Phylogenetic Typology

Gerhard Jäger

Tübingen University

Bochum Language Colloquium

April 27, 2021
• common practice since Greenberg (1963):
  • collect a sample of languages
  • classify them according to some typological feature
    ⇒ skewed distribution indicates something interesting going on

• Problem: languages are not independent samples
• skewed distribution may reflect
  • skewed diversification rate across families
  • properties of an ancestral bottleneck
• balanced sampling mitigates the first, but not the second problem
Maslova (2000):

“If the A-distribution for a given typology cannot be assumed to be stationary, a distributional universal cannot be discovered on the basis of purely synchronic statistical data.”

“In this case, the only way to discover a distributional universal is to estimate transition probabilities and as it were to ‘predict’ the stationary distribution on the basis of the equations in (1).”
Distribution of verb-object/object verb vs. noun-relative clause/relative clause-noun
this study:
- word-order data from WALS
- 1,060 languages
- 94 families + 81 isolates = 175 lineages
Steps of (Bayesian) model validation

- exploratory data exploration → descriptive statistics
- specification of (a) generative probabilistic model(s)
- prior predictive simulation
- model fitting
- posterior predictive simulation
- model comparison

(cf., eg., Gelman et al. 2014)
Descriptive statistics

- Each language can be represented as a binary vector over 4 variables (for the four combinations of OV/VO and NRc/RcN).
- The **total variance** is the sum of the variance of those for binary variables.
- The **mean lineage-wise variance** is the average total variance per lineage.
- The **between-family variance** is the total variance between the centroids for each family.
Descriptive statistics
Defining models

- feature values evolve according to a *continuous time Markov chain* (CTMC)
- evolution along a phylogeny
- phylogenetic tree is only partially known - represented here as posterior distribution of Bayesian phylogenetic inference from lexical data (from ASJP)
Figure: Schematic structure of the phylogenetic CTMC model. Independent but identical instances of a CTMC run on the branches of a phylogeny.
Figure: a. CTMC b. Equilibrium distribution c. Fully specified history of a phylogenetic Markov chain d. Marginalizing over events at branches e. Marginalizing over states at internal nodes
Figure: Phylogenetic Markov CTMC with a collection of phylogenies
Figure: CTMC for a possibly correlated feature pair
Figure: Universal vs. lineage-specific model
From words to trees

Swadesh lists

sound similarities

word alignments

cognate classes

character matrix

phylogenetic tree

training pair-Hidden Markov Model

applying pair-Hidden Markov Model

classification/ clustering

feature extraction

Bayesian phylogenetic inference
From words to trees

- Swadesh lists
- Sound similarities
- Word alignments
- Cognate classes
- Character matrix
- Phylogenetic tree

Training pair-Hidden Markov Model

Applying pair-Hidden Markov Model

Classification/clustering

Feature extraction

Bayesian phylogenetic inference

<table>
<thead>
<tr>
<th>concept</th>
<th>Latin</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>ego</td>
<td>Ei</td>
</tr>
<tr>
<td>you</td>
<td>tu</td>
<td>yu</td>
</tr>
<tr>
<td>we</td>
<td>nos</td>
<td>wi</td>
</tr>
<tr>
<td>one</td>
<td>unus</td>
<td>w3n</td>
</tr>
<tr>
<td>two</td>
<td>duo</td>
<td>tu</td>
</tr>
<tr>
<td>person</td>
<td>persona, homo</td>
<td>pers3n</td>
</tr>
<tr>
<td>fish</td>
<td>piskis</td>
<td>fiS</td>
</tr>
<tr>
<td>dog</td>
<td>kanis</td>
<td>dag</td>
</tr>
<tr>
<td>louse</td>
<td>pedikulus</td>
<td>laus</td>
</tr>
<tr>
<td>tree</td>
<td>arbor</td>
<td>tri</td>
</tr>
<tr>
<td>leaf</td>
<td>foly~u*</td>
<td>lif</td>
</tr>
<tr>
<td>skin</td>
<td>kutis</td>
<td>skin</td>
</tr>
<tr>
<td>blood</td>
<td>saNgw~is</td>
<td>bl3d</td>
</tr>
<tr>
<td>bone</td>
<td>os</td>
<td>bon</td>
</tr>
<tr>
<td>horn</td>
<td>kornu</td>
<td>horn</td>
</tr>
<tr>
<td>ear</td>
<td>auris</td>
<td>ir</td>
</tr>
<tr>
<td>eye</td>
<td>okulus</td>
<td>Ei</td>
</tr>
</tbody>
</table>
From words to trees

