The Inner-Outer hypothesis of Indo-Aryan: a computational study

Many Indo-Aryan dialect groupings have been advanced in the literature, primarily on the basis of sound changes, but no single proposal has emerged as the winner. Millennia of contact, diglossia, migration, and cultural exchange greatly complicate the picture of relationships between Indo-Aryan speech varieties. This paper employs a data-driven Bayesian methodology in order to uncover shared dialectal patterns across the Indo-Aryan languages. In particular, we attempt to operationalize Southworth's (2005) hypothesis that two large dialect groups existed, and that communication within these groups was greater historically than communication between them. We find partial support for this hypothesis, according to our implementation.

Grierson (1967 [1903–28]) first proposed that there were two fundamental IA dialectal groups (inner and outer), given the fact that languages in the extreme east and west appear to have undergone the same phonological and morphological changes. Chatterji (1926) argued against this hypothesis, convincingly demonstrating that many of these changes took place at a relatively late date and were hence not probative with respect to dialectology. Southworth (2005) revives the Inner-Outer hypothesis, adducing new pieces of evidence and suggesting that the "sheer number of innovations between east and southwest ... would argue against their being totally independent of each other" (p. 147). This view stands somewhat in contrast to that of Masica (1991), who suggests that the IA realm consists of multiple overlapping dialect groups, lacking clear-cut divisions. Following Masica's (1991) (albeit qualified) observation that "non-arbitrary way of [establishing dialect groups] might appear to lie in giving priority to phonology" (p. 457), we attempt to investigate this issue using Modern IA forms extracted from Turner's (1966) dictionary. We phonologically align each ModIA form with the Sanskrit form (or reconstructed etymon) which it continues using a version of the Needleman-Wunsch algorithm, which allows us to extract the sound changes that have given rise to each ModIA form, as well as the immediate contexts in which they occur (e.g., $r > \emptyset$ / m _ a). We discard forms involving systematic morphological restructuring (e.g., infinitive verbs), leaving 50706 words in 87 languages.

We use a series of Bayesian shared-admixture models in order to test the Inner-Outer hypothesis. Our basic model, which draws upon Latent Dirichlet Allocation, a methodology popular in the field of topic modeling, assumes that each linguistic feature found in a given language is generated by one of two unobserved dialectal components. We implement two versions of this model: in Model 1, individual sound changes are treated as the linguistic feature of interest, given the importance awarded by Southworth to lexical diffusion of changes in shaping interdialectal IA patterns. In Model 2, we assume that each dialectal component generates whole WORDS on the basis of the sound changes they display. Models of this sort assume that each language has a probability distribution over components, and that each component has a distribution over sound changes. A common choice for these distributions is the Dirichlet distribution, which generates probability simplices given one (in our case) or more CONCENTRATION PARAMETERS (henceforth β), which can be fixed or inferred from the data. If $\beta > 1$, smooth or relatively uniform probability distributions are generated; if $\beta < 1$, sparse distributions are generated, with one outcome dominating probability mass. We infer the value of the concentration parameter of the language-component distribution. If Southworth's view is correct, then β should be below 1, indicating that languages take most of their features from one distribution, compatible with the notion that dialect groups maintained their integrity over a long period of time.

Preliminary results show that in Model 1, all posterior values of β are greater than 1; in Model 2, however, roughly one fourth of the posterior values are less than 1 indicating some degree of support for the Inner-Outer hypothesis, though this evidence is not statistically significant. It is worth noting that a model designed to represent lexical diffusion of sound changes along provides no support for the Inner-Outer model, given Southworth's discussion of this sociolinguistic process. We discuss the implications of this model and the distribution of dialect components across languages, as well as the sound changes that we find associated with each dialect group.

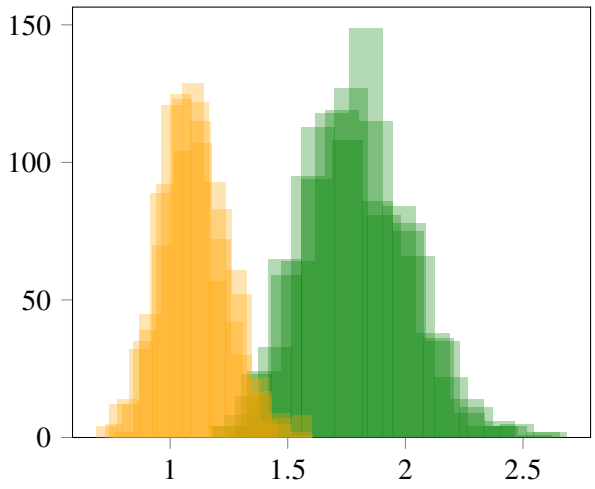

Figure 1: Posterior distributions for β for model 1 (green) and model 2 (orange)

References

Chatterji, S. K. (1926). *The Origin and Development of the Bengali Language*. Calcutta: Calcutta University Press.

Grierson, G. A. (1967 [1903-28]). *Linguistic Survey of India*. Delhi: Motilal Banarsidass.

Masica, C. P. (1991). *The Indo-Aryan languages*. Cambridge: Cambridge University Press.

Southworth, F. C. (2005). *Linguistic Archaeology of South Asia*. London: Routledge.

Turner, R. L. (1962–1966). *A comparative dictionary of Indo-Aryan languages*. London: Oxford University Press.