Computational analysis of Gondi dialects

Taraka Rama and **Çağrı Çöltekin** and **Pavel Sofroniev**
Department of Linguistics
University of Tübingen, Germany
taraka-rama.kasicheyanula@uni-tuebingen.de
cagri.coeltekin@sfs.uni-tuebingen.de
pavel.sofroniev@student.uni-tuebingen.de

Abstract

This paper presents a computational analysis of Gondi dialects spoken in central India. We present a digitized data set of the dialect area, and analyze the data using different techniques from dialectometry, deep learning, and computational biology. We show that the methods largely agree with each other and with the earlier non-computational analyses of the language group.

1 Introduction

Gondi languages are spoken across a large region in the central part of India (cf. figure 1). The languages belong to the Dravidian language family and are closely related to Telugu, a major literary language spoken in South India. The Gondi languages received wide attention in comparative linguistics (Burrow and Bhattacharya, 1960; Garapati, 1991; Smith, 1991) due to their dialectal variation. On the one hand, the languages look like a dialect chain while, on the other hand, some of the dialects are shown to exhibit high levels of mutual unintelligibility (Beine, 1994).

Smith (1991) and Garapati (1991) perform classical comparative analyses of the dialects and classify the Gondi dialects into two subgroups: Northwest and Southeast. Garapati (1991) compares Gondi dialects where most of the dialects belong to Northwest subgroup and only three dialects belong to Southeast subgroup. In a different study, Beine (1994) collected lexical word lists transcribed in International Phonetic Alphabet (IPA) for 210 concepts belonging to 46 sites and attempted to perform a classification based on word similarity. Beine (1994) determines two words to be cognate (having descended from the same common ancestor) if they are identical in form and meaning. The average similarity between two sites is determined as the average number of identical words between the two sites. The author describes the experiments of the results qualitatively and does not perform any quantitative analysis. Until now, there has been no computational analysis of the lexical word lists to determine the exact relationship between these languages. We digitize the dataset and then perform a computational analysis.

Recent years have seen an increase in the number of computational methods applied to the study of both dialect and language classification. For instance, Nerbonne (2009) applied Levenshtein distance for the classification of Dutch and German dialects. Nerbonne finds that the classification offered by Levenshtein distance largely agrees with the traditional dialectological knowledge of Dutch and German areas. In this paper, we apply the dialectometric analysis to the Gondi language word lists.

In the related field of computational historical linguistics, Gray and Atkinson (2003) applied Bayesian phylogenetic methods from computational biology to date the age of Proto-Indo-European language tree. The authors use cognate judgments given by historical linguists to infer both the topology and the root age of the Indo-European family. In parallel to this work, Kondrak (2009) applied phonetically motivated string similarity measures and word alignment inspired methods for the purpose of determining if two words are cognates or not. This work was followed by List (2012) and Rama (2015) who employed statistical and string kernel methods for determining cognates in multilingual word lists.

In typical dialectometric studies (Nerbonne, 2009), the assumption that all the pronunciations of a particular word are cognates is often justified by the data. However, we cannot assume that this is the case in Gondi dialects since there are sig-

Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 26–35,
Valencia, Spain, April 3, 2017.

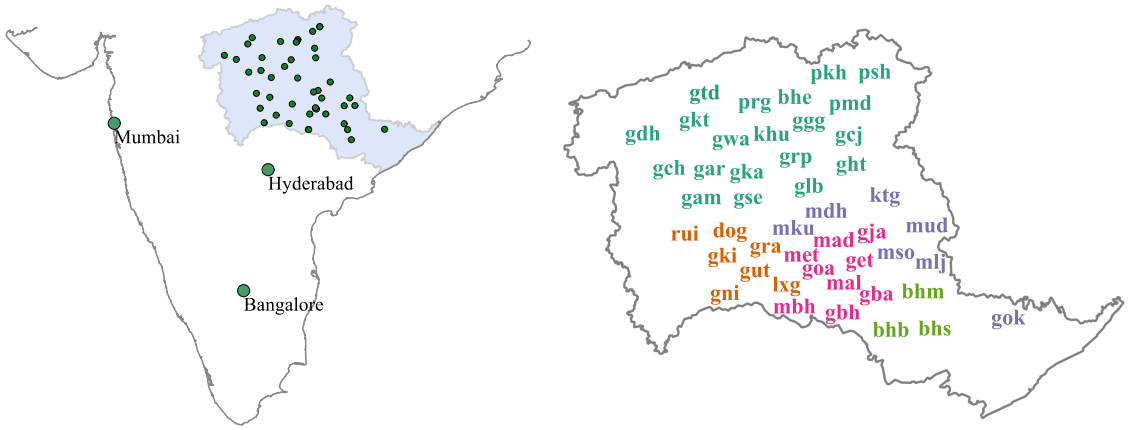

Figure 1: The Gondi language area with major cities in focus. The dialect/site codes and the geographical distribution of the codes are based on Beine (1994).

nificant amount of lexical replacement due to borrowing (from contact) and internal lexical innovations. Moreover, the previous comparative linguistic studies classify the Gondi dialects using sound correspondences and lexical cognates. In this paper, we will use the Pointwise Mutual Information (Wieling et al., 2009) method for obtaining sound change matrices and use the matrix to automatically identify cognates.

