Disentangling dialects: a neural approach to Indo-Aryan historical phonology and subgrouping

Chundra A. Cathcart[1,2] **and Taraka Rama**[3]

[1]Department of Comparative Language Science, University of Zurich
[2]Center for the Interdisciplinary Study of Language Evolution, University of Zurich
[3]Department of Linguistics, University of North Texas
`chundra.cathcart@uzh.ch`, `taraka.kasicheyanula@unt.edu`

Abstract

This paper seeks to uncover patterns of sound change across Indo-Aryan languages using an LSTM encoder-decoder architecture. We augment our models with embeddings representing language ID, part of speech, and other features such as word embeddings. We find that a highly augmented model shows highest accuracy in predicting held-out forms, and investigate other properties of interest learned by our models' representations. We outline extensions to this architecture that can better capture variation in Indo-Aryan sound change.

1 Introduction

The Indo-Aryan languages, comprising Sanskrit (otherwise known as Old Indo-Aryan, or OIA) and its descendant languages, including medieval languages like Pāḷi and modern languages such as Hindi/Urdu, Panjabi, and Bangla, form a well-studied subgroup of the Indo-European language family. At the same time, many aspects of the Indo-Aryan languages' history remain poorly understood. One reason is that there are large historical gaps in the attestation of Indo-Aryan languages, making it challenging to document when certain shared innovations took place. Additionally, while the operation of sound changes are a diagnostic for subgrouping that historical linguistic often employ, Indo-Aryan languages have remained in close contact for millennia, borrowing words from each other and making it difficult to establish subgroup-defining sound laws using the traditional comparative method of historical linguistics.

While a number of large digitized multilingual resources pertaining to the Indo-Aryan languages exist, these data sets have not been widely used in studies, and our understanding of Indo-Aryan dialectology stands to benefit greatly from the application of deep learning techniques. This paper seeks to move further towards closing this gap. We use an LSTM-based encoder-decoder architecture to analyze a large data set of OIA etyma (ancestral forms) and medieval/modern Indo-Aryan reflexes (descendant forms) extracted from a digitized etymological dictionary, with the goal of inferring patterns of sound change from input/output string pairs. We use language embeddings with the goal of capturing individual languages' historical phonological behavior. We augment this basic model with additional embeddings that may help in capturing irregular patterns of sound change not captured by language embeddings; additionally, we compare the performance of these models against a baseline model that is embedding-free.

We evaluate the performance of models with different embeddings by assessing the accuracy with which held-out forms in medieval/modern Indo-Aryan languages are predicted on the basis of the OIA etyma from which they descend, and carry out a linguistically informed error analysis. We provide a quantitative evaluation of the degree of agreement between the genetic signal of each model's embeddings and a reference taxonomy of the Indo-Aryan languages. We find that a model with embeddings representing data points' language ID, part of speech, semantic profile and etymon ID predicts held-out forms that are closest to the ground truth forms, but that a tree constructed from language embeddings learned by this model shows lower agreement with a reference taxonomy of Indo-Aryan than a tree constructed on the basis of a model with only language embeddings, and that in general, the ability of our models to recapitulate uncontroversial genetic signal is mixed. Finally, we carry out experiments designed to investigate the information captured by specific embeddings used in our models; we find that our models learn meaningful information from augmented representations, and outline directions for future research.

Proceedings of the 24th Conference on Computational Natural Language Learning, pages 620–630
Online, November 19-20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

2 Background: Indo-Aryan dialectology

Despite a long history of scholarship, there is no general consensus regarding the subgrouping of Indo-Aryan languages comparable to that regarding other branches of Indo-European, such as Slavic or Germanic. Scholars argue for a core-periphery (Hoernle, 1880; Grierson, 1967 [1903-28]; Southworth, 2005; Zoller, 2016) or East-West split between the languages (Montaut, 2009, 2017), or are agnostic to the higher-order subgrouping of Indo-Aryan, given the many challenges involved in establishing such groups (for discussion, see Southworth 1964; Jeffers 1976; Masica 1991; Toulmin 2009). Disagreement between these groups stems largely from the fact that the different hypotheses are based on different linguistic features, and there is no agreed upon way in which to establish that individual features shared across languages are inherited from a common ancestor rather than due to parallel innovation. The traditional comparative method of historical linguistics (Hoenigswald, 1960; Weiss, 2015) tends to establish linguistic subgroups on the basis of innovations in morphology as well as shared sound changes, some of which are thought to be unlikely to operate independently. Indeed, many scholars have agreed that Indo-Aryan subgrouping should be established according to sound change; however, the establishment of regular sound changes has proved challenging given the high degree of irregularity in the data (Masica, 1991). Our method has the potential to detect regularities and bear on the questions described above.

3 Related work

Traditional computational dialectology (Kessler, 1995; Nerbonne and Heeringa, 2001) identifies dialect clusters using edit distance; more recent work uses neural architectures for dialect classification based on social media data for languages such as English (Rahimi et al., 2017b,a) and German (Hovy and Purschke, 2018). Computational methods have been applied to the related field of historical linguistics to identify cognates (words that go back to a common ancestor) and infer relationships between languages (Rama et al., 2018) as well as the reconstruction of ancestral words through Bayesian methods (Bouchard-Côté et al., 2013), gated neural networks (Meloni et al., 2019) and non-neural sequence labeling methods (Ciobanu and Dinu, 2020).

