Toward a deep dialectological representation of Indo-Aryan

Chundra A. Cathcart
Department of Comparative Linguistics
University of Zurich
chundra.cathcart@uzh.ch

Abstract

This paper presents a new approach to disentangling inter-dialectal and intra-dialectal relationships within one such group, the Indo-Aryan subgroup of Indo-European. I draw upon admixture models and deep generative models to tease apart historic language contact and language-specific behavior in the overall patterns of sound change displayed by Indo-Aryan languages. I show that a "deep" model of Indo-Aryan dialectology sheds some light on questions regarding inter-relationships among the Indo-Aryan languages, and performs better than a "shallow" model in terms of certain qualities of the posterior distribution (e.g., entropy of posterior distributions), and outline future pathways for model development.

1 Introduction

At the risk of oversimplifying, quantitative models of language relationship fall into two broad categories. At a wide, family-level scale, phylogenetic methods adopted from computational biology have had success in shedding light on the histories of genetically related but significantly diversified speech varieties (Bouckaert et al., 2012). At a shallower level, the subfield of dialectometry has used a wide variety of chiefly distance-based methodologies to analyze variation among closely related dialects with similar lexical and typological profiles (Nerbonne and Heeringa, 2001), though this work also emphasizes the importance of hierarchical linguistic relationships and the use of abstract, historically meaningful features (Prokić and Nerbonne, 2008; Nerbonne, 2009). It is possible, however, that neither methodology is completely effective for for language groups of intermediate size, particularly those where certain languages have remained in contact to an extent that blurs the phylogenetic signal, but have experienced great enough diversification that dialectometric approaches are not appropriate. This paper presents a new approach to disentangling inter-dialectal and intra-dialectal relationships within one such group, the Indo-Aryan subgroup of Indo-European.

Indo-Aryan presents many interesting puzzles. Although all modern Indo-Aryan (henceforth NIA) languages descend from Sanskrit or Old Indo-Aryan (henceforth OIA), their subgrouping and dialectal interrelationships remain somewhat poorly understood (for surveys of assorted problems, see Emeneau 1966; Masica 1991; Toulmin 2009; Smith 2017; Deo 2018). This is partly due to the fact that these languages have remained in contact with each other, and this admixture has complicated our understanding of the languages' history. Furthermore, while most NIA languages have likely gone through stages closely resembling attested Middle Indo-Aryan (MIA) languages such as Prakrit or Pali, no NIA language can be taken with any certainty to be direct descendants of an attested MIA variety, further shrouding the historical picture of their development.

The primary goal of the work described in this paper is to build, or work towards building, a model of Indo-Aryan dialectology that incorporates realistic assumptions regarding historical linguistics and language change. I draw upon admixture models and deep generative models to tease apart historic language contact and language-specific behavior in the overall patterns of sound change displayed by Indo-Aryan languages. I show that a "deep" model of Indo-Aryan dialectology sheds some light on questions regarding inter-relationships among the Indo-Aryan languages, and performs better than a "shallow" model in terms of certain qualities of the posterior distribution (e.g., entropy of posterior distributions). I provide a comparison with other met-

Proceedings of VarDial, pages 110–119
Minneapolis, MN, June 7, 2019

rics, and outline future pathways for model development.

2 Sound Change

The notion that sound change proceeds in a regular and systematic fashion is a cornerstone of the comparative method of historical linguistics. When we consider cognates such as Greek *p^herō* and Sanskrit *bharā(mi)* 'I carry', we observe regular sound correspondences (e.g., *p^h*:*bh*) which allow us to formulate sound changes that have operated during the course of each language's development from their shared common ancestor. Under ideal circumstances, these are binary yes/no questions (e.g., Proto-Indo-European **b^h* > Greek *p^h*). At other times, there is some noise in the signal: for instance, OIA *kṣ* is realized as *kh* in most Romani words (e.g., *akṣi-* 'eye' > *jakh*), but also as *čh* (*kṣurikā-* > *čhuri* 'knife'), according to Matras (2002, 41). This is undoubtedly due to relatively old language contact (namely lexical borrowing) between prehistoric Indo-Aryan dialects, as opposed to different conditioning environments which trigger a change *kṣ* > *kh* in some phonological contexts but *kṣ* > *čh* in others. The idea that Indo-Aryan speech varieties borrowed forms from one another on a large scale is well established (Turner, 1975 [1967], 406), as is often the case in situations where closely related dialects have developed in close geographic proximity to one another (cf. Bloomfield, 1933, 461–495). An effective model of Indo-Aryan dialectology must be able to account this sort of admixture. Phylogenetic methods and distance-based methods provide indirect information regarding language contact (e.g., in the form of uncertain tree topologies), but do not explicitly model intimate borrowing.

