Continuous Space Representations of Linguistic Typology and their Application to Phylogenetic Inference

Yugo Murawaki (Kyushu University, Japan)

Background

Computational approaches to phylogenetic inference

- Huge success in the last decade
  - Indo-European
  - Austronesian
  - Bantu, etc
- Previous studies are cognate-based, using either
  1. regular sound changes, or
  2. the rates of birth & death of cognates
- Only applicable to known language families
  - Known because they share cognates!
- *Language isolates* lack cognates to be compared cross-linguistically
  - Ainu
  - Basque
  - Japanese

Linguistic typology as the last hope

- An arbitrary pair of languages can be compared

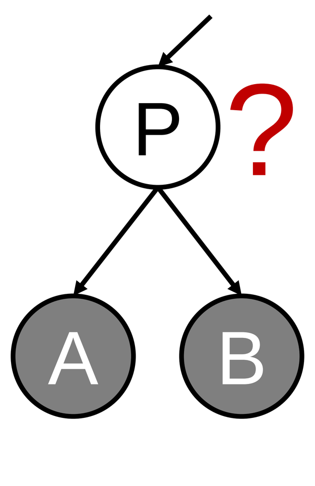
Feature 81A
Order of SOV
- 0: SOV
- 1: SVO
- 2: VSO
- ⋮

Japanese:	1	1	2	⋯	0	4
Korean:	2	2	1	⋯	0	4
Ainu:	0	1	1	⋯	0	4

- Generally much more stable than cognates
  - Possibly even on the order of 10,000 years

What can be said about the ancestor?

- We know too little about how typological features change over time
- The ancestor would be close to its descendants
- The ancestor must be a *typologically natural* language



- Previous phylogenetic models assume *independence* of features
- Typological studies show they are *interdependent*
  - Object-Verb implies Adjective-Noun [Greenberg, 1963]
- A naïve application of phylogenetic models leads to the reconstruction of unnatural ancestors

Continuous space representations that capture typological naturalness

Typological naturalness

$$P(\mathrm{x}) = \frac{\exp(f(h))}{\sum_{\mathrm{x}'} \exp(f(\mathrm{h}'))}$$

v': 2 | 2 | 0 | ⋯ | 3 | 3

⇧ Debinarize the binary vector

x'': 0 | 0 | 1 | 0 | 0 | 0 | ⋯ | 0

⇧ Binarize accordoing to categorical constraints

x': 0.01 | 0.00 | 0.92 | 0.02 | 0.01 | 0.01 | ⋯ | 0.00

⇧ Decode (with errors) $\mathrm{x}' = \sigma(\mathrm{W}_\mathrm{d}\mathrm{h} + \mathrm{b}_\mathrm{d})$

h: 0.21 | 0.84 | ⋯ | 0.03

⇧ Encode $\mathrm{h} = \sigma(\mathrm{W}_\mathrm{e}\mathrm{x} + \mathrm{b}_\mathrm{e})$

