Jessica K. Ivani*, Netra Paudyal and John Peterson

Indo-Aryan – a house divided? Evidence for the east–west Indo-Aryan divide and its significance for the study of northern South Asia

https://doi.org/10.1515/jsall-2021-2029
Published online August 30, 2021

Abstract: In this study, we investigate the possible presence of an east–west divide in Indo-Aryan languages suggested in previous literature (Peterson, John. 2017a. Fitting the pieces together – towards a linguistic prehistory of eastern-central South Asia (and beyond). *Journal of South Asian Languages and Linguistics* 4(2). 211–257.), with the further hypothesis that this divide may be linked to the influence of the Munda languages, spoken in the eastern part of the subcontinent. Working with 217 fine-grained variables on a sample of 27 Indo-Aryan and Munda languages, we test the presence of a geographical divide within Indo-Aryan using computational methods such as cluster analysis in combination with visual statistical inference. Our results confirm the presence of a geographical divide for the whole dataset and most of the individual features. We then proceed to compute the degree of similarity between the Indo-Aryan languages and Munda, using a Bayesian alternative to a t-test. The results for most features support the claim that the languages identified in the eastern clusters are indeed more similar to Munda, thereby opening up further research scenarios for the history of this region.

Keywords: historical linguistics; Indo-Aryan; language contact; Munda

1 Introduction[1]

There are at least 600 languages belonging to no fewer than six families spoken today in South Asia, a region which includes Pakistan, India, Nepal, Bangladesh,

1 This research was funded by the Deutsche Forschungsgemeinschaft (DFG) Project Grant 326697274 ("Towards a linguistic prehistory of eastern-central South Asia (and beyond)"), which

***Corresponding author: Jessica K. Ivani,** University of Zürich, Zürich, Switzerland,
E-mail: jessica.ivani@uzh.ch
Netra Paudyal and John Peterson, Kiel University, Kiel, Germany,
E-mail: n.paudyal@isfas.uni-kiel.de (N. Paudyal), jpeterson@isfas.uni-kiel.de (J. Peterson)

Sri Lanka and the Maldives. These languages belong to the following families: Indo-European (including Indo-Aryan, Iranian and Nuristani), Dravidian, Andamanese,[3] Tibeto-Burman, Tai-Kadai, and Austro-Asiatic (including Munda). In addition, at least three isolates are also spoken in the subcontinent: Burushaski, in Pakistan, Kusunda in central Nepal, and Nihali in central India. Figure 1 provides an overview of these, excluding Nihali, Kusunda and Tai-Kadai languages. In

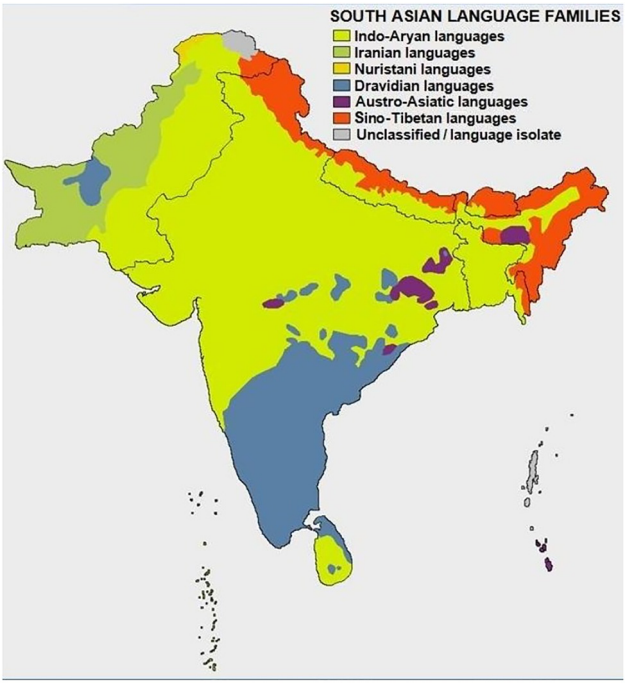


Figure 1: South Asian language families.[2]

we gratefully acknowledge here. John Peterson is the project principal investigator; he wrote the introduction and the conclusions. Jessica K. Ivani oversaw the data collection, performed the analysis and wrote the sections on data, methodology and results. Netra Paudyal contributed with field data. Data collection was performed by Jessica K. Ivani, assisted by a team of student assistants: Annika Besser, Nellia Bleyer, Lennart Chevallier, Nikita König, Johanna Schwarz, and with the support of Erika Just, PhD student at the University of Kiel.