A number of studies have used mixed-membership models such as the Structure model (Pritchard et al., 2000) in order to explicitly model admixture between languages (Reesink et al., 2009; Syrjänen et al., 2016). Under this approach, individual languages receive their linguistic features from latent ancestral components with particular feature distributions. A key assumption of the Structure model is the relative invariance and stability of the features of interest (e.g., allele frequencies, linguistic properties). However, sound change is a highly recurrent process, with many telescoped and intermediate changes, and it is not possible to treat sound changes that have operated as stable, highly conservative features.[1]

Intermediate stages between OIA and NIA languages are key for capturing similarities in cross-linguistic behavior, and we require a model that teases apart dialect group-specific trends and language-level ones. Consider the following examples:

- Assamese /x/, the reflex of OIA *s, ś, ṣ*, is thought to develop from intermediate **ś* (Kakati, 1941, 224). This isogloss would unite it with languages like Bengali, which show /ʃ/ for OIA *s, ś, ṣ*.
- Some instances of NIA *bh* likely come from an earlier **mh* (Tedesco 1965, 371; Oberlies 2005, 48) (cf. Oberlies 2005:48).
- The Marathi change *ch* > *s* affects certain words containing MIA **ch* < OIA *kṣ* as well as OIA *ch* (Masica, 1991, 457); *ch* ∼ *kh* < OIA *kṣ* variation is of importance to MIA and NIA dialectology (compare the Romani examples given above).

In all examples, a given NIA language shows the effects of chronologically deep behavior which serves as an isogloss uniting it with other NIA languages, but this trend is masked by subsequent language-specific changes.[2] Work on probabilistic reconstruction of proto-word forms explicitly appeals to intermediate chronological stages where linguistic data are unobserved (Bouchard-Côté et al., 2007); however, unlike the work cited, this paper does not assume a fixed phylogeny, and hence I cannot adopt many of the simplifying conventions that the authors use.

3 Data

I extracted all modern Indo-Aryan forms from Turner's (1962–1966) *Comparative Dictionary of the Indo-Aryan Languages* (henceforth CDIAL),[3]

[1] Cathcart (to appear) circumvents this issue in a mixed-membership model of Indo-Aryan dialectology by considering only sound changes thought *a priori* in the literature to be relatively stable and of importance to dialectology.

[2] Some similar-looking sound changes can be shown to be chronologically shallow. For instance, the presence of *ṣ* for original *kh* in Old Braj, taken by most scholars to represent a legitimate sound change and not just an orthographic idiosyncrasy, affects Persian loans such as *ṣaracu* 'expense' ← Modern Persian *xirč* (McGregor, 1968, 125). This orthographic behavior is found in Old Gujarati as well (Baumann, 1975, 9). For further discussion of this issue, see Strnad 2013, 16ff.

[3] Available online at http://dsal.uchicago.edu/dictionaries/soas/

along with the Old Indo-Aryan headwords (henceforth ETYMA) from which these reflexes descend. Transcriptions of the data were normalized and converted to the International Phonetic Alphabet (IPA). Systematic morphological mismatches between OIA etyma and reflexes were accounted for, including stripping the endings from all verbs, since citation forms for OIA verbs are in the 3sg present, while most NIA reflexes give the infinitive. I matched each dialect with corresponding languoids in Glottolog (Hammarström et al., 2017) containing geographic metadata, resulting in the merger of several dialects. I excluded cognate sets with fewer than 10 forms, yielding 33231 modern Indo-Aryan forms. I preprocessed the data, first converting each segment into its respective sound class, as described by List (2012), and subsequently aligning each converted OIA/NIA string pair via the Needleman-Wunsch algorithm, using the Expectation-Maximization method described by Jäger (2014), building off of work by Wieling et al. (2012). This yields alignments of the following type: e.g., OIA /aːntra/ 'entrails' > Nepali /aːn∅ro/, where ∅ indicates a gap where the "cursor" advances for the OIA string but not the Nepali string. Gaps on the OIA side are ignored, yielding a one-to-many OIA-to-NIA alignment; this ensures that all aligned cognate sets are of the same length.

4 Model

The basic family of model this paper employs is a Bayesian mixture model which assumes that each word in each language is generated by one of K latent dialect components. Like Structure (and similar methodologies like Latent Dirichlet Allocation), this model assumes that different elements in the same language can be generated by different dialect components. Unlike the most basic type of Structure model, which assumes a two-level data structure consisting of (1) languages and the (2) features they contain, our model assumes a three-level hierarchy, where (1) languages contain (2) words, which display the operation of different (3) sound changes; latent variable assignment happens at the word level.