The comparative linguistic research classified the Gondi dialects into five different genetic groups (cf. table 1). However, the exact branching of the Gondi dialects is yet a open question. In this paper, we apply both dialectometric and phylogenetic approaches to determine the exact branching structure of the dialects.

The paper is organized as followed. In section 2, we describe the dataset and the gold standard dialect classification used in our experiments. In section 3, we describe the various techniques for computing and visualizing the dialectal differences. In section 4, we describe the results of the different analyses. We conclude the paper in section 5.

2 Datasets

The word lists for our experiments are derived from the fieldwork of Beine (1994). Beine (1994) provides multilingual word lists for 210 meanings in 46 sites in central India which is shown in figure 1. In the following sections, we use the Glottolog classification (Nordhoff and Hammarström, 2011) as gold standard to evaluate the various analyses. Glottolog is a openly available resource that summarizes the genetic relationships of the world's dialects and languages from published scholarly linguistic articles. For reference, we provide the Glottolog classification[1] of the Gondi dialects in table 1. The Glottolog classification is derived from comparative linguistics (Garapati, 1991; Smith, 1991) and dialect mutual intelligibility tests (Beine, 1994).

Dialect codes	Classification
gdh, gam, gar, gse, glb, gtd, gkt, gch, prg, gka, gwa, grp, khu, ggg, gcj, bhe, pmd, psh, pkh, ght	Northwest Gondi, Northern Gondi
rui, gki, gni, dog, gut, gra, lxg	Northwest Gondi, Southern Gondi
met, get, mad, gba, goa, mal, gja, gbh, mbh	Southeast Gondi, General Southeast Gondi, Hill Maria-Koya, Hill Maria
mku, mdh, ktg, mud, mso, mlj, gok	Southeast Gondi, General Southeast Gondi, Muria
bhm, bhb, bhs	Southeast Gondi, General Southeast Gondi, Bison Horn Maria

Table 1: Classification of the 46 sites according to Glottolog (Nordhoff and Hammarström, 2011).

The whole dialect region is divided into two major groups: Northwest Gondi and Southeast Gondi which are divided into five major subgroups: Northern Gondi, Southern Gondi, Hill Maria, Bison Horn Maria, Muria where Northern Gondi and Southern Gondi belong to the Northwest Gondi branch whereas the rest of the subgroups belong to Southeast Gondi branch. It has

[1] http://glottolog.org/resource/languoid/id/gond1265

to be noted that there is no gold standard about the internal structure of dialects belonging to each dialect group.

3 Methods for comparing and visualizing dialectal differences

We use the IPA transcribed data to compute both unweighted and weighted string similarity/distance between two words. We use the same IPA data to train LSTM autoencoders introduced by Rama and Çöltekin (2016) and project the autoencoder based distances onto a map.

As mentioned earlier, the dialectometric analyses typically assume that all words that share the same meaning are cognates. However, as shown by Garapati (1991), some Gondi dialects exhibit a clear tree structure. Both dialectometric and autoencoder methods only provide an aggregate amount of similarity between dialects and do not work with cognates directly. The methods are sensitive to lexical differences only through high dissimilarity of phonetic strings. Since lexical and phonetic differences are likely to indicate different processes of linguistic change, we also analyze the categorical differences due to lexical borrowings/changes. For this purpose, we perform automatic cognate identification and then use the inferred cognates to perform both Bayesian phylogenetic analysis and dialectometric analysis.

3.1 Dialectometry

3.1.1 Computing aggregate distances

In this subsection, we describe how Levenshtein distance and autoencoder based methods are employed for computing site-site distances.

Levenshtein distance: Levenshtein distance is defined as the minimum number of edit operations (insertion, deletion, and substitution) that are required to transform one string to another. We use the Gabmap (Nerbonne et al., 2011) implementation of Levenshtein distance to compute site-site differences.

Autoencoders: Rama and Çöltekin (2016) introduced LSTM autoencoders for the purpose of dialect classification. Autoencoders were employed by Hinton and Salakhutdinov (2006) for reducing the dimensionality of images and documents. Autoencoders learn a dense representation that can be used for clustering the documents and images.