Other recent work infers language embeddings from large parallel corpora using different neural architectures (Östling and Tiedemann, 2017; Johnson et al., 2017; Tiedemann, 2018; Rabinovich et al., 2017). These embeddings tend to produce hierarchical clustering configurations that are close to the language classification trees inferred from historical linguistic research. These claims have been tested by Bjerva et al. (2019) who find that the distances between learned language representations may not be reflective of genetic relationship but of structural similarity. It is not always straightforward to interpret the sources of differentiation among these embeddings; typically, embeddings based on synchronic patterns of language use in corpora may be due to word order patterns, phonotactic patterns, or a number of other interrelated language-specific distributions. Cathcart and Wandl (2020) investigate the patterns of sound change captured by a neural encoder-decoder architecture trained on Proto-Slavic and contemporary Slavic word forms, and find that embeddings dispay at least partial genetic signal, but also note a negative relationship between overall model accuracy and the degree to which embeddings reflect the *communis opinio* subgrouping of Slavic.

4 Data and rationale for model design

We use data from an etymological dictionary of the Indo-Aryan languages (Turner, 1962–1966).[1] We extract OIA etyma and their corresponding reflexes in medieval and modern Indo-Aryan languages (e.g., OIA *vākya* 'speech, words' develops to Pāḷi *vākya*, Kashmiri *wākh*, etc.). As the traditional Indological orthography used to transcribe forms in the dictionary is phonemic, we retain this representation and convert characters with diacritics to a Normalization Form Canonical Decomposition (NFD) Unicode representation in order to reduce the number of input and output character types. Additionally, we extract glosses provided for OIA etyma (at the time of writing, the extraction of reflex glosses cannot be straightforwardly automated due to the unstructured nature of the markup language, plus the absence of glosses for certain reflexes). We match languages in the dictionary with the closest matching glottocode from the Glottolog database (Hammarström et al., 2017), and omit languages with fewer than 100 entries. This results in a data set of 82431 forms in 61 languages; the number of

[1] Online at `https://dsalsrv04.uchicago.edu/dictionaries/soas/`

forms in each language can be seen in Table 1. The most frequent language is the medieval language Prakrit, followed by Hindi, the medieval language Pāḷi, Marathi, and Panjabi.

As mentioned above, a goal of this study is to employ language embeddings in a neural model in order to capture language-level regularities in sound change from which genetic information can be extracted. However, there are many factors in our data set that lead to irregularity in sound change. Some irregularity is due to contact between Indo-Aryan languages (Turner, 1975 [1967]) as well as analogical change; other instances of irregularity are due to artifacts of the way that data are presented in the etymological dictionary. A key source of systematic morphological non-congruence is the fact that for verbal forms, the OIA third-person present singular is often paired with modern Indo-Aryan infinitives. For instance, OIA *vaś**ati*** 'wishes, wills' is paired with reflexes such as Assamese *bahā**iba*** (non-cognate verbal endings are in bold), whereas a non-verbal form with a similar ending, OIA *ūnaviṁśati* 'nineteen', is paired with reflexes such as Assamese *unaix*, which does not contain a morphological mismatch. We do not wish for our our system to learn that the first pattern is a sound change. For this reason we code OIA etyma according to whether or not they are verbal, potentially allowing our system to disentangle morphological mismatches from legitimate sound changes.

A more interesting and poorly understood point (that is not merely an artifact of the data) is that etyma with certain semantic profiles may be more prone to certain analogical changes. For instance, nouns of certain semantic fields may be more likely to receive a diminutive suffix, which may then be reanalyzed as part of the noun stem; additionally, semantically related nouns are known to undergo analogical contamination (Malkiel, 1962) or develop specific patterns of sound symbolism (Carling and Johansson, 2014; Blasi et al., 2016). Finally, particular etyma may favor a particular "prototype" showing specific patterns of sound change. An example of this phenomenon can be seen in reflexes of OIA *vismarati* 'forgets'. OIA *sm* usually changes to *m(h)* or *s(s)* in descendant languages; however, only one reflex of *vismarati* shows *m(h)* (Prakrit *vimharai*), while the rest show *s(s)*. It is possible that an early variant **visarati* was diffused among neighboring dialects and an early date. All in all, while we do not explicitly model contact, accounting for the factors described above can improve model accuracy and allow us to tease apart legitimate patterns of sound change from orthogonal factors.

In order to achieve this goal, we augment a basic model using language embeddings with different embedding types designed to account for idiosyncrasies of data collection as well as potential real-world sources of irregularity like those described above. We make use of embeddings that represent the part of speech (POS) of the OIA etymon. Additionally, we represent the semantic profile of each OIA etymon by generating embeddings of each etymon's English language gloss using a pretrained BERT model (Devlin et al., 2019; Wolf et al., 2019), though this approach does not fully encapsulate the OIA word's semantics. Finally, we wish to take into account idiosyncratic patterns displayed by individual etyma (such as *vismarati*, as discussed above). A one-hot encoding of etymon IDs is costly, as there are 13580 unique etyma in our dataset; instead, we combine information from BERT embeddings and the input string in order to produce a unique embedding for each etymon in the data set. In sum, these augmentations provide a way for our model to disentangle the orthogonal forces of sound change and other factors.