I contrast the behavior of a DEEP model with that of a SHALLOW model. The deep model draws inspiration from Bayesian deep generative models (Ranganath et al., 2015), which incorporate intermediate latent variables which mimic the architecture of a neural network. This structure allows us to posit an intermediate representation between the sound patterns in the OIA etymon and the sound patterns in the NIA reflex, allowing the model to pick up on shared dialectal similarities between forms in languages as opposed to language-specific idiosyncrasies. The shallow model, which serves as a baseline of sorts, conflates dialect group-level and language-level trends; it contains a flat representation of all of the sound changes taking place between a NIA word and its ancestral OIA etymon, and in this sense is halfway between a Structure model and a Naïve Bayes classifier (with a language-specific rather than global prior over component membership).

4.1 Shallow model

Here, I describe the generative process for the shallow model, assuming W OIA etyma, L languages, K dialect components, I unique OIA inputs, O unique NIA outputs, and aligned OIA-NIA word pair lengths $T_w : w \in \{1, ..., W\}$. For each OIA etymon, an input $x_{w,t}$ at time point $t \in \{1, ..., T_w\}$ consists of a trigram centered at the timepoint in question (e.g., ntr in OIA /aːntra/ 'entrails'), and the NIA reflex's output $y_{w,l,t}$ contains the segment(s) aligned with timepoint t (e.g., Nepali ∅). $x_{w,t} : t = 0$ is the left word boundary, while $x_{w,t} : t = T_w + 1$ is the right word boundary. Accordingly, sound change in the model can be viewed as a rewrite rule of the type A > B / C _ D. The model has the following parameters:

- Language-level weights over dialect components: $U_{l,k}; l \in \{1, ..., L\}, k \in \{1, ..., K\}$
- Dialect component-level weights over sound changes: $W_{k,i,o}; k \in \{1, ..., K\}, i \in \{1, ..., I\}, o \in \{1, ..., O\}$

The generative process is as follows:

For each OIA etymon $x_w \in \{1, ..., W\}$

  For each language $l \in \{1, ..., L\}$ in which the etymon survives, containing a reflex $y_{w,l}$

    Draw a dialect component assignment $z_{w,l} \sim \text{Categorical}(f(U_{l,\cdot}))$

    For each time point $t \in \{1, ..., T_w\}$

      Draw a NIA sound $y_{w,l,t} \sim \text{Categorical}(f(W_{z_{w,l},x_{w,t},\cdot}))$

All weights in U and W are drawn from a Normal distribution with a mean of 0 and standard deviation of 10; $f(\cdot)$ represents the softmax function (throughout this paper), which transforms these weights to probability simplices. The generative process yields the following joint log likelihood of the OIA etyma $\boldsymbol{x}$ and NIA reflexes $\boldsymbol{y}$ (with the discrete latent variables $\boldsymbol{z}$ marginalized out:

$$P(\boldsymbol{x}, \boldsymbol{y}|U, W) = \prod_{w=1}^{W} \prod_{l=1}^{L} \sum_{k=1}^{K} \left[f(U_{l,k}) \prod_{t=1}^{T_w} f(W_{k,x_{w,l,t},y_{w,l,t}}) \right] \quad (1)$$

As readers will note, this model weights all sound changes equally, and makes no attempt to distinguish between dialectologically meaningful changes and noisy, idiosyncratic changes.

4.2 Deep model

The deep model, like the shallow model, is a mixture model, and as such retains the language-level weights over dialect component membership U. However, unlike the shallow model, in which the likelihood of an OIA etymon and NIA reflex under a component assignment $z = k$ is dependent on a flat representation of edit probabilities between OIA trigrams and NIA unigrams associated with dialect component k. Here, I attempt to add some depth to this representation of sound change by positing a hidden layer of dimension J between each $x_{w,t}$ and $y_{w,l,t}$. The goal here is to mimic a "noisy" reconstruction of an intermediate stage between OIA and NIA represented by dialect group k. This reconstruction is not an explicit, linguistically meaningful string (as in Bouchard-Côté et al. 2007, 2008, 2013); furthermore, it is re-generated for each individual reflex of each etymon, and not shared across data points (such a model would introduce deeply nested dependencies between variables, and enumerating all possible reconstructions would be computationally infeasible).

For parsimony's sake, I employ a simple Recurrent Neural Network (RNN) architecture to capture rightward dependencies (Elman, 1990). Figure 1 gives a visual representation of the network, unfolded in time. This model exchanges W, the dialect component-level weights over sound changes, for the following parameters:

- Dialect component-level weights governing hidden layer unit activations by OIA sounds: $W^x_{k,i,j}; k \in \{1, ..., K\}, i \in \{1, ..., I\}, j \in \{1, ..., J\}$
- Dialect component-level weights governing hidden layer unit activations by previous hidden layers: $W^h_{k,i,j}; k \in \{1, ..., K\}, i \in \{1, ..., J\}, j \in \{1, ..., J\}$
- Language-level weights governing NIA output activations by hidden layer units: $W^y_{l,j,o}; l \in \{1, ..., L\}, j \in \{1, ..., J\}, o \in \{1, ..., O\}$