An autoencoder network consists of two parts: *encoder* and *decoder*. The encoder network takes a word as an input and transforms the word to a fixed dimension representation. The fixed dimension representation is then supplied as an input to a decoder network that attempts to reconstruct the input word. In our paper, both the encoder and decoder networks are Long-Short Term Memory networks (Hochreiter and Schmidhuber, 1997).

In this paper, each word is represented as a sequence $(x_1, \ldots x_T)$ of one-hot vectors of dimension $|P|$ where P is the total number (58) of IPA symbols across the dialects. The encoder is a LSTM network that transforms each word into $h \in \mathbb{R}^k$ where k is predetermined beforehand (in this paper, k is assigned a value of 32). The decoder consists of another LSTM network that takes h as input at each timestep to predict an output representation. Each output representation is then supplied to a softmax function to yield $\hat{x_t} \in \mathbb{R}^{|P|}$. The autoencoder network is trained using the binary cross-entropy function $(-\sum_t x_t log(\hat{x_t}) + (1 - x_t) log(1 - \hat{x_t}))$ where, x_t is a 1-hot vector and $\hat{x_t}$ is the output of the softmax function at timestep t to learn both the encoder and decoder LSTM's parameters. Following Rama and Çöltekin (2016), we use a bidirectional LSTM as the encoder network and a unidirectional LSTM as the decoder network. Our autoencoder model was implemented using Keras (Chollet, 2015) with Tensorflow (Abadi et al., 2016) as the backend.

3.1.2 Visualization

We use Gabmap, a web-based application for dialectometric analysis for visualizing the site-site distances (Nerbonne et al., 2011; Leinonen et al., 2016).[2] Gabmap provides a number of methods for analyzing and visualizing dialect data. Below, we present maps and graphics that are results of *multi-dimensional scaling* (MDS) *clustering*.

For all analyses, Gabmap aggregates the differences calculated over individual items (concepts) to a site-site distance matrix. With phonetic data, it uses site-site differences based on string edit distance with a higher penalty for vowel–consonant alignments and a lower penalty for the alignments of sound pairs that differ only in IPA diacritics. With binary data, Gabmap uses Hamming distances to compute the site-site differences. The cognate clusters obtained from the automatic iden-

[2] Available at `http://gabmap.nl/`.

tification procedure (section 2.2) forms categories (cognate clusters) which are analyzed using binary distances. Finally, we also visualize the distances from the autoencoders (section 2.1) using Gabmap.

Gabmap provides various agglomerative hierarchical clustering methods for clustering analyses. In all the results below, we use Ward's method for calculating cluster differences. For our analyses, we present the clustering results on (color) maps and dendrograms. Since the clustering is known to be relatively unstable, we also present probabilistic dendrograms that are produced by noisy clustering (Nerbonne et al., 2008). In noisy clustering, a single cluster analysis is performed a large number of times ($\sim$ 100) by adding a small noise to the distance matrix that is proportional to the standard deviation of the original distance matrix. The combined analysis then provides statistical support for the branches in a dendrogram.

The multi-dimensional scaling (MDS) is a useful analysis/visualization technique for verifying the clustering results and visualizing the dialect continuum. A site-site (linguistic) distance matrix represents each site on a multi-dimensional space. MDS 'projects' these distances to a smaller dimensional space that can be visualized easily. In dialect data, the distances in few most-important MDS dimensions correlate highly with the original distances, and these dimensions often correspond to linguistically meaningful dimensions. Below, we also present maps where areas around the linguistic similar locations are plotted using similar colors.

3.2 Phylogenetic approaches

3.2.1 Automatic cognate detection

Given a multilingual word list for a concept, the automatic cognate detection procedure (Hauer and Kondrak, 2011) can be broken into two parts:

1. Compute a pairwise similarity score for all word pairs in the concept.
2. Supply the pairwise similarity matrix to a clustering algorithm to output clusters that show high similarity with one another.

Needleman-Wunsch algorithm (NW, Needleman and Wunsch (1970); the similarity counterpart of Levenshtein distance) is a possible choice for computing the similarity between two words. The NW algorithm maximizes similarity whereas Levenshtein distance minimizes the distance between two words. The NW algorithm assigns a score of 1 for character match and a score -1 for character mismatch. Unlike Levenshtein distance, NW algorithm assigns a penalty score for opening a gap (deletion operation) and a penalty for gap extension which models the fact that deletion operations occur in chunks (Jäger, 2013).