x: 0 | 0 | 1 | 0 | 0 | 0 | ⋯ | 0

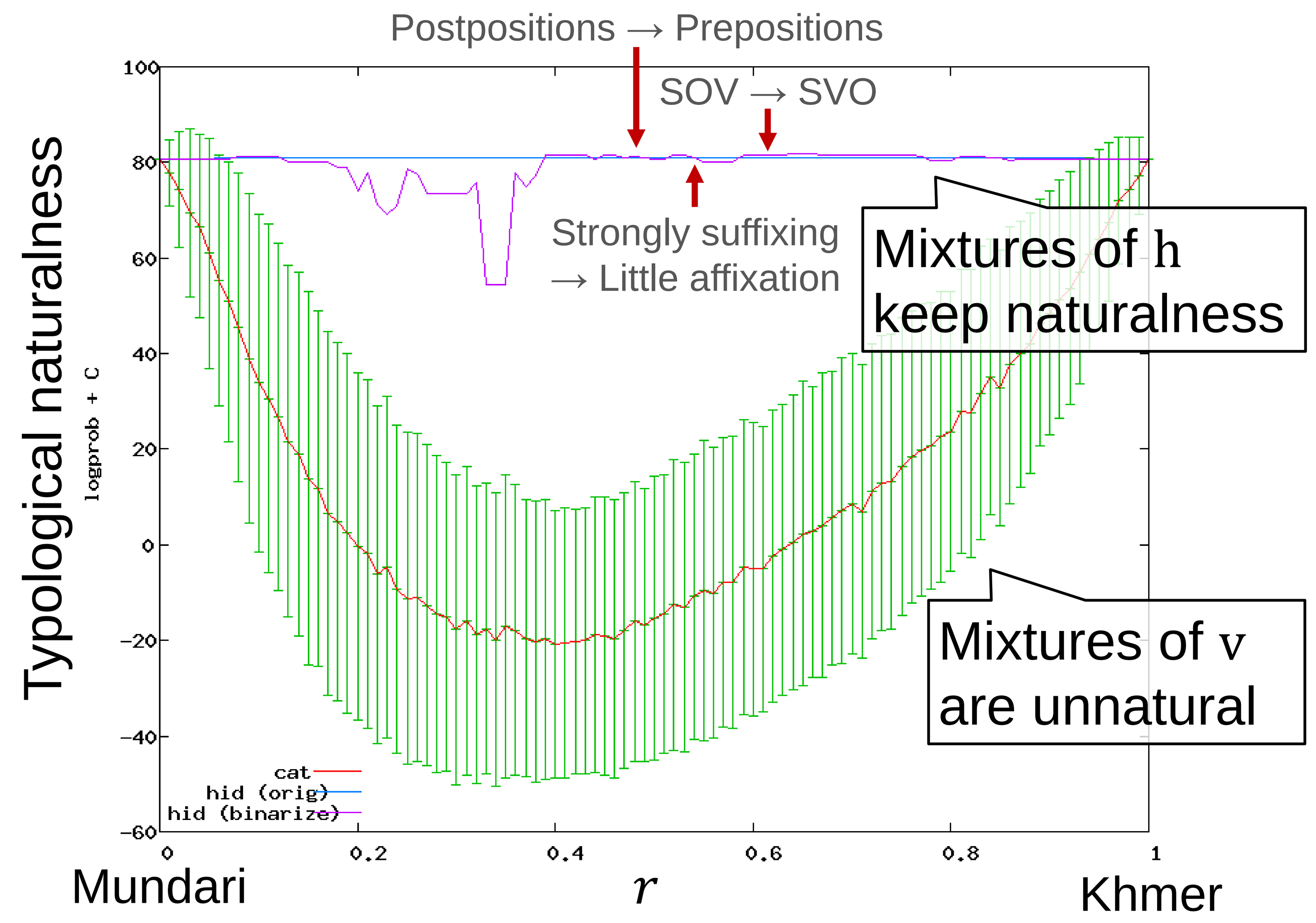
⇧ Binarize the categorical vector (1-of-K)