For a given mixture component $z = k$, the activation of the hidden layer at time t, h_t, depends on two sets of parameters, each associated with component k: the weights $W^x_{k,x_{x_t},\cdot}$, associated with the OIA input at time t; and W^h_k, the weights associated with the previous hidden layer h_{t-1}'s activations, for all $t > 1$. Given a hidden layer h_t, the weights W^l can be used to generate a probability distribution over possible outcomes in NIA language l. The forward pass of this network can be viewed as a generative process, denoted $\boldsymbol{y}_{w,t} \sim \mathsf{RNN}(\boldsymbol{x}_{w,l}, W^x_k, W^h_k, W^l)$ under the parameters for component k and language l; under such a process, the likelihood of $\boldsymbol{y}_{w,l}$ can be computed as follows:

$$P_{\mathsf{RNN}}(\boldsymbol{y}_{w,l}|\boldsymbol{x}_w, W^x_k, W^h_k, W^l) = \prod_{t=1}^{T_w} f(h_t^\top W^l)_{y_{w,l,t}} \quad (2)$$

where

$$h_t = \begin{cases} f(W^x_{k,x_{w,t},\cdot}), & \text{if } t = 1 \\ f(h_{t-1}^\top W^h \oplus W^x_{k,x_{w,t},\cdot}), & \text{if } t > 1 \end{cases} \quad (3)$$

The generative process for this model is nearly identical to the process described in the previous sections; however, after the dialect component assignment ($z_{w,l} \sim \text{Categorical}(f(U_{l,\cdot}))$) is drawn, the NIA string $\boldsymbol{y}_{w,l}$ is sampled from $\mathsf{RNN}(\boldsymbol{x}_w, W^x_{z_{w,l}}, W^h_{z_{w,l}}, W^l)$. The joint log likelihood of the OIA etyma $\boldsymbol{x}$ and NIA reflexes $\boldsymbol{y}$ (with the discrete latent variables $\boldsymbol{z}$ marginalized out is the following:

$$P(\boldsymbol{x}, \boldsymbol{y}|U, W^x, W^h, W^y) = \prod_{w=1}^{W} \prod_{l=1}^{L} \sum_{k=1}^{K} \left[f(U_{l,k}) P_{\mathsf{RNN}}(\boldsymbol{y}_{w,l}|\boldsymbol{x}_w, W^x_k, W^h_k, W^l) \right] \quad (4)$$

The same $\mathcal{N}(0, 10)$ prior as above is placed over U, W^x, W^h, W^y. J, the dimension of the hidden layer, is fixed at 100. This model bears some similarities to the mixture of RNNs described by Kim et al. (2018).

I have employed a simple RNN (rather than a more state-of-the art architecture) for several reasons. The first is that I am interested in the consequences of expanding a flat mixture model to contain a simple, slightly deeper architecture. Additionally, I believe that the fact that the hidden layer of an RNN can be activated by a softmax function is more desirable from the perspective of representing sound change as a categorical or multinomial distribution, as all layer unit activations sum to one, as opposed to the situation with Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), which traditionally use sigmoid or hyperbolic tangent functions to activate the hidden layer. Furthermore, long-distance dependencies are not particularly widespread in Indo-Aryan sound change, lessening the need for more complex architectures. At the same time, the RNN is a crude approximation to the reality of language change. RNNs and related models draw a single arc between a hidden layer at time t and the corresponding output. It is perhaps not appropriate to envision this single dependency unless the dimensionality of the hidden layer is large enough to absorb potential contextual information that is crucial to sound change. To put it simply, emission probabilities in sound change are sharper than transitions common in most NLP applications (e.g., sentence prediction), and it may not be correct to envision y_t given $h_{t'<t}, h_t$ as a function of an additive combination of weights, though in practice, I find it too computationally costly to enumerate all possible value combinations the hidden layer at multiple consecutive time points. This issue requires further exploration, and I employ what seems to be the most computationally tractable approach for the moment.

5 Results

I learn each model's MAP configuration using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of .1.[4] I run the optimizer for 10000 iterations over three random initializations, fitting the model on mini-batches of 100 data points, and monitor convergence by observing the trace of the log posterior (Figure 2).