5 Model

Our experiments use an LSTM Encoder-Decoder with 0th-order nonmonotonic hard attention (Wu and Cotterell, 2019). The authors' architecture works as follows: for each input x (for our purposes an OIA etymon), a latent representation $\boldsymbol{h}_j^{\text{enc}} \in \mathbb{R}^{2D}$ is learned for each time step $j \in \{1, ..., |x|\}$ via a bidirectional LSTM on the basis of the input symbol at time step j. For each output y (for our purposes a medieval/modern Indo-Aryan reflex), a latent representation $\boldsymbol{h}_i^{\text{dec}} \in \mathbb{R}^D$ is learned via a forward LSTM for each time step $i \in \{1, ..., |y|\}$ on the basis of the output symbol at time step $i - 1$. The probability that the output is aligned with the jth input symbol at time i is equal to $\text{softmax}(\boldsymbol{h}_i^{\text{dec}\top} \boldsymbol{T} \boldsymbol{h}_j^{\text{enc}})$, where $\boldsymbol{T} \in \mathbb{R}^{D \times 2D}$ is a learned parameter. The emission probability of the output symbol at time i given such an alignment is proportional to $\exp(\boldsymbol{W} \tanh(\boldsymbol{S}[\boldsymbol{h}_i^{\text{dec}}; \boldsymbol{h}_j^{\text{enc}}]))$, and is hence also dependent on the previous output symbols ($\boldsymbol{W} \in \mathbb{R}^{\Sigma_y \times 3D}$ and $\boldsymbol{S} \in \mathbb{R}^{3D \times 3D}$ are learned parameters). Structural zeros are used in order to ensure that the alignment between the input and

Language	Glottocode	N
Prakrit(Maharashtri)	maha1305	8118
Hindi	hind1269	5948
Pali	pali1273	5225
Marathi	mara1378	4895
EasternPanjabi	panj1256	4622
Gujarati	guja1252	4490
Sindhi	sind1272	4020
Odia	oriy1255	3925
Nepali	nepa1254	3807
Sinhala	sinh1246	3791
Bengali	beng1280	3109
WesternPanjabi	west2386	3060
Kumaoni	kuma1273	2857
Kashmiri	kash1277	2659
Assamese	assa1263	2543
Maithili	mait1250	1466
Shina	shin1264	1152
Bagheli	bagh1251	1086
Bhadrawahi	bhal1244	814
Khowar	khow1242	797
Kachchi	kach1277	789
Dhivehi	dhiv1236	775
Bhojpuri	bhoj1244	750
Konkani	konk1267	672
Garhwali	garh1243	672
Phalura	phal1254	648
Awadhi	gang1265	607
Dameli	dame1241	607
Bhadrawahi	bhad1241	602
NortheastPashayi	nort2666	525
Gawar-Bati	gawa1247	520
Kalami	kala1373	488
Kalasha	kala1372	407
VlaxRomani	vlax1238	397
MahasuPahari	maha1287	381
Torwali	torw1241	374
Kalasha	sout2669	329
Shumashti	shum1235	316
NorthwestPashayi	laur1248	292
Jaunsari	jaun1243	269
Wotapuri-Katarqalai	wota1240	258
Domari	nawa1257	245
NortheastPashayi	aret1240	243
Domaaki	doma1260	239
IndusKohistani	indu1241	224
Savi	savi1242	207
Tirahi	tira1253	186
KohistaniShina	kohi1248	186
Churahi	chur1258	181
Marwari(India)	marw1260	166
NorthwestPashayi	nort2665	148
NortheastPashayi	kura1247	147
Lomavren	loma1235	146
WesternPanjabi	mult1243	143
Chambeali	cham1307	142
NortheastPashayi	wega1238	140
WelshRomani	wels1246	138
Braj	braj1242	135
SoutheastPashayi	sout2672	130
Khetrani	khet1238	120
Pangwali	pang1282	103

Table 1: Number of reflexes for languages in the data

output string is strictly monotonically increasing.[2]

In our experiments, we concatenate embeddings encoding the features described in the previous section to our input at each time step, namely (L)anguage ID, (P)art of speech, (S)emantic profile, and (E)tymon. We use a one-hot encoding of language ID and POS ID, and employ BERT embeddings (reduced from 768 to 128 dimensions using principal component analysis) to represent an etymon's semantic profile. Embeddings for etyma are represented by contatenating the first and last states of a Bidirectional LSTM encoding of the etymon string to the BERT-based semantic embedding (denoted by $\boldsymbol{e}(\mathsf{gloss}_i)$)). Formally, these embeddings consist of the following, for a given data point index $i \in \{1, ..., |\mathsf{data}|\}$:

- L: $z_i^{\mathsf{lang}} = \text{MLP}(\mathsf{lang}_i)$
- P: $z_i^{\mathsf{POS}} = \text{MLP}(\mathsf{POS}_i)$
- S: $z_i^{\mathsf{sem}} = \text{MLP}(\boldsymbol{e}(\mathsf{gloss}_i))$
- E: $z_i^{\mathsf{etym}} = \text{MLP}([\text{MLP}([\text{LSTM}(x_{i,1:|x_i|})_{|x_i|}; \text{LSTM}(x_{i,|x_i|:1})_{|x_i|}]); z_i^{\mathsf{sem}}])$

After one or more of these embeddings are concatenated to an input token, the resulting concatenation is passed to another MLP layer, which is then fed to the encoder-decoder architecture. We increment our models by concatenating embeddings to the input in a stepwise fashion, yielding four models (L, LP, LPS, LPSE). Additionally, we compare our models against a baseline that does not use any embeddings.

We set the embedding and hidden layer dimension size to 128. We carry out K-fold ($K = 8$) cross-validation to assess model accuracy, training our models on mini-batches of 64 data points for a maximum of 200 epochs, validating on 10% of the training data and stopping early if the validation loss does not decrease over twenty consecutive epochs. We use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of .001. We greedily decode test data using the fitted models, generating each output token on the basis of the previous generated token.[3]

[2] Although Wu and Cotterell (2019) report superior performance of a 1st-order hard monotonic model that penalizes alignments which jump more than w time steps; we did not implement this model since we could not make a principled decision regarding the value of w.

[3] Code and results are available at https://github.com/chundrac/ia-conll-2020

6 Results

We assess the accuracy with which our models predict held-out medieval/modern Indo-Aryan forms on the basis of the corresponding OIA etymon input by measuring the phoneme error rate (PER), which we define as the Levenshtein distance between the predicted and true form divided by the length of the longer string (normalized Levenshtein distance), and the word error rate (WER), or the proportion of held-out forms where one or more errors occurs in the predicted form. Mean PER and WER values are found for each model in Table 2.

As expected, the baseline model performs the worst according to these metrics. Of the non-baseline models, the model with language and POS embeddings shows the worst overall performance; highly augmented models such as the model which uses language, POS, semantic and etymon embeddings shows the best performance in terms of PER. However, the model with only language embeddings shows the best performance in terms of WER, indicating that the LPSE model introduces errors in a higher number of individual predicted words even if it produces fewer errors overall. We carry out pairwise Wilcoxon signed-rank tests to assess the significance of differences in PER between models, using the Bonferroni correction for multiple comparisons; all between-model differences are highly significant or significant, with the exception of the difference between the L and LPS models.

Figure 1 displays the relationship between language-level PER and the number of training examples for a given language. The correlation between these two variables is negative and significant (Spearman's ρ is between $-.39$ and $-.54$) for all models. However, as shown by the lines of best fit plotted in the figure, this correlation is considerably weaker for the baseline model than for the other models. Interestingly, the omission of language embeddings seems to have resulted in higher error rates for languages with larger numbers of training examples; if no information regarding language ID is fed to the encoder-decoder, there seems to be no way to keep highly influential languages from interfering in the patterns learned for other highly influential languages.

Model	PER	WER
L	.257	**.808**
LP	.262	.818
LPS	.256	.813
LPSE	**.255**	.809
Baseline	.346	.940

Table 2: Phoneme error rates and word error rates for each model. All pairwise PER comparisons between models are highly significant ($p < 0.001$) according to a Wilcoxon signed-rank test with Bonferroni correction, with the exception of L–LPS ($p = 1$) and LPS–LPSE ($p = 0.02$).

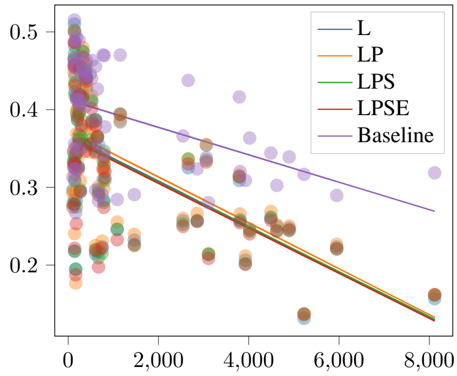

Figure 1: Language-level PER plotted by the number of training examples for each language, for each model.

7 Error analysis

PER based on unweighted Levenshtein distance is agnostic to error type. In a task such as ours, some error types will indicate strongly that our models have failed to learn meaningful generalizations regarding sound change. Additionally, certain errors produced by a model may be due to errors in the data when the model has in fact learned a meaningful pattern of change.