The NW algorithm is not sensitive to different sound segment pairs, but a realistic algorithm should assign higher similarity score to sound correspondences such as /l/ $\sim$ /r/ than the sound correspondences /p/ $\sim$ /r/. The weighted Needleman-Wunsch algorithm takes a segment-segment similarity matrix as input and then aligns the two strings to maximize the similarity between the two words.

In dialectometry (Wieling et al., 2009), the segment-segment similarity matrix is estimated using *pointwise mutual information* (PMI). The PMI score for two sounds x and y is defined as followed:

$$pmi(x, y) = log\frac{p(x, y)}{p(x)p(y)} \quad (1)$$

where, $p(x, y)$ is the probability of x, y being matched in a pair of cognate words, whereas, $p(x)$ is the probability of x. A positive PMI value between x and y indicates that the probability of x being aligned with y in a pair of cognates is higher than what would be expected by chance. Conversely, a negative PMI value indicates that an alignment of x with y is more likely the result of chance than of shared inheritance.

The PMI based computation requires a prior list of plausible cognates for computing a weighted similarity matrix between sound segments. In the initial step, we extract cross-lingual word pairs that have a Levenshtein distance less than 0.5 and treat them as a list of plausible cognates. The PMI estimation procedure is described as followed:

1. Compute alignments between the word pairs using a vanilla Needleman-Wunsch algorithm.[3]
2. The computed alignments from step 1 are then used to compute similarity between segments x, y according to the following formula:

[3]We set the gap-opening penalty to -2.5 and gap extension penalty to -1.75.

3. The PMI matrix obtained from step 2 is used to realign the word pairs and generate a new list of segment alignments. The new list of alignments is employed to compute a new PMI matrix.
4. Steps 2 and 3 are repeated until the difference between PMI matrices reach zero.

In our experience, five iterations were sufficient to reach convergence. At this stage, we use the PMI matrix to compute a word similarity matrix between the words belonging to a single meaning. The word similarity matrix was converted into a word distance matrix using the following transformation: $(1 + exp(x))^{-1}$ where, x is the PMI score between two words. We use the InfoMap clustering algorithm (List et al., 2016) for the purpose of identifying cognate clusters.

3.2.2 Bayesian phylogenetic inference

The Bayesian phylogenetics originated in evolutionary biology and works by inferring the evolutionary relationship (trees) between DNA sequences of species. The same method is applied to binary traits of species (Yang, 2014). A binary trait is typically a presence or absence of a evolutionary character in an biological organism. Computational biologists employ a probabilistic substitution model θ that models the transition probabilities from $0 \rightarrow 1$ and $1 \rightarrow 0$. The substitution matrix would be a 2×2 matrix in the case of a binary data matrix.

A evolutionary tree that explains the relationship between languages consist of topology (τ) and branch lengths ($\mathbf{T}$). The likelihood of the binary data to a tree is computed using the pruning algorithm (Felsenstein, 1981). Ideally, identifying the best tree would involve exhaustive enumeration of the trees and calculating the likelihood of the binary matrix for each tree. However, the number of possible binary tree topologies grows factorially ($(2n-3)!!$ where, n is the number of languages) and, hence intractable even for a small number (20) of languages. The inference problem would be to estimate the joint posterior density of $\tau, \theta, \mathbf{T}$.

The Bayesian phylogenetic inference program (*MrBayes*;[4] Ronquist and Huelsenbeck (2003)) requires a binary matrix (languages $\times$ number of clusters) of 0s and 1s, where, each column shows if a language is present in a cluster or not. The cognate clusters are converted into a binary matrix of 0s and 1s in the following manner. A word for a meaning would belong to one or more cognate sets. For example, in the case of German, Swedish, and Hindi, the word for *dog* in German 'Hund' and Swedish 'hund' would belong to the same cognate set, while Hindi 'kutta' would belong to a different category. The binary trait matrix for these languages for a single meaning, *dog*, would be as in table 2. A Bayesian phylogenetic analysis employs a Markov-Chain Monte-Carlo procedure to navigate across the tree space. In this paper, we ran two independent runs until the trees inferred by the two runs do not differ beyond a threshold of 0.01. In summary, we ran both the chains for 4 million states and sampled trees at every 500 states to avoid auto-correlation. Then, we threw away the initial one million states as burn-in and generated a summary tree of the post burn-in runs (Felsenstein, 2004). The summary tree consists of only those branches which have occurred more than 50% of the time in the posterior sample, consisting of 25000 trees.

German	*Hund*	1	0
Swedish	*hund*	1	0
Hindi	*kutta*	0	1

Table 2: Binary matrix for meaning "dog".

4 Results

In this section, we present visualizations of differences in the language area using MDS and noisy clustering.