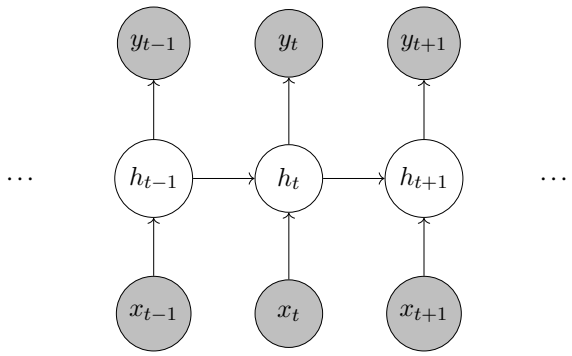

Figure 1: RNN representation, unfolded in time: hidden layers depend on OIA inputs $x_1, ..., x_{T_w}$ and previous hidden layers (for $t > 1$); NIA outputs $y_1, ..., y_{T_w}$ depend on hidden layers. Hidden layer activations are dependent on dialect component-specific parameters, while activations of the output layer are dependent on individual NIA language-specific parameters.

The flat model fails to pick up on any major differences between languages, finding virtually identical posterior values of $f(U_l)$, the language-level distribution over dialect component membership, for all $l \in \{1, ..., L\}$. According to the MAP configuration, each language draws forms from the same dialect group with $> .99$ probability, essentially undergoing a sort of "component collapse" that latent variable models sometimes encounter (Bowman et al., 2015; Dinh and Dumoulin, 2016). It is likely that bundling together sound change features leads to component-level distributions over sound changes with high entropy that are virtually indistinguishable from one another.[5] While this particular result is disappointing in the lack of information it provides, I observe some properties of our models' posterior values in order to diagnose problems that can be addressed in future work (discussed below).

The deep model, on the other hand, infers highly divergent language-level posterior distributions over cluster membership. Since these distributions are not identical across initializations due to the label-switching problem, I compute the Jensen-Shannon divergence between the language-level posterior distributions over cluster membership for each pair of languages in our sample for each initialization. I then average these divergences across initializations. These averaged

[4] Code for all experiments can be found at https://github.com/chundrac/IA_dial/VarDial2019.

[5] I made several attempts to run this model with different specifications, including different prior distributions, but achieved the same result.

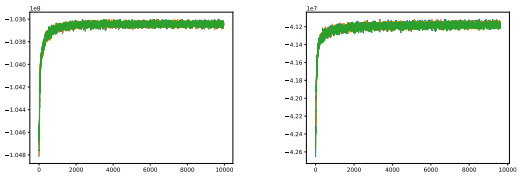

Figure 2: Log posteriors for shallow model (left) and deep model (right) for 10000 iterations over three random initializations.

divergences are then scaled to three dimensions using multidimensional scaling. Figure 3 gives a visualization of these transformed values via the red-green-blue color vector, plotted on a map; languages with similar component distributions display similar colors. With a few exceptions (that may be artifacts of the fact that certain languages have only a small number of data points associated with them), a noticieable divide can be seen between languages of the main Indo-Aryan speech region on one hand, and languages of northwestern South Asia (dark blue), the Dardic languages of Northern Pakistan, and the Pahari languages of the Indian Himalayas, though this division is not clear cut. Romani and other Indo-Aryan varieties spoken outside of South Asia show affiliation with multiple groups. While Romani dialects are thought to have a close genetic affinity with Hindi and other Central Indic languages, it was likely in contact with languages of northwest South Asian during the course of its speakers' journey out of South Asia (Hamp, 1987; Matras, 2002). However, this impressionistic evaluation is by no means a confirmation that the deep model has picked up on linguistically meaningful differences between speech varieties. In the following sections, some comparison and evaluation metrics and checks are deployed in order to assess the quality of these models' behavior.

5.1 Entropy of distributions

I measure the average entropy of the model's posterior distributions in order to gauge the extent to which the models are able to learn sparse, informative distributions over sound changes, hidden state activations, or other parameters concerning transitions through the model architecture. Normalized entropy is used in order to make entropies of distributions of different dimension comparable; a distribution's entropy can be normalized by dividing by its maximum possible entropy.

As mentioned above, our data set consists of OIA trigrams and the NIA segment corresponding to the second segment in the trigram, representing rewrite rules operating between OIA and the NIA languages in our sample. It is often the case that more than one NIA reflex is attested for a given OIA trigram. As such, the sound changes that have operated in an NIA language can be represented as a collection of categorical distributions, each summing to one. I calculate the average of the normalized entropies of these sound change distributions as a baseline against which to compare entropy values for the models' parameters. The pooled average of the normalized entropies across all languages is .11, while the average of averages for each language is .063.

For the shallow model, the parameter of interest is $f(V)$, the dialect component-level collection of distributions over sound changes, the mean normalized entropy of which, averaged across initializations but pooled across components within each initialization, is 0.91 (raw values range from 0.003 to 1). For the deep model, the average entropy of the dialect-level distributions over hidden-layer activations, $f(W^x)$, is only slightly lower, at 0.86 (raw values range from close to 0 to 1).