v: 2 | 2 | 0 | ⋯ | 3 | 3

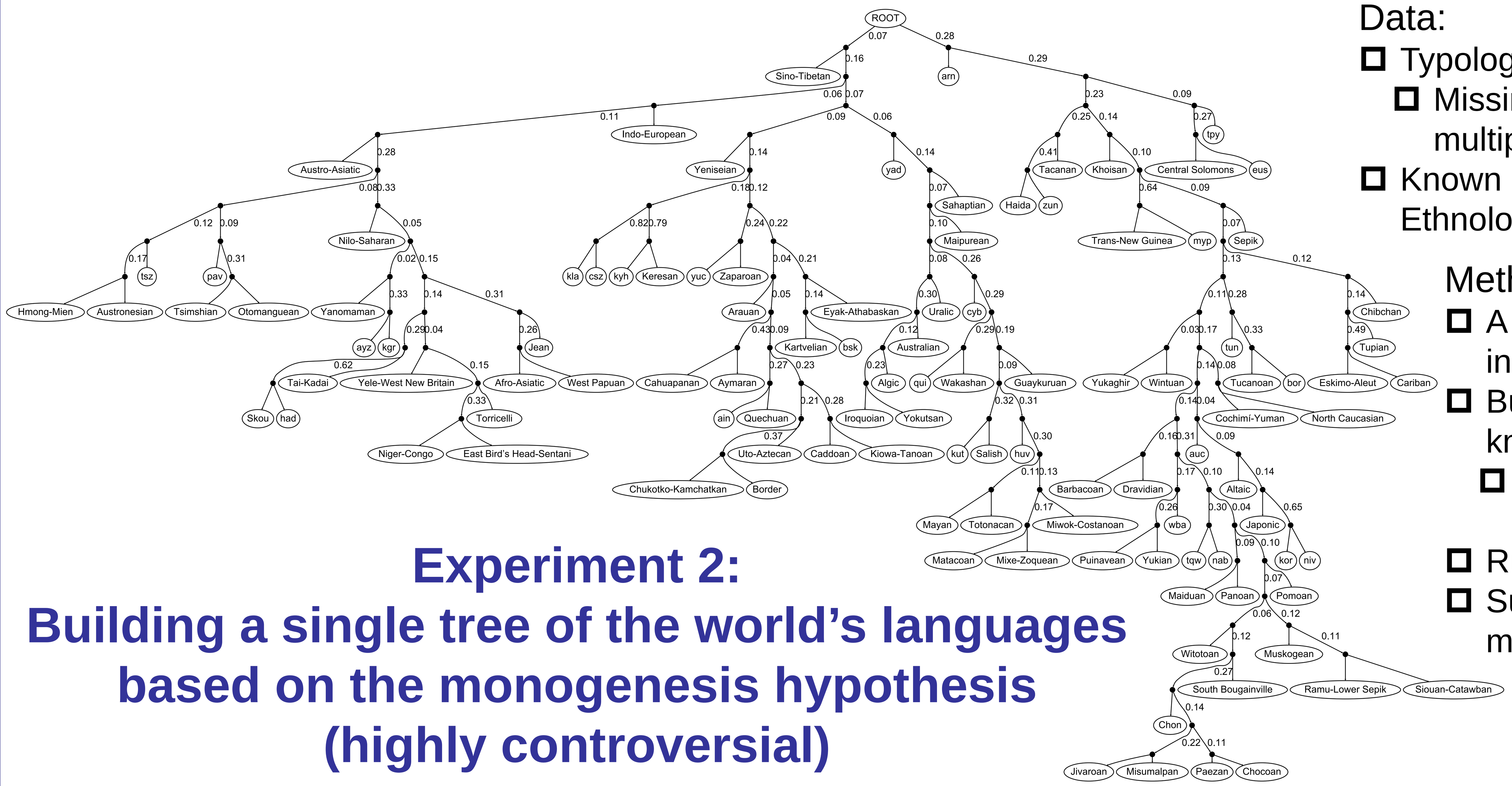
- The Autoencoder non-linearly maps the feature vector into the continuous space
  - The matrix W_e captures dependencies among features
  - The decoder ensures that the original vector is reconstructible
- Energy-based model $\mathrm{P}(\mathrm{x})$ assesses naturalness
  - Trained such that observed languages are distinguished from other possible combinations of features

Experiment 1: Mixing languages

- Assumption: the ancestor is close to an intermediate state between two descendants, a and b
- Compare two ways of mixing languages:
  1. Mixtures of h: linear interpolation in the continuous space $\mathrm{h} = (1-r)\mathrm{h}_a + r\mathrm{h}_b$
  2. Mixtures of v: randomly replace the elements of the categorical vector v_a with those of v_b with prob. r



Experiment 2: Building a single tree of the world's languages based on the monogenesis hypothesis (highly controversial)



Data:
- Typology Database: WALS
  - Missing values are imputed by multiple correspondence analysis
- Known language families: Ethnologue

Method:
- A simple generative tree model in the continuous space
- Build a binary tree on top of known language families
  - Components' stability is learned from known trees
- Run MCMC sampling
- Summarize samples by a maximum clade credibility tree