- Swadesh lists
- Sound similarities
- Word alignments
- Cognate classes
- Character matrix
- Phylogenetic tree

- Training
- Applying pair-Hidden Markov Model
- Classification/clustering
- Feature extraction
- Bayesian phylogenetic inference

Swadesh lists

[Diagram of a network]
From words to trees

<table>
<thead>
<tr>
<th>Language</th>
<th>fish:ə</th>
<th>tongue</th>
<th>smokeːt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abui-Atangmelang</td>
<td>-af-u</td>
<td>tal-i-fi-</td>
<td>aum-::-b-a-a-o-o-7o-</td>
</tr>
<tr>
<td>Abui-Fuimelang</td>
<td>-af-u</td>
<td>tal-E-b-</td>
<td>aum-::-b-a-a-o-o-7o-</td>
</tr>
<tr>
<td>Blagar-Bakalog</td>
<td>-ab-</td>
<td>-j-e-bur-</td>
<td>-ad-::-b-a-a-(a)-ka-</td>
</tr>
<tr>
<td>Blagar-Bama</td>
<td>-ab-</td>
<td>teg-e-bur-</td>
<td>-bad-::-b-a-a-(a)-ka-</td>
</tr>
<tr>
<td>Blagar-Kulijahi</td>
<td>-ab-</td>
<td>tej-e-bur-</td>
<td>-bad-::-b-a-a-(a)-ka-</td>
</tr>
<tr>
<td>Blagar-Nule</td>
<td>-ab-</td>
<td>tej-e-bur-</td>
<td>-ad-::-b-a-a-(a)-ka-</td>
</tr>
<tr>
<td>Blagar-Tuntuli</td>
<td>-ab-</td>
<td>tej-e-bur-</td>
<td>a-ad(a)-(a)-(a)-q-</td>
</tr>
<tr>
<td>Blagar-Warsalelang</td>
<td>-ab-</td>
<td>tel-e-bur-</td>
<td>a-ad-::-b-a-a-(a)-x-</td>
</tr>
<tr>
<td>Bunaq</td>
<td>-ab-</td>
<td>-</td>
<td>-bad-::-b-a-a-(a)-x-</td>
</tr>
<tr>
<td>Deing</td>
<td>haf-</td>
<td></td>
<td>-bad-::-b-a-a-(a)-x-</td>
</tr>
<tr>
<td>Hamap</td>
<td>7ab-</td>
<td>mar-P-hui-</td>
<td>-bad-::-b-a-a-(a)-x-</td>
</tr>
<tr>
<td>Kabola</td>
<td>hab-</td>
<td>tal-e-b-</td>
<td>aval-::-b-e-a-e-7o-</td>
</tr>
<tr>
<td>Kaera-Padansul</td>
<td>-ab-</td>
<td>tale-e-b-</td>
<td>a-ad-::-b-e-a-a-x-</td>
</tr>
<tr>
<td>Kafoa</td>
<td>-af(u)</td>
<td>tal-i-p-</td>
<td>-fo-an-</td>
</tr>
<tr>
<td>Kamang</td>
<td>-ap-i</td>
<td>nai-::-ju-</td>
<td>-p-a-s-</td>
</tr>
<tr>
<td>Kiraman</td>
<td>-bo-</td>
<td>nai-i-ber-</td>
<td>-ar-::-b-a-a-o-k-</td>
</tr>
<tr>
<td>Klon</td>
<td>-ob-</td>
<td>gel-E-b-</td>
<td>-ed-::-b-o-o-o-</td>
</tr>
<tr>
<td>Kui</td>
<td>-eb-</td>
<td>tal-i-ber-</td>
<td>-ar-::-b-o-o-o-k-</td>
</tr>
<tr>
<td>Kula</td>
<td>-ap-i</td>
<td>-il-I-p-</td>
<td>-p-a-s-</td>
</tr>
<tr>
<td>Nedebang</td>
<td>aaf-i</td>
<td>gel-e-fu-</td>
<td>-ar-::-b-u-u-</td>
</tr>
<tr>
<td>Reta</td>
<td>aaf-</td>
<td>nai-e-bul-</td>
<td>a-md-::-b-o-o-o-</td>
</tr>
</tbody>
</table>
| Sar-Adiabang     | haf-   | -p-e-fal- | -ar-::-b-u-
| Sar-Nule         | haf-   | nai-e-fai- | -|
| Sawila           | -ap-i  | gal-impur- | -p-e-a-a-ka-|
| Telwa-Madar      | xaf-   | gel-i-yi-  | -bun-|
| Wersing          | -ap-i  | nej-e-bur- | nd-p-e-a-a-k-|
| Wpantar          | haf-   | nai-e-bu-  | -b-un-|