For each $k \in \{1, ..., K\}$, I compute the forward pass of $\mathsf{RNN}(\boldsymbol{x}_{w,l}, W^x_k, W^h_k, W^l)$ for each etymon w and each language l in which the etymon survives using the inferred values for W^x_k, W^h_k, W^l and compute the entropy of each $f(h_t^\top W^l)$, yielding an average of .74 (raw values range from close to 0 to 1). While these values are still very high, it is clear that the inclusion of a hidden layer has learned sparser, potentially more meaningful distributions than the flat approach, and that increasing the dimensionality of the hidden layer will likely bring about even sparser, more meaningful distributions. The entropies cited here are considerably higher than the average entropy of languages' sound change distributions, but the latter distributions do little to tell us about the internal clustering of the languages.

5.2 Comparison with other linguistic distance metrics

Here, I compare the cluster membership inferred by this paper's models against other measures of linguistic distance. Each method yields a pairwise inter-language distance metric, which can be compared against a non-linguistic measure. I measure

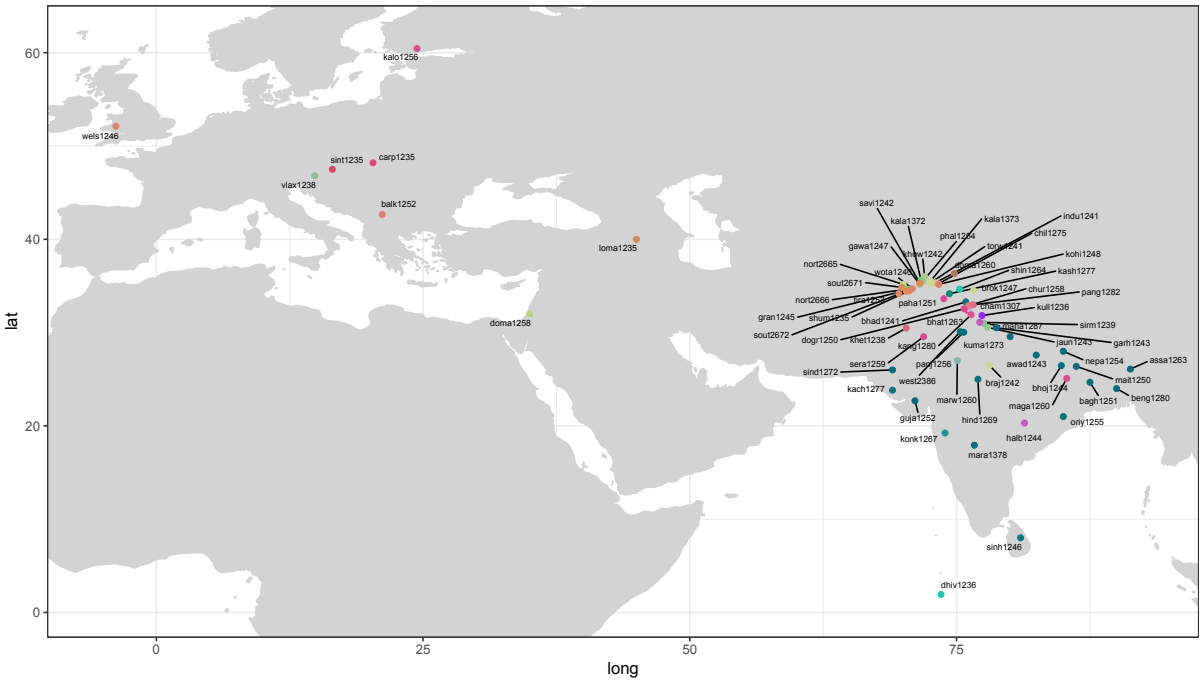

Figure 3: Dialect group makeup of languages in sample under deep model

the correlation between each linguistic distance measure as well as great circle geographic distance and patristic distance according to the Glottolog phylogeny using Spearman's ρ.

5.2.1 Levenshtein distance

Borin et al. (2014) measure the normalized Levenshtein distances (i.e., the edit distance between two strings divided by the length of the longer string) between words for the same concept in pairs of Indo-Aryan languages, and find that average normalized Levenshtein distance correlates significantly with patristic distances in the Ethnologue tree. This paper's dataset is not organized by semantic meaning, so for comparability, I measure the average normalized Levenshtein distance between cognates in pairs of Indo-Aryan languages, which picks up on phonological divergence between dialects, as opposed to both phonological and lexical divergence.

5.2.2 Jensen-Shannon divergence

Each language in our dataset attests one or more (due to language contact, analogy, etc.) outcomes for a given OIA trigram, yielding a collection of sound change distributions, as described above. For each pair of languages, I compute the Jensen-Shannon divergence between sound change distributions for all OIA trigrams that are continued in both languages, and average these values. This gives a measure of pairwise average diachronic phonological divergence between languages.