With this in mind we turn to a quantitative but linguistically informed error analysis designed to investigate the types of errors made by each model and assess the degree to which each model makes certain errors. One error type consists of errors in vowel quality or quantity: the LPS model erroneously generates Kashmiri *muhun* (< OIA *mōháyati* 'bewilders') as *mŏhun*. Errors of this sort affect consonants as well: the Pāḷi form *addana* (< OIA *ardana* 'tormenting') is erroneously generated as *aḍḍana* by the LP model. Errors of this sort, particularly vowel type errors, often occur at the right word edge, perhaps due to confusion resulting from the restructuring of the OIA case and gender systems in many Indo-Aryan languages; e.g., Sindhi *vali* (< OIA *vallī* 'creeper') is generated as *vala*

Language	Glottocode	True	L	LP	LPS	LPSE	Baseline
Bagheli	bagh1251	machlī	machilā	mā̃s	machā	māchī	macch
Jaunsari	jaun1243	māchā	māch	mā̃ś	bhēċū	māchā	macch
Khowar	khow1242	maċí	muċh	mač	máču	mučh	macch
VlaxRomani	vlax1238	mačo	māch	machaṛ	māċh	maċhi	macch
Bhadrawahi	bhal1244	maċhli	maċh	machli	maċhi	machli	māchā
Bhadrawahi	bhad1241	machlī	meċhlī	meċhl	machlī	meċhlī	mācho
Kachchi	kach1277	macch	machī	machi	machi	machī	mācho
Odia	oriy1255	mācha	mācha	mācha	macha	mācha	mācho
Pali	pali1273	maccha	maccha	maccha	maccha	maccha	mācho
EasternPanjabi	panj1256	macch	masch	macch	macchā	macchā	mācho
Dhivehi	dhiv1236	mas	mais	mais	mahi	mati	māch
Hindi	hind1269	machlī	māch	māch	māch	māch	māch
Konkani	konk1267	māsḷī	māsi	māċ	māċa	māċu	māch
Sindhi	sind1272	machu	machu	machu	machu	machu	māch
NortheastPashayi	aret1240	māċ	mȭċ	māčī	mȭč	mačoṭ	māch
Hindi	hind1269	māch	machlī	māch	machlī	machī	māch
Marathi	mara1378	māsā	māċh	mās	mās	mās	māch
Bengali	beng1280	māch	māchlā	meċā	māch	māchā	macha
Gawar-Bati	gawa1247	maċoṭá	māċh	māċ	māċ	maċ	macha
Prakrit(Maharashtri)	maha1305	maccha	maccha	maccha	maccha	maccha	macha
Sinhala	sinh1246	masā	mas	masa	mas	masā	macha
Bhadrawahi	bhad1241	machli	machlī	maċhlī	machlī	machlī	machī
Garhwali	garh1243	māchu	māchu	maċhlu	māchī	māchu	machī

Table 3: Selected held-out forms generated on the basis of OIA *mátsya* 'fish' for several languages. The true held-out form is presented alongside forms generated by the L, LP, LPS, LPSE and baseline models.

by the LP and LPS models. Some errors involve excrescence or insertion, where extra phonological information is erroneously produced in the predicted form, e.g., LPS *kiriruvalaṇa* for Sinhala *kiriväla* (< OIA *kṣīravallī* 'Batatus paniculata'). Elsewhere, we find erroneous deletion of phonological material, e.g. LPS *lūṛh* in place of Bagheli *loṛhnihār* (< OIA *luṭháti* 'rolls, wallows'). Some errors involve the generation of output that is phonemically analyzable as the ground truth form, e.g., L *kachwā* for Hindi *kachuā* (< OIA *kacchapa* 'turtle, tortoise'). Along with source errors or morphological mismatches that simply cannot be detected by our model architecture, the errors mentioned above make up the bulk of errors produced by models.

We align tokens in held-out forms with tokens in predicted forms using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), allowing us to automatically extract errors made by each model (e.g., instances where *ḷ* was generated instead of *l*). We classify mismatches between held-out and predicted tokens according to whether they involve an insertion, a deletion, a change affecting a vowel (other than deletion), or a change affecting a consonant (other than deletion). Proportions of these changes across models are given in Table 4. As can be observed, error type rates are similar across models; however, the baseline model has lower rates of erroneous deletion than other models, and higher rates of incorrect substitutions affecting consonants. A possible reason for this behavior is that the baseline model, which lacks language embeddings, likely comes under influence from Prakrit, which contains the most training examples in the data set (well-attested languages like Hindi contain fewer training examples due to partial replacement of vocabulary inherited from OIA by Persian and Arabic loanwords). Prakrit did not undergo vowel or consonant deletion to the extent that later-attested Indo-Aryan languages did; hence, the overwhelming influence of this language on the model would make deletion a less likely change overall. Prakrit did however undergo drastic changes to consonants, such as full assimilation of clusters other than nasal-plosive sequences; if the model is influenced by this behavior, it may account for some of the instances of incorrect consonant substitution not seen in the other models. As an example, Marathi *khābārī* (< OIA *kārṣmaryā* 'the tree Gmelina arborea') is incorrectly predicted to be *khā̃varī* by the LP model, but the baseline model produces the more conservative *kāmbhārī*.