4.1 String edit distance

In the left map in Figure 2, the first three MDS dimensions are mapped to RGB color space, visualizing the differences between the locations. Note that the major dialectal differences outlined in table 1 are visible in this visualization. For example, the magenta and yellow-green regions separate the Bison Horn Maria and the Hill Maria groups from the surrounding areas with sharp contrasts. The original linguistic distances and the distances based on first three MDS dimensions correlate with $r = 0.90$, hence, retaining about 81% of the variation in the original distances. The middle map in figure 2 displays only the first dimension, which seems to represent a difference between north and south. On the other hand, the

[4] http://mrbayes.sourceforge.net/

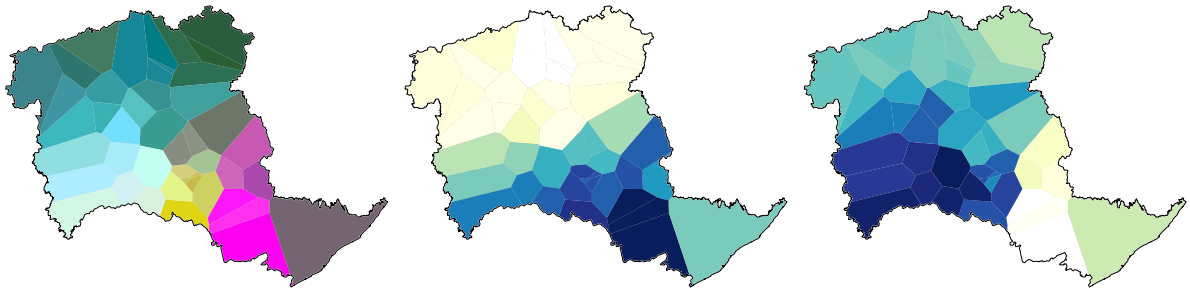

Figure 2: MDS analysis performed by Gabmap with string edit distance. The left map shows first three MDS dimensions mapped to RGB color space. The middle map shows only the first dimension, and the right map shows the second MDS dimension. The first three dimensions correlate with the original distances with $r = 0.73$, $r = 0.55$ and $r = 0.41$, respectively, and first three dimensions with $r = 0.90$.

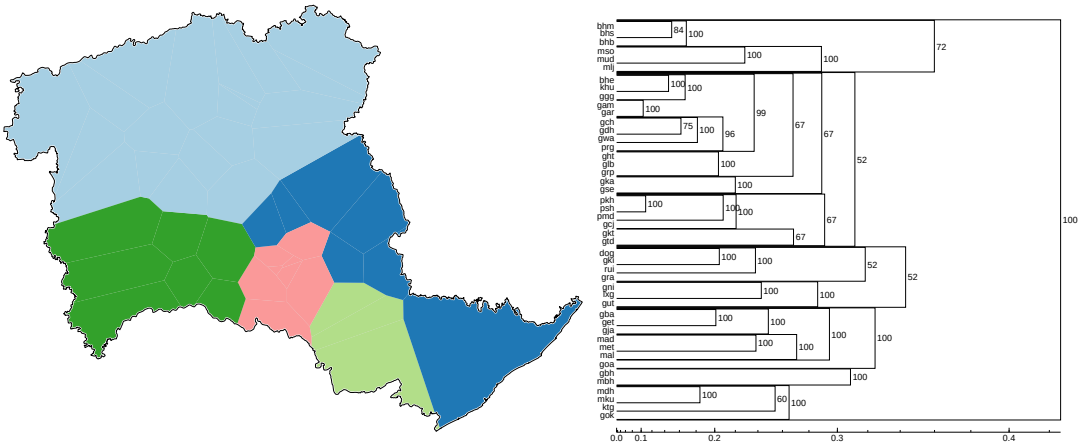

Figure 3: Clustering analysis performed by Gabmap with string edit distance using Ward's method and the color in the map indicate 5 dialect groups. Probabilistic dendrogram from the default Gabmap analysis (string edit distance).

right map (second MDS dimension) seems to indicate a difference between Bison Horn Maria (and to some extent Muria) and the rest.

The clustering results are also complementary to the MDS analysis. The 5-way cluster map presented in figure 3 indicates the expected dialect groups described in table 1. Despite some unexpected results in the detailed clustering, the probabilistic dendrogram presented in figure 3 also shows that the main dialect groups are stable across noisy clustering experiments. For instance, the Bison Horn Maria group (bhm, bhs, bhb) presented on the top part of the dendrogram indicates a very stable group: these locations are clustered together in all the noisy clustering experiments. Similarly, the next three locations (mco, mud, mlj, belonging to Muria area) also show a very strong internal consistency, and combine with the Bison Horn Maria group in 72% of the noisy clustering experiments. However, other members of Muria group (mdh, mku, ktg, gok at the bottom of the probabilistic dendrogram) seem to be placed often apart from the rest of the group.