5.2.3 LSTM Autoencoder

Rama and Çöltekin (2016) and Rama et al. (2017) develop an LSTM-based method for representing the phonological structure of individual word forms across closely related speech varieties. Each string is fed to a unidirectional or bidirectional LSTM autoencoder, which learns a continuous latent multidimensional representation of the sequence. This embedding is then used to reconstruct the input sequence. The latent values in the embedding provide information that can be used to compute dissimilarity (in the form of cosine or Euclidean distance) between strings or across speech varieties (by averaging the latent values for all strings in each dialect or language). I use the bidirectional LSTM Autoencoder described in the work cited in order to learn an 8-dimensional latent representation for all NIA forms in the dataset, training the model over 20 epochs on batches of 32 data points using the Adam optimizer to minimize the categorical cross-entropy between the input sequence and the NIA reconstruction predicted by the model. I use the learned model parameters to generate a latent representation for each form. The latent representations are averaged across forms within each language, and pairwise linguistic Euclidean distances are computed between each av-

	Geographic	Genetic
Shallow JSD	-0.01	-0.03
Deep JSD	$0.147*$	0.008
LDN	$0.346*$	0.013
Raw JSD	$0.302*$	$-0.051*$
LSTM AE	$0.158*$	$-0.068*$
LSTM ED	$0.084*$	0.0001

Table 1: Spearman's ρ values for correlations between each linguistic distance metric (JSD = Jensen-Shannon Divergence, LDN = Levenshtein Distance Normalized, AE = Autoencoder, ED = Encoder-Decoder) and geographic and genetic distance. Asterisks represent significant correlations.

eraged representation.

5.2.4 LSTM Encoder-Decoder

For the sake of completeness, I use an LSTM encoder-decoder to learn a continuous representation for every OIA-NIA string pair. This model is very similar to the LSTM autoencoder, except that it takes an OIA input and reconstructs an NIA output, instead of taking an NIA form as input and reconstructing the same string. I train the model as described above.

5.3 Correlations

Table 1 gives correlation coefficients (Spearman's ρ) between linguistic distance metrics and non-linguistic distance metrics. In general, correlations with Glottolog patristic distance are quite poor. This is surprising for Levenshtein Distance Normalized, given the high correlation with patristic distance reported by Borin et al. (2014). Given that the authors measured Levenshtein distance between identical concepts in pairs of languages, and not cognates, as I do here, it is possible that lexical divergence carries a stronger genetic signal than phonological divergence, at least in the context of Indo-Aryan (it is worth noting that I did not balance the tree, as described by the authors; it is not clear that this would have yielded any improvement). On the other hand, the Levenshtein distance measured in this paper correlates significantly with great circle distance, indicating a strong geographic signal. Average Jensen-Shannon divergence between pairs of languages' sound change distributions shows a strong association with geographic distance as well.

Divergence/distances based on the deep model, the LSTM Autoencoder, and the LSTM Encoder-Decoder show significant correlations with geospatial distance, albeit lower ones. It is not entirely clear what accounts for this disparity. Intuitively, we expect more shallow chronological features to correlate with geographic distance. It is possible that the LSTM and RNN architectures are picking up on chronologically deeper information, and show a low geographic signal for this reason, though this highly provisional idea is not borne out by any genetic signal.

It is not clear how to assess the meaning of these correlations at this stage. Nevertheless, deep architectures provide an interesting direction for future research into sound change and language contact, as they have the potential to disaggregate a great deal of information regarding interacting forces in language change that is censored when raw distance measures are computed directly from the data.

6 Outlook

This paper explored the consequences of adding hidden layers to models of dialectology where the languages have experienced too much contact for phylogenetic models to be appropriate, but have diversified to the extent that traditional dialectometric approaches are not applicable. While the model requires some refinement, its results point in a promising direction. Modifying prior distributions could potentially produce more informative results, as could tweaking hyperparameters of the learning algorithms employed. Additionally, it is likely that the model will benefit from hidden layers of higher dimension J, as well as bidirectional approaches, and despite the misgivings regarding LSTM and GRUs stated above, future work will probably benefit from incorporating these and related architectures (e.g., attention). Additionally, the models used in this paper assumed discrete latent variables, attempting to be faithful to the traditional historical linguistic notion of intimate borrowing between discrete dialect groups. However, continuous-space models may provide a more flexible framework for addressing the questions asked in this paper (cf. Murawaki, 2015).

This paper provides a new way of looking at dialectology and linguistic affiliation; with refinement and expansion, it is hoped that this and related models can further our understanding of the history of the Indo-Aryan speech community and can generalize to new linguistic scenarios. It is

hoped that methodologies of this sort can join forces with similar tools designed to investigate interaction of regularly conditioned sound change and chronologically deep language contact in individual languages' histories.

References

George Baumann. 1975. *Drei Jaina-Gedichte in Alt-Gujarātī: Edition, Übersetzung, Grammatik, und Glossar*. Franz Steiner, Wiesbaden.