Selected held-out reflexes of OIA *mátsya* 'fish' are provided in Table 3 along with their predicted counterparts for each model, illustrating the challenges that our models face when predicting held-out forms. The reflexes provided all descend from *mátsya*, but some have gained extra morphology during their development, most frequently the suffix *-la*, a morpheme added to a number of medieval/modern Indo-Aryan nouns. For morphological irregularities of this type, models endowed with semantic information have the potential to infer that certain semantically related nouns acquire the suffix *-la* during their development; at the same time, the L model, which lack semantic information, generates forms reflecting a *-la* suffix, which may indicate that it has learned certain phonotactic patterns from the target-side language model, which is shared across all languages. The baseline model consistently produces a limited number of reflex types, likely informed by the most frequent languages in the sample (i.e., Prakrit and Hindi).

	I	D	V	C
L	0.19	0.24	0.35	0.22
LP	0.20	0.22	0.35	0.22
LPS	0.20	0.23	0.35	0.22
LPSE	0.20	0.23	0.35	0.22
B	0.20	0.19	0.36	0.26

Table 4: Proportions of error types (erroneous insertion [I], deletion [D], vowel substitution [V], consonant substitution [C]) produced by each model.

8 Genetic signal

We investigate the degree to which the language-level embeddings learned by our models represent the genetic relatedness of Indo-Aryan languages in our sample. For the language embeddings produced by each model, we compute the cosine distances between each pair of embeddings, and use these distances to construct language trees using neighbor joining (NJ, Saitou and Nei, 1987). We compare each tree to a reference taxonomy of the Indo-Aryan languages taken from Glottolog (Hammarström et al., 2017), which contains relatively uncontroversial language groupings but does not resolve all subgroupings, and hence contains numerous polytomies (i.e., non-binary branchings).

We measure the distance between two language trees is measured using Generalized Quartet Distance (GQD, Pompei et al., 2011; Rama et al., 2018). A quartet in a phylogenetic language tree consists of four languages and can either be resolved ("butterfly") or unresolved ("star"). The generalized quartet distance is the ratio of the number of butterfly quartets which differ across trees to the total number of butterfly quartets in the reference tree. While our reference tree is non-binary, the trees inferred from our models' embeddings are binary. Accordingly, the GQD measure does not penalize the inferred tree, ignoring the star quartets found in the reference tree. The GQD scores for all our models are given in Table 5, along with scores for a baseline tree constructed using averaging the normalized Levenshtein distance (LDN) between cognate forms for pairs of languages (lower values indicate greater agreement).

Model	GQD
L	0.509
LP	0.559
LPS	0.514
LPSE	0.569
LDN	**0.304**

Table 5: GQD between the inferred tree and the reference tree for each model

Notably, despite its good performance according to the PER metric, the embeddings produced by the LPSE model show the lowest agreement with a reference taxonomy of Indo-Aryan out of all of the models used in our experiments, particularly when compared with the model that uses only language embeddings. A possible explanation is that by including additional embeddings in our models designed to capture different patterns of sound change in different morphological, semantic and etymological scenarios, we have filtered out critical information relevant to subgrouping, removing valuable genetic signal displayed by morphological traits, which may explain why the model with language embeddings outperforms the other models. A similar negative relationship between model accuracy and genetic signal displayed by embeddings was found by Cathcart and Wandl (2020).

At the same time, all models are significantly outperformed by the LDN tree, indicating that string distances between contemporary forms capture inter-language relationships at more levels of granularity than the distances computed from the embeddings learned by our models. All models, including the LPSE model, are successful at learning patterns of change within individual lineages and recapitulating shallow subgroups. For visualization, we map LPSE embeddings to three-dimensional space using multidimensional scaling

