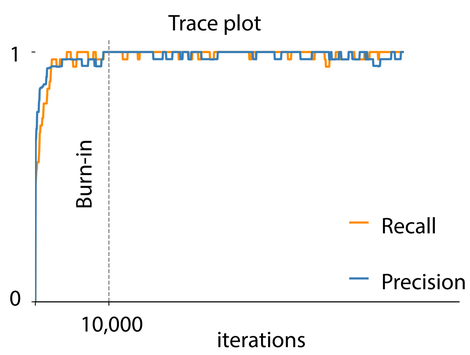
Experiments



Currently, we perform intensive simulations to test and verify the model. We distribute languages in space and simulate linguistic features and contact areas. Then, we run the algorithm and try to identify the simulated contact areas (see left).

The figure below reports the precision and recall of the experiment. After around 10,000 iterations the MCMC converges and almost perfectly identifies the simulated contact areas.

Once testing is completed, we will explore the algorithm on real world data in South America where language contact, together with inheritance, is hypothesized to explain much of the current linguistic variation.



References: [1] Sagart, Laurent, et al. "Dated language phylogenies shed light on the ancestry of Sino-Tibetan." Proceedings of the National Academy of Sciences (2019): 201817972. [2] Bouckaert, Remco R., Claire Bowern, and Quentin D. Atkinson. "The origin and expansion of Pama–Nyungan languages across Australia." Nature ecology & evolution (2018): 1. [3]: Chang, Will, et al. "Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis." Language 91.1 (2015): 194-244. [4] Dryer, Matthew S., and Martin Haspelmath. "The world atlas of language structures online." (2013). [5] Nichols, Johanna, Alena Witzlack-Makarevich, and Balthasar Bickel. "The AUTOTYP genealogy and geography database: 2013 release." Zurich: University of Zurich (2013). [6]: Daumé III, Hal. "Non-parametric Bayesian areal linguistics." arXiv preprint arXiv:0906.5114 (2009).