Leonard Bloomfield. 1933. *Language*. Holt, Rinehart and Winston, New York.

Lars Borin, Anju Saxena, Taraka Rama, and Bernard Comrie. 2014. Linguistic landscaping of south asia using digital language resources: Genetic vs. areal linguistics. In *Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3137–3144.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110:4224–4229.

Alexandre Bouchard-Côté, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896, Prague. Association for Computational Linguistics.

Alexandre Bouchard-Côté, Percy S Liang, Dan Klein, and Thomas L Griffiths. 2008. A probabilistic approach to language change. In *Advances in Neural Information Processing Systems*, pages 169–176.

R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *Proceedings of the Twentieth Conference on Computational Natural Language Learning (CoNLL)*.

Chundra Cathcart. to appear. A probabilistic assessment of the Indo-Aryan Inner-Outer Hypothesis. *Journal of Historical Linguistics*.

Ashwini Deo. 2018. Dialects in the Indo-Aryan landscape. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pages 535–546. John Wiley & Sons, Oxford.

Laurent Dinh and Vincent Dumoulin. 2016. Training neural Bayesian nets. http://www.iro.umontreal.ca/bengioy/cifar/NCAP2014-summerschool/slides/Laurent_dinh_cifar_presentation.pdf.

Jeffrey Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Murray B. Emeneau. 1966. The dialects of Old-Indo-Aryan. In Jaan Puhvel, editor, *Ancient Indo-European dialects*, pages 123–138. University of California Press, Berkeley.

Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2017. Glottolog 3.3. Max Planck Institute for the Science of Human History.

Eric P Hamp. 1987. On the sibilants of romani. *Indo-Iranian Journal*, 30(2):103–106.

Gerhard Jäger. 2014. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In *Quantifying Language Dynamics*, pages 155–204. Brill.

Banikanta Kakati. 1941. *Assamese, its formation and development*. Government of Assam, Gauhati.

Yoon Kim, Sam Wiseman, and Alexander M Rush. 2018. A tutorial on deep latent variable models of natural language. *arXiv preprint arXiv:1812.06834*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Johann-Mattis List. 2012. SCA. Phonetic alignment based on sound classes. In M. Slavkovik and D. Lassiter, editors, *New directions in logic, language, and computation*, pages 32–51. Springer, Berlin, Heidelberg.

Colin P. Masica. 1991. *The Indo-Aryan languages*. Cambridge University Press, Cambridge.

Yaron Matras. 2002. *Romani – A Linguistic Introduction*. Cambridge University Press, Cambridge.

R. S. McGregor. 1968. *The language of Indrajit of Orcha*. Cambridge University Press, Cambridge.

Yugo Murawaki. 2015. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 324–334.

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

John Nerbonne and Wilbert Heeringa. 2001. Computational comparison and classification of dialects. *Dialectologia et Geolinguistica*, 9:69–83.

Thomas Oberlies. 2005. *A historical grammar of Hindi*. Leykam, Graz.

Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Jelena Prokić and John Nerbonne. 2008. Recognising groups among dialects. *International journal of humanities and arts computing*, 2(1-2):153–172.

Taraka Rama and Çağrı Çöltekin. 2016. Lstm autoencoders for dialect analysis. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 25–32.

Taraka Rama, Çağrı Çöltekin, and Pavel Sofroniev. 2017. Computational analysis of gondi dialects. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 26–35.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei. 2015. Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771.

Ger Reesink, Ruth Singer, and Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology*, 7:e1000241.

Caley Smith. 2017. The dialectology of Indic. In Jared Klein, Brian Joseph, and Matthias Fritz, editors, *Handbook of Comparative and Historical Indo-European Linguistics*, pages 417–447. De Gruyter, Berlin, Boston.

Jaroslav Strnad. 2013. *Morphology and Syntax of Old Hindi*. Brill, Leiden.

Kaj Syrjänen, Terhi Honkola, Jyri Lehtinen, Antti Leino, and Outi Vesakoski. 2016. Applying population genetic approaches within languages: Finnish dialects as linguistic populations. *Language Dynamics and Change*, 6:235–283.

Paul Tedesco. 1965. Turner's Comparative Dictionary of the Indo-Aryan Languages. *Journal of the American Oriental Society*, 85:368–383.

Matthew Toulmin. 2009. *From linguistic to sociolinguistic reconstruction: the Kamta historical subgroup of Indo-Aryan*. Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University, Canberra.

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

Ralph L. Turner. 1975 [1967]. Geminates after long vowel in Indo-aryan. In *R.L. Turner: Collected Papers 1912–1973*, pages 405–415. Oxford University Press, London.

Martijn Wieling, Eliza Margaretha, and John Nerbonne. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2):307–314.