4.2 Binary distances

Next, we present the MDS analysis based on lexical distances in figure 4. For this analysis, we identify cognates for each meaning (cf. section 2.2), and treat the cognate clusters found in each location as the only (categorical) features for analysis. The overall picture seems to be similar to the analysis based on the phonetic data, although the north-south differences are more visible in this analysis. Besides the first three dimensions (left map), both first (middle map) and second (right map) dimensions indicate differences between north and south. The left figure shows that there is a gradual transition from the Northern dialects (gtd, gkt, prg, ggg, khu, bhe, gcj, pmd, psh,

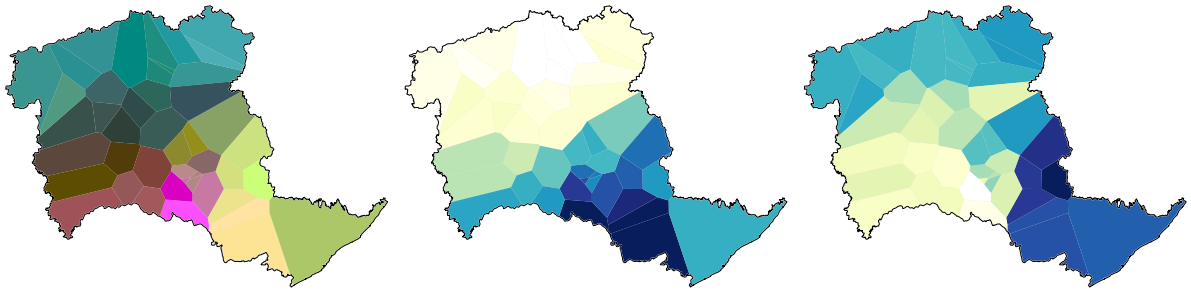

Figure 4: MDS analysis performed by Gabmap with categorical differences. The left map shows first three MDS dimensions mapped to RGB color space. The middle map shows only the first dimension, and the right map shows the second MDS dimension. The first three dimensions correlate with the original distances with $r = 0.77$, $r = 0.53$ and $r = 0.41$, respectively, and first three dimensions with $r = 0.94$.

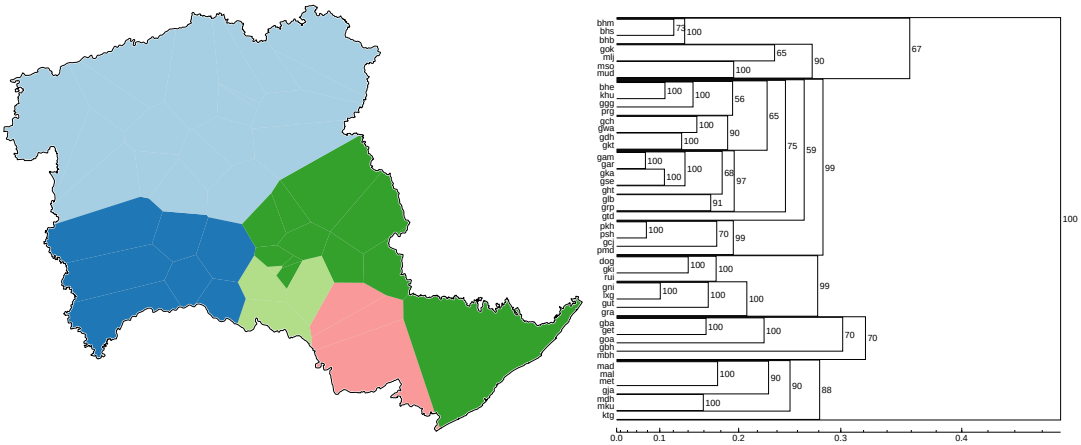

Figure 5: The dendrogram shows the results of the hierarchical clustering (left) based on binary matrix. Probabilistic dendrogram from the Gabmap analysis with Hamming distances.

pkh) to the rest of the northern dialects that share borders with Muria and Southern dialects. There is a transition between Southern dialects to the Hill Maria dialects where, the Hill Maria dialects do not show much variation.

The clustering analysis of the binary matrix from cognate detection step is projected on the geographical map of the region in figure 5. The map retrieves the five clear subgroups listed in table 1. Then, we perform a noisy clustering analysis of the Hamming distance matrix which is shown in the same figure. The dendrogram places Bison-Horn Maria dialects (bhm, bhs, bhb) along with the eastern dialects of Muria subgroup. It also places all the Northern Gondi dialects into a single cluster with high confidence. The dendrogram also places all the southern dialects into a single cluster. On the other hand, the dendrogram incorrectly places the Hill Maria dialects along with the western dialects of Muria subgroup. With slight variation in the detail, the cluster analysis and the probabilistic dendrogram presented in figure 5 are similar to the analysis based on phonetic differences.

