Gaussian Process Models of Sound Change in Indo-Aryan Dialectology

Chandra A. Cathcart
Department of Comparative Linguistics
University of Zurich

LChange 2019, Florence
2 August 2019

Introduction

Digitized etymological resources and quantitative models can aid us in better understanding language relationships.

Indo-Aryan languages are a group with a long history of scholarship. It is generally agreed that all contemporary Indo-Aryan languages descend from some attested variant of Sanskrit or Old Indo-Aryan (OIA, Emeneau 1966), though some minority views exist (e.g., Kogan, 2005). No truly conclusive understanding of Indo-Aryan subgrouping: a number of hypotheses have been advanced, but no single proposal has emerged as the winner (for summary of challenges, see Southworth 1964; Jeffers 1976; Toomlin 2009).

Sound change is assigned a great deal of explanatory power in Indo-Aryan dialectology (Masica, 1991); a number of sound changes thought to be probative with respect to Indo-Aryan dialectology have been put forth (Hock, 2016). This is in part due to the fact that IA languages have developed in close contact with each other, and intimate lexical borrowing between closely related languages has been widespread (Turner, 1967).

Broad goal of current work: use sound changes extracted from a large etymological database to shed light on IA dialectology.

Narrow goal: use a Gaussian Process prior over sound changes to capture shared database to shed light on IA dialectology.

Hypotheses

Hoernle (1880): four groups nested within two higher-order groups

Grierson, Linguistic Survey of India: Inner-Outter hypothesis, on the basis of several morpho-syntactic and phonological innovations

Chatterji (1926): argues that innovations proposed by Grierson are too chronologically shallow to be meaningful for subgrouping purposes

Zografi (1976): skeptical of I-O hypothesis, inconclusive on I-A subgrouping

Masica (1991): has a history of use in quantitative approaches to sound change (Southworth, 2005; Hock, 2016). This work contrasts two priors over sound change: A Dirichlet prior, which has been shown to perform well in terms of posterior predictive checks (see figure).

We extracted all modern Indo-Aryan (NIA) forms from Turner (1966b) along with the OIA headwords from which these reflexes descend (Middle Indo-Aryan languages such as Prakrit and Pali were excluded).

It is generally agreed that all contemporary Indo-Aryan languages descend from a number of sound changes thought to be probative with respect to Indo-Aryan dialectology (Masica, 1991); a number of sound changes thought to be probative with respect to Indo-Aryan dialectology have been put forth (Hock, 2016). This is in part due to the fact that IA languages have developed in close contact with each other, and intimate lexical borrowing between closely related languages has been widespread (Turner, 1967).

We restrict our analysis to changes affecting OIA segments, such as vowels, diphthongs, and nasals. They are thought to play a meaningful role in Indo-Aryan dialectology (Southworth, 2005; Hock, 2016).

Model and results

We employ mixed-membership models in order to tease apart admixture between IA languages on the basis of sound changes; model assumes that each word in each language is generated by one of K latent dialect components, according to the relevant sound changes whose operation the word displays. Key parameters are \( \theta \) (language-level distributions over dialect components) and \( \phi \) (component-level collections of distributions over sound changes). The stochastic generative process we assume to underlie the data looks as follows:

1. For each word in each OIA etymon:
2. For each language \( l \in \{1, \ldots, L\} \) continuing:
3. \( z_l \sim \text{Categorical}(\theta_l) \) [Draw a dialect component label]
4. For each OIA input \( w_{il} \) in etymon \( w_l \) at index \( i \in \{1, \ldots, |w_{il}|\} \):
5. \( y_{il} \sim \text{Categorical}(\phi_{z_l, w_{il}}) \) [Generate each output]

Across models, \( \phi \) consists of a logistic normal prior over sound change with three different types of covariates: (1) diagonal (cf. Srivastava and Sutton, 2017); “binary” GP prior with ARD kernel, dependent on whether sounds involved were the same or different; and (2) “granular” GP prior, dependent on featural dissimilarity of sounds involved.

Ultimately, the binary GP model performs the best in terms of posterior predictive checks (see figure). If we figure out the best way to probabilistically represent sound change, digitized etymological resources can tell us a great deal about the evolution of linguistic groups such as Indo-Aryan.

References