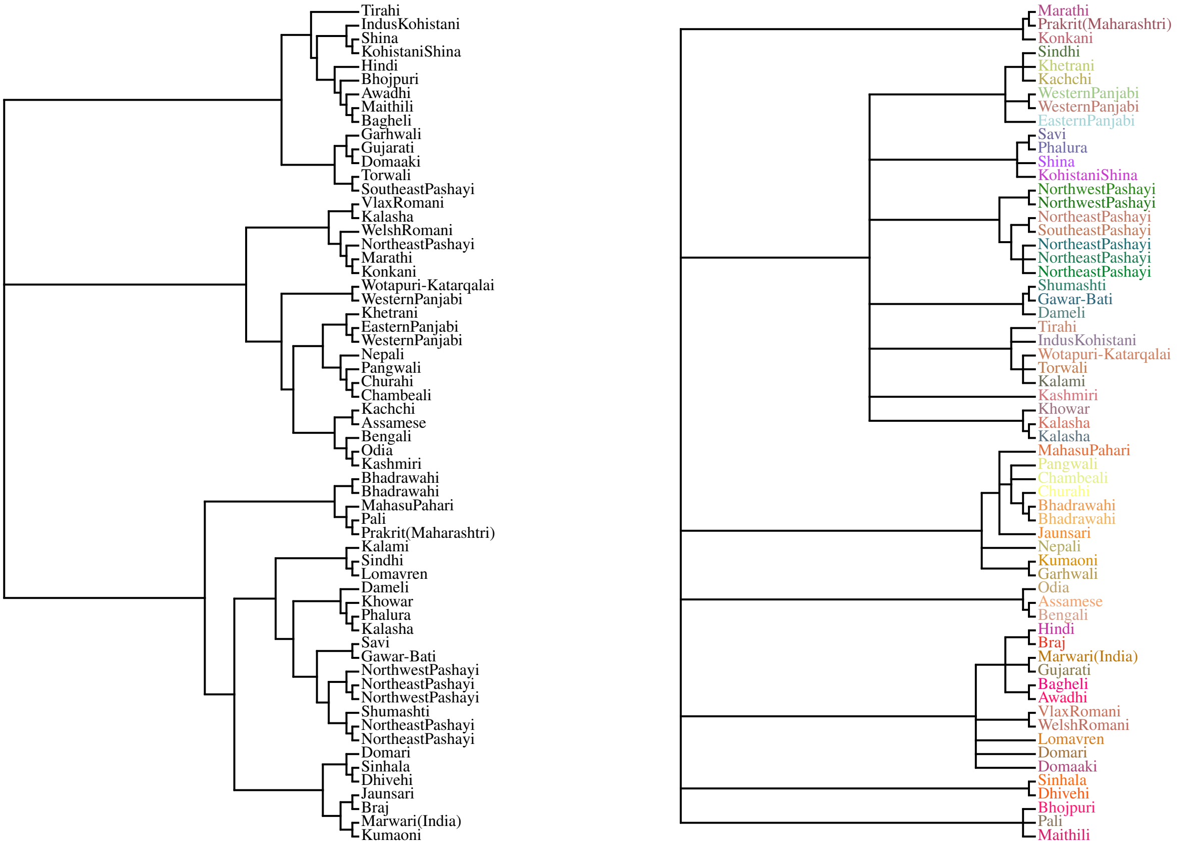

Figure 2: Tree constructed from LPSE model embeddings using neighbor joining (left) and reference taxonomy of the Indo-Aryan languages in our sample, taken from Glottolog (right). Colors of languages represent the positions of their respective language embeddings from the LPSE model in three-dimensional space.

and color the taxa of the reference tree using RGB color vectors produced by these normalized values, as shown in Figure 2. Similarly colored taxa are close to each other in embedding space. In general, very close relatives appear to occupy the same region of embedding space, but small groups within large taxa may be distant from each other in the same space, indicating premature but not altogether unpromising results for our attempt to capture genetic signal in the data we analyze.

9 Representations learned by embeddings

Here, we investigate whether our architecture has learned properties of the different patterns displayed by verbal and non-verbal forms. OIA verbs end in *-ti* and *-tē*; virtually no OIA nouns end in *-tē*, and while some OIA nouns end in *-ti*, they tend to have a different phonotactic profile from OIA verb forms. Hence, it is entirely possible that encoder-decoder models with embeddings that encode part of speech ignore this information and can simply learn mappings such as OIA *-ati* > Assamese *-iba* (e.g., from OIA *vaśati* : Assamese *bahāiba*), since *-iba* is a suffix that freqently co-occurs with the OIA sequence *-ati*. The fact that a model with only language embeddings outperforms a model with language embeddings and part-of-speech embeddings provides a reason to suspect that our architecture does not learn anything about the morphological mismatches found between OIA verb citation forms and medieval/modern Indo-Aryan forms, simply because it does not need to — surface patterns in phonology may be sufficient to learn the correct mapping between input and output.

To investigate whether our models have learned anything from part-of-speech embeddings, we carry out an experiment where we feed held-out data to the LP, LPS, and LPSE models, holding the language ID and semantic and etymological information constant but perturbing the POS ID (i.e., changing it to VERB for non-verbal data and changing it to NON-VERB for verbal data). We measure the normalized Levenshtein distance between forms decoded with true POS IDs and forms decoded with perturbed POS IDs, and average these PER values for verbal and non-verbal forms. As shown in Table 6, changing a verbal POS ID to a non-verbal one results in a more dissimilar form, whereas the opposite change has less of an effect (all differences are highly significant according to a Mann-Whitney test, $p < .001$), since our system never encounters certain noun suffixes with a verbal POS ID and hence doesn't learn variable patterns for these suffixes. Differences are most pronounced for the LP model, but when more information is concatenated to the input, the effect of using an embedding for POS ID appears to be absorbed by the other embeddings, and the disparity levels out. This indicates that our augmented models do learn meaningful information from the POS ID, but that models endowed with semantic information learn more fine-grained patterns and produce more accurate results than the LP model.

	LDN		Match	
Model	N→V	V→N	N→V	V→N
LP	0.226	0.291	0.604	0.535
LPS	0.189	0.204	0.668	0.619
LPSE	0.170	0.220	0.700	0.628

Table 6: Average normalized Levenshtein distance between predicted held-out forms with unperturbed and perturbed POS IDs ([N]on-verb and [V]erb), grouped by model and POS (left); average proportion of leftmost matching segments found between these pairs of forms.