2 Source: https://commons.wikimedia.org/wiki/File:Indo-Aryan_language_map.svg, by user C1MM. Creative Commons license: https://creativecommons.org/licenses/by-sa/3.0/.

3 Abbi (2009) argues that there are in fact two genealogically unrelated language families in the Andaman Islands, whose protolanguages she refers to as "Proto Ang" and "Proto Great Andamanese".

the present study we focus on the Indo-Aryan and Munda languages, whose internal relationships are illustrated in Figure 2 and Figure 4, respectively.

There is still much debate on the internal groupings of Indo-Aryan, especially with respect to the status of the so-called "outer languages". Figure 2 presents a simplified and slightly adapted version of the genealogical scheme given in Eberhard and Gary (2019), with a few of the better-known representatives of each branch, assuming an outer branch here for ease of presentation. The authors of the present article are entirely neutral on this issue.[4]

With respect to the Munda languages, there is general consensus that there is a North Munda branch which is clearly distinguishable from the remaining languages, although the internal groupings of this second group are still a matter of intense debate. For ease of reference, this latter group is referred to as South Munda, although it is not a clearly defined group like the northern branch. Figure 4, from Zide (1969: 412), is by far the most influential model for this family, despite all refinements to this tree which have been proposed in the last decades.[5]

South Asia has long been considered a textbook-example of a "Sprachbund" or language area. Bloch (1934: 322–328) was perhaps the first author to note structural similarities among South Asian languages belonging to different stocks, and some 20 years later, Emeneau (1956) brought the high level of convergence between the languages of South Asia – regardless of language family – to the attention of a larger linguistic audience. In the following years numerous further traits were suggested in individual studies such as Gumperz and Wilson (1971), before Masica (1976) appeared, a book-length study examining South Asia as a

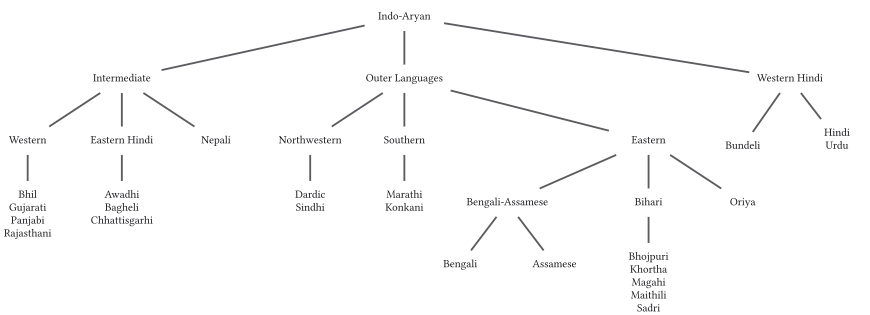


Figure 2: The Indo-Aryan languages, simplified and adapted from the classification in Eberhard and Gary (2019).

4 For further classification systems of Indo-Aryan, see Masica (2001: 446–463).

5 Cf. Peterson (2015b) for further discussion of classificatory systems of Munda.

linguistic area, summarizing earlier research on the topic and examining others such as word order, causative verb morphology, general converbs, etc., in greater detail. One of Masica's (1976) suggestions for future work was discovering possible sub-regions in South Asia, and this has become the main emphasis of research in contact linguistics in the region since then. Among the many studies in this area – far too many to mention here – are a number dealing with eastern-central South Asia, the focus of the present study. These include Abbi (1997), Ebert (1993, 1999), Osada (1991) and Peterson (2010b, 2015a, 2017b).

Recent research in areal linguistics however is increasingly doing away with the notion of "linguistic areas" in any meaningful sense of the term, with the focus instead being placed on the details of areal diffusion itself (Enfield 2005: 191). This is the view taken in Peterson (2017a) and is also assumed in the present study.