Swadesh lists
sound similarities
word alignments
cognate classes
character matrix
phylogenetic tree

Bayesian
phylogenetic
inference

feature extraction
classification/
clustering
applying
pair-Hidden Markov Model

training
pair-Hidden Markov Model
Swadesh lists

sound similarities

word alignments

cognate classes

character matrix

phylogenetic tree

training pair-Hidden Markov Model

applying pair-Hidden Markov Model

classification/ clustering

feature extraction

Bayesian phylogenetic inference

From words to trees

TNG. ENGAN. MAIBI
TNG. ENGAN. POLE
TNG. ENGAN. SAU
TNG. ENGAN. YARIBA
TNG. FASU. FASU
TNG. FASU. NAMASI
TNG. FINISTERRE-HUON. AMARA
TNG. FINISTERRE-HUON. BORONG
TNG. FINISTERRE-HUON. BURUM
TNG. FINISTERRE-HUON. BURUM MIND
TNG. FINISTERRE-HUON. DEDUA
TNG. FINISTERRE-HUON. HUBE
TNG. FINISTERRE-HUON. KATE
TNG. FINISTERRE-HUON. KOMBA
TNG. FINISTERRE-HUON. KOYCORONG
TNG. FINISTERRE-HUON. MAPE
TNG. FINISTERRE-HUON. MAPE 2
TNG. FINISTERRE-HUON. MIGABAC
TNG. FINISTERRE-HUON. MINDIK
TNG. FINISTERRE-HUON. MONOLILI
TNG. FINISTERRE-HUON. NABAK
TNG. FINISTERRE-HUON. NANKINA
TNG. FINISTERRE-HUON. NEK
TNG. FINISTERRE-HUON. NUKNA
TNG. FINISTERRE-HUON. SELEPET
TNG. FINISTERRE-HUON. TIMBE
TNG. FINISTERRE-HUON. TOBO
TNG. FINISTERRE-HUON. WANTOAT
TNG. FINISTERRE-HUON. YOPNO
TNG. GOILALAN. AFDA
TNG. GOILALAN. KUNIMAIPA
TNG. GOILALAN. MAFULU
From words to trees

Swadesh lists

sound similarities

word alignments

cognate classes

character matrix

feature extraction

Bayesian phylogenetic inference

training pair-Hidden Markov Model

applying pair-Hidden Markov Model

classification/clustering

Phylogenetic tree
Workflow

(data from all 94 families in database; ca. 1,060 languages in total)

- estimate posterior tree distributions with MrBayes for each family, using Glottolog as constraint tree
- estimate transition rates
- estimate stationary distribution of major word order categories
• all models use the same prior for rates:

\[ \text{rate}_i \sim \text{LogNormal}(0, 1) \]

• universal models: one set of rates across lineages
• lineage-dependent models: different set of rates for each lineage
• dependent features model: 8 rates per set
• independent features model: 4 rates per set
universal model

lineage-specific model

Figure: Prior predictive simulations
Run MCMC to infer posterior distribution

- here: done with Johannes Wahle’s *Julia* package *Julia_Tree*
- based on *Mamba* (https://mambajl.readthedocs.io/en/latest/)
- currently under submission
- If you want to give it a try yourself, get in touch with Johannes
• use parameters from posterior sample
• simulate mock data using these parameters
Figure: Posterior predictive simulations: total variance. Horizontal lines indicate the empirical value. The thick vertical lines show the 50% highest-density intervals and the thin lines the 95% highest-density intervals of the posterior predictive distributions.
Figure: Posterior equilibrium probabilities and linear regression
Figure: Correlation coefficients for feature pairs. White dots indicate the median, thick lines the 50% and thin lines the 95% HPD intervals.
Figure: Feature-pairs with credible evidence for a correlation.
• All these techniques assess the **predictive performance** of models
• A good predictive model may be a poor scientific model though.
• Good predictive performance is a necessary but not a sufficient condition for model evaluation.