4.3 Autoencoder distances

The MDS analysis of autoencoder-based distances are shown in figure 6. The RGB color map of the first three dimensions shows the five dialect regions. The figure shows a clear boundary between Northern and Southern Gondi dialects. The map shows the Bison Horn Maria region to be of distinct blue color that does not show much variance. The autoencoder MDS dimensions correlate the highest with the autoencoder distance matrix. The first dimension (middle map in figure 6) clearly distinguishes the Northern dialects from the rest. The second dimension distinguishes Southern Gondi dialects and Muria dialects from the rest of the dialects.

The clustering analysis of the autoencoder dis-

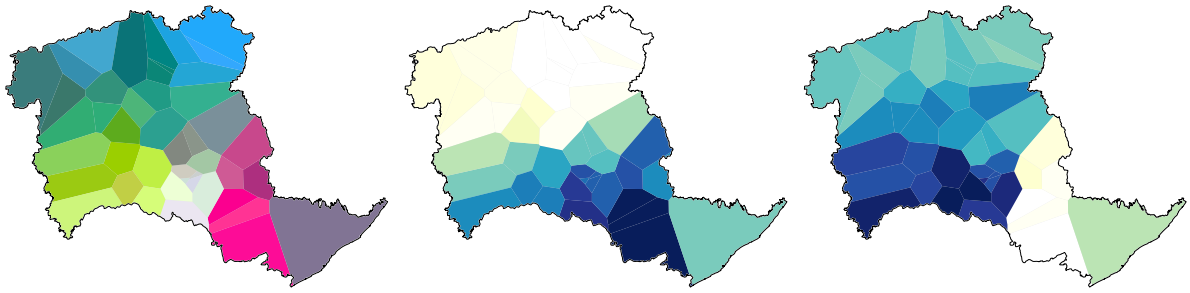

Figure 6: MDS analysis performed by Gabmap with autoencoder differences. The left map shows first three MDS dimensions mapped to RGB color space. The middle map shows only the first dimension, and the right map shows the second MDS dimension. The first three dimensions correlate with the original distances with $r = 0.74$, $r = 0.57$ and $r = 0.49$, respectively, and first three dimensions with $r = 0.92$.

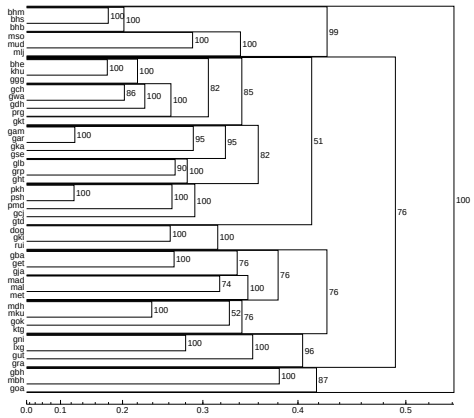

Figure 7: Probabilistic dendrogram from the Gabmap analysis with autoencoder distances. The clustering result is similar to the left map in figure 3.

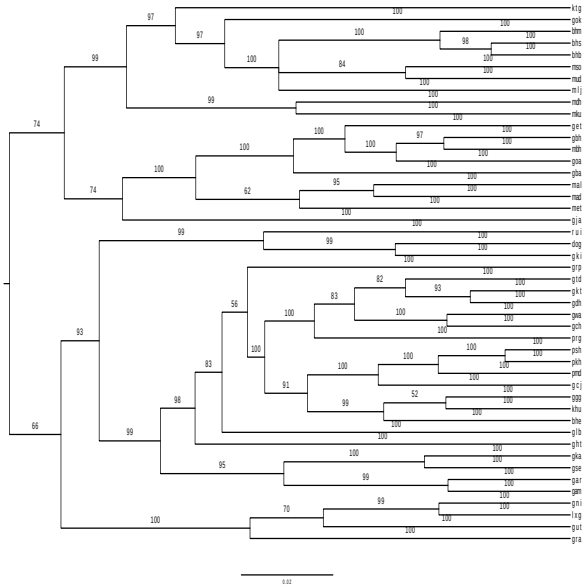

Figure 8: The majority consensus tree of the Bayesian posterior trees.

tances are projected on to the geographical map in figure 7. The map retrieves the five subgroups in table 1. The noisy clustering clearly puts the Bison Horn Maria group into a single cluster. It also places all the northern dialects into a single group with 100% confidence. On the other hand, the dendrogram splits the Southern Gondi dialects into eastern and western parts. The eastern parts are placed along with the Hill Maria dialects. The clustering analysis also splits the Muria dialects into three parts. However, the dendrogram places *gok* (a eastern Muria dialect) incorrectly with Far Western Muria (mku).