In the majority of Indo-Aryan languages, verbal morphology (such as the infinitive ending) consists of suffixes with straightforwardly affixal behavior (Masica, 1991, 321ff.); stem alternations are rare outside of valency changing processes. It is therefore worthwhile to know the extent to which our perturbations reflect this quality: do the differences in output caused by perturbing the POS ID accumulate at the right word edge, or are they distributed throughout the word? In order to investigate this question, we tabulate the length of the leftmost matching substring across each output string and its perturbed counterpart, and divide this number by the mean of the two strings; higher values indicate that the two strings differ only according to a suffix. As shown in the right half of Table 6, mean values for this quantity tend to be greater than .5, indicating that perturbing the POS ID tends to result in different suffixation. Further investigation of these distributions and metrics designed to capture this quality is outside the scope of this paper but will provide stimulating future research.

10 Discussion and Outlook

In this paper, we investigated the ability of LSTM-based encoder-decoder architectures to capture recurrent patterns of sound change between OIA and medieval/modern Indo-Aryan languages, as well as encode information regarding the genetic relationships between languages. We found that a model augmented with information regarding forms' language ID, POS ID, semantic profile, and etymon ID showed the lowest phoneme error rate out of all models, but that language embeddings learned by this model showed low agreement with a reference taxonomy of Indo-Aryan, and that in general, our models struggled to capture uncontroversial genetic signal. This issue may be in part due to architectural choices we made, along with challenging aspects of the data set we used, drawn from a diverse group of languages.

Our experiments show that the use of different embeddings appears to allow our models to learn deviations from regular sound change that are found in words with certain parts of speech, semantic profiles, or that reflect particular etyma. At the same time, there are many avenues for improving the performance of models on this highly challenging data set. In future work, we plan to obtain glosses for reflex forms in the dictionary in order to determine whether the semantic distance between an etymon and a reflex can capture vagaries of analogical change that we were unable to model in this paper. Additionally, our models learned embeddings for POS IDs that did not vary across languages, rather than language-specific ones, and we built our models incrementally in a stepwise function rather than considering all possible subsets of predictors of interest, decisions that may have influenced our results. Greater flexibility will play an important role in future deep learning approaches that hope to capture the multifaceted diachronic processes that yield synchronic linguistic similarity and dissimilarity.

Acknowledgements

We thank anonymous CoNLL reviewers for helpful comments, and Sidaarth Sabhnani for assistance with data extraction.

References

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019. What do language representations really represent? *Computational Linguistics*, 45(2):381–389.

Damián E. Blasi, Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences*, 10.1073/pnas.1605782113.

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.

Gerd Carling and Niklas Johansson. 2014. Motivated language change: processes involved in the growth and conventionalization of onomatopoeia and sound symbolism. *Acta Linguistica Hafniensia*, 46(2):199–217.

Chundra Cathcart and Florian Wandl. 2020. In search of isoglosses: continuous and discrete language embeddings in Slavic historical phonology. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 233–244, Online. Association for Computational Linguistics.

Alina Maria Ciobanu and Liviu P Dinu. 2020. Automatic identification and production of related words for historical linguistics. *Computational Linguistics*, 45(4):667–704.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George A. Grierson. 1967 [1903-28]. *Linguistic Survey of India*. Motilal Banarsidass, Delhi.

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

Henry M. Hoenigswald. 1960. *Language change and linguistic reconstruction*. University of Chicago Press, Chicago.

A. F. Rudolf Hoernle. 1880. *A comparative grammar of the Gaudian languages*. Trübner and Co., London.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Robert J Jeffers. 1976. The position of the Bihārī dialects in Indo-Aryan. *Indo-Iranian Journal*, 18(3-4):215–225.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Brett Kessler. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 60–67, Dublin. EACL.

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

Yakov Malkiel. 1962. Weak phonetic change, spontaneous sound shift, lexical contamination. *Lingua*, 11:263–275.

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

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2019. Ab antiquo: Proto-language reconstruction with rnns. *arXiv preprint arXiv:1908.02477*.

Annie Montaut. 2009. Ergative and pre-ergative patterns in Indo-Aryan as predications of localization.

Annie Montaut. 2017. Grammaticalization of participles and gerunds in Indo-Aryan: Preterite, future, infinitive. *Unity and diversity in grammaticalization scenarios*, 16:97.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–53.

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

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 644–649. Association for Computational Linguistics.

Simone Pompei, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PloS one*, 6(6):e20109.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada. Association for Computational Linguistics.

Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. 2017a. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 167–176.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2017b. A neural model for user geolocation and lexical dialectology. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 209–216.

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400.

N. Saitou and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Franklin Southworth. 1964. Family-tree diagrams. *Language*, 40(4):557–565.

Franklin C. Southworth. 2005. *Linguistic Archaeology of South Asia*. Routledge, London.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. In Eetu Mäkelä, Mikko Tolonen, and Jouni Tuominen, editors, *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference (DHN 2018)*, pages 188–197.

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.

Michael Weiss. 2015. The comparative method. In Claire Bowern and Bethwyn Evans, editors, *The Routledge handbook of historical linguistics*, pages 127–145. Routledge, London.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy. Association for Computational Linguistics.

Claus-Peter Zoller. 2016. Outer and Inner Indo-Aryan, and northern India as an ancient linguistic area. *Acta Orientalia*, 77:71–132.