According to Peterson (2017a), there are two main geographical zones in northern South Asia with respect to the distribution of linguistic density, namely the Indo-Gangetic Plain and the mountains and hills which border this region on all sides. The first of these, the Indo-Gangetic Plain (see Figure 3), is a classic spread zone in the terminology of Nichols (1992), i.e., an area of rapid language spread from a historical perspective, with little genealogical diversity, etc. It is bounded in the north and northwest by the Himalayan range and Hindu Kush, respectively, in the south by the Vindhya and Satpura ranges of central India, and in the southeast by the Chotanagpur Plateau. These mountainous and hill areas, by

Figure 3: The Indo-Gangetic Plain and neighboring geological regions.[6]

6 From Peterson (2017a: 218), reprinted with kind permission by Mouton de Gruyter.

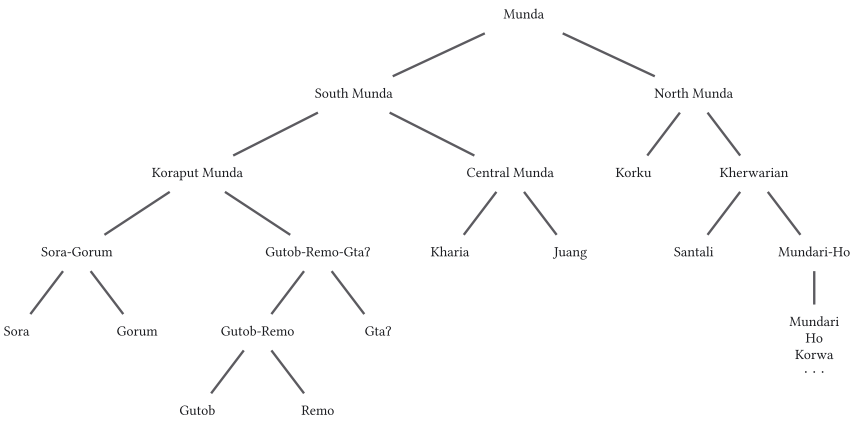


Figure 4: The Munda languages according to Zide (1969).

contrast, are residual or accretion zones (Nichols 1992, 1997). In these last three regions numerous languages belonging to the Dravidian and Munda families are found, as well as the isolate Nihali, spoken in Maharashtra, and it is almost certain that there were once also languages belonging to other language families which have since disappeared (cf. e.g. the discussion of Kurmali in Paudyal and Peterson, this issue). Figure 5, from Anderson (2008: 2), shows the relative position of the Munda languages throughout this region.

Based on the modern distribution of the Munda languages in these accretion zones, Peterson (2017a) suggests that earlier forms of these Austro-Asiatic languages were once much more widespread than they are today, and that speakers of these languages at some time either switched to Indo-Aryan languages, or survived with their traditional languages only in these more remote mountainous regions which were of little interest to pastoralists and farmers. If so, we would reasonably expect to see signs of this substrate in the languages of the eastern half of the Gangetic Plain.

To this end, Peterson (2017a) analyzed data on 28 structural features from 29 languages belonging to the Munda, Indo-Aryan and Dravidian families, using programs borrowed from genetics such as UPGMA and NeighborNet to analyze the data, checking for clusters which would either support or refute a likely substrate in eastern Indo-Aryan. His results show a clear divide between eastern and western Indo-Aryan, with the eastern Indo-Aryan languages grouping together with Munda, suggesting that there is in fact a Munda substrate in eastern Indo-Aryan. However, his study is admittedly based on a very small number of features, hence the present study is designed to test this hypothesis once again with a much larger data base, to which we now turn.

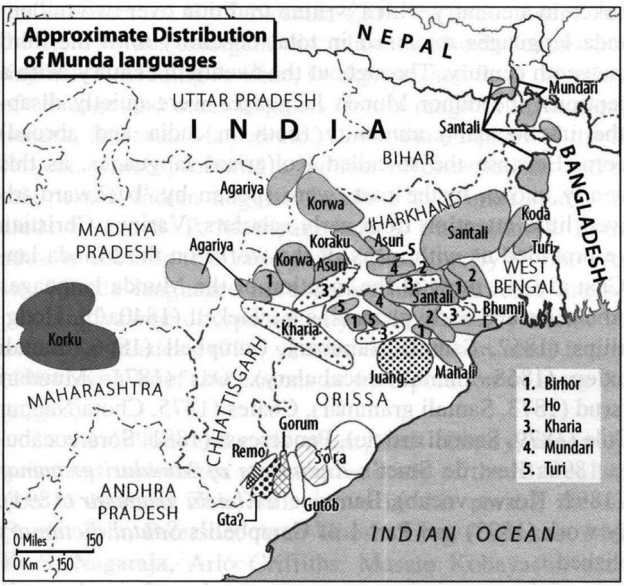


Figure 5: Approximate distribution of the Munda languages.[7]

2 Language sample

We surveyed a sample of 27 languages in total, covering two main genealogical groups: Indo-Aryan and Munda (Austro-Asiatic). The language sample is built around two main criteria: geographical distribution of the languages and the availability of descriptive sources. The targeted geographic area is northern and central India, and we have prioritized the Indo-Aryan and Munda languages spoken in the region. However, the full sample (not used here) includes several Dravidian languages from northern central India, as well as isolates like Nihali, Burushaski and Kusunda, collected for additional analysis and future research.