4.4 Bayesian analysis

The summary tree of the Bayesian analysis is shown in figure 8. The figure also shows the percentage of times each branch exists in the posterior sample of trees. The tree clearly divided Northwest Gondi from Southeast Gondi groups. The tree places all the Northern Gondi dialects into a single group in 99% of the trees. The southern dialects are split into two different branches with *rui, dog, gki* branching later from the common Northwest Gondi later than the rest of the Southern Gondi dialects. The tree clearly splits the Hill Maria dialects from rest of Southeast Gondi dialects. The tree also places all the Bison Horn Maria dialects into a single group but does not put them into a different group from the rest of the Muria dialects.

5 Conclusion

In this paper, we performed analysis using tools from dialectometry and computational historical

linguistics for the analysis of Gondi dialects. The dialectometric analysis rightly retrieves all the subgroups in the region. However, both edit distance and autoencoder distances differ in the noisy clustering analysis. On the other hand, the noisy clustering analysis on the binary cognate matrix yields the best results. The Bayesian tree based on cognate analysis also retrieves the top level subgroups right but does not clearly distinguish Bison Horn Maria group from Muria dialects. As a matter of fact, the Bayesian tree agrees the highest with the gold standard classification from Glottolog.

The contributions of the paper is as followed. We digitized a multilingual lexical wordlist for 46 dialects and applied both dialectometric and phylogenetic methods for the classification of dialects and find that phylogenetic methods perform the best when compared to the gold standard classification.

Acknowledgments

The authors thank the five reviewers for the comments which helped improve the paper. The first and third authors are supported by the ERC Advanced Grant 324246 EVOLAEMP, which is gratefully acknowledged.

The code and the data for the experiments is available at `https://github.com/PhyloStar/Gondi-Dialect-Analysis`

References

Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

David K. Beine. 1994. A sociolinguistic survey of the Gondi-speaking communities of central india. Master's thesis, San Diego State University, San Diego.

Thomas Burrow and S. Bhattacharya. 1960. A comparative vocabulary of the Gondi dialects. *Journal of the Asiatic Society*, 2:73–251.

François Chollet. 2015. Keras. *GitHub repository: https://github. com/fchollet/keras*.

Joseph Felsenstein. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Joseph Felsenstein. 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Umamaheshwar Rao Garapati. 1991. Subgrouping of the Gondi dialects. In B. Lakshmi Bai and B. Ramakrishna Reddy, editors, *Studies in Dravidian and general linguistics: a festschrift for Bh. Krishnamurti*, pages 73–90. Centre of Advanced Study in Linguistics, Osmania University.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 865–873, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Gerhard Jäger. 2013. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291.

Grzegorz Kondrak. 2009. Identification of cognates and recurrent sound correspondences in word lists. *Traitement Automatique des Langues et Langues Anciennes*, 50(2):201–235, October.

Therese Leinonen, Çağrı Çöltekin, and John Nerbonne. 2016. Using Gabmap. *Lingua*, 178:71–83.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Berlin, Germany, August. Association for Computational Linguistics.

Johann-Mattis List. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France, April. Association for Computational Linguistics.

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(3):443–453.

John Nerbonne, Peter Kleiweg, Wilbert Heeringa, and Franz Manni. 2008. Projecting dialect differences to geography: Bootstrap clustering vs. noisy

clustering. In Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt, and Reinhold Decker, editors, *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society*, pages 647–654, Berlin. Springer.

John Nerbonne, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap – a web application for dialectology. *Dialectologia*, Special Issue II:65–89.

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

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the First International Workshop on Linked Science*, volume 783, pages 53–58.

Taraka Rama and Çağri 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, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Taraka Rama. 2015. Automatic cognate identification with gap-weighted string subsequences. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*, pages 1227–1231.

Fredrik Ronquist and John P Huelsenbeck. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574.

Ian Smith. 1991. Interpreting conflicting isoglosses: Historical relationships among the Gondi dialects. In B. Lakshmi Bai and B. Ramakrishna Reddy, editors, *Studies in Dravidian and general linguistics: a festschrift for Bh. Krishnamurti*, pages 27–48. Centre of Advanced Study in Linguistics, Osmania University.

Martijn Wieling, Jelena Prokić, and John Nerbonne. 2009. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, pages 26–34. Association for Computational Linguistics.

Ziheng Yang. 2014. *Molecular evolution: A statistical approach*. Oxford University Press, Oxford.