The sample is not genealogically balanced; Indo-Aryan languages are more represented in the sample, compared to the languages of the Munda group; this distribution roughly reflects the presence of these language families in the Indian subcontinent and the overall number of languages attested for each language family.

Given that one of our main interests is the fine-grained analysis of specific grammatical topics, we required detailed and reliable descriptive sources, more

7 From Anderson (2008), reprinted with kind permission of Mouton de Gruyter.

extensive than grammar sketches and minimal linguistic outlines. This imposed certain limits on the sample building process as most of the languages of South Asia are under-described and a large body of the available data, especially on Indo-Aryan and Dravidian languages, is dated and unreliable.

Following these requirements, we have approached the language selection systematically. To ensure that our sample contains enough linguistic variation, we aimed at including at least one language for each sub-branch of all the families. In case a grammar description was not available for the chosen language, or the description did not meet the minimal required quality, we moved to another language belonging to the same sub-branch, until the coverage for the branch was completed, moving then to the next sub-branch. Languages spoken in North Central India were preferred. The geographical distribution of the languages is shown in Figure 6.

Figure 6: The language sample.

3 Structural features

We performed data collection based on a custom designed parameter set of 217 morphosyntactic features from 27 languages (16 Indo-Aryan languages, 11 Munda

languages), for a total of about 5,859 datapoints. The full set of parameters is organized into three main questionnaires: Questionnaire A (broad typological features), Questionnaire B (fine-grained features) and Questionnaire C (features specific to South Asia). Each questionnaire is discussed in detail in the respective subsection below. The choice behind these three different questionnaires reflects practical and methodological requirements. The broad typological features in Questionnaire A are generic enough to provide a bird's eye view of the area and highlight relevant typological trends; on the other hand, the detailed features in Questionnaire B allow us to capture the microvariation within a specific region and among related languages. In addition, the specific South Asian features (Questionnaire C) aim at expanding the findings in recent literature on this topic (e.g., Peterson 2017a). This multifaceted approach seeks to be explorative — without being too vague — and specific at the same time, by reducing the risks of "cherry-picking" the parameters.

Most coded variables were binary, such as the presence or absence of a particular feature. All such variables were annotated as either symmetric or asymmetric for the purpose of data analysis (cf. Section 4). A few of the variables were categorial, representing a choice between two or more types. In addition, most variables in Questionnaire B are logically dependent on each other, capturing features at different levels of detail.

The parameters were collected by a team that consists of BA and MA linguistics students who received training in data collection, with a focus on the prominent grammatical features in South Asia (as outlined e.g. in Emeneau 1956; Masica 1976). The data was collected in distinct phases. For Questionnaire A, two students worked independently of one another on an assigned language. The data was then reprocessed and checked by the team coordinator, and eventual ambiguities were solved by group discussion or by consulting language experts, where available. An analogous procedure was undertaken for Questionnaire C. Questionnaire B was developed with the specific purpose of collecting fine-grained linguistic information; the relevant linguistic data was first collected in a detailed language report. The individual language reports contain constructions, examples and other pertinent information to the most fine-grained level, taken from the grammar or some other linguistic source (such as field notes). Each language report uniquely identifies the language it describes with metadata such as name(s), geographic coordinates, ISO code[8] and Glottocode (Hammarström et al. 2019). The reports are written in Markdown,[9] a simple, easy-to-use text format that can be converted to other document formats and further processed. After a data check, revision and

8 https://www.loc.gov/standards/iso639-2/.

9 Originally released as https://daringfireball.net/projects/markdown/.

stabilization, the data from each language report were gathered in a spreadsheet that mapped the values to a tabular format compatible with other questionnaires.

The data contained in the language reports are aggregated in datasheets that describe the values in a binary state, to make them compatible with the other parameter sets and to set up the analysis phase. The data coverage, for each of the three questionnaires, is shown in Figure 7. The figure represents the available data for each language in the sample according to each Questionnaire (A, B, C). One cell represents one variable per language. Red cells indicate available data values, while blue cells show missing data (not mentioned in the source or the grammar description). Based on this representation, the median data coverage per language in our sample is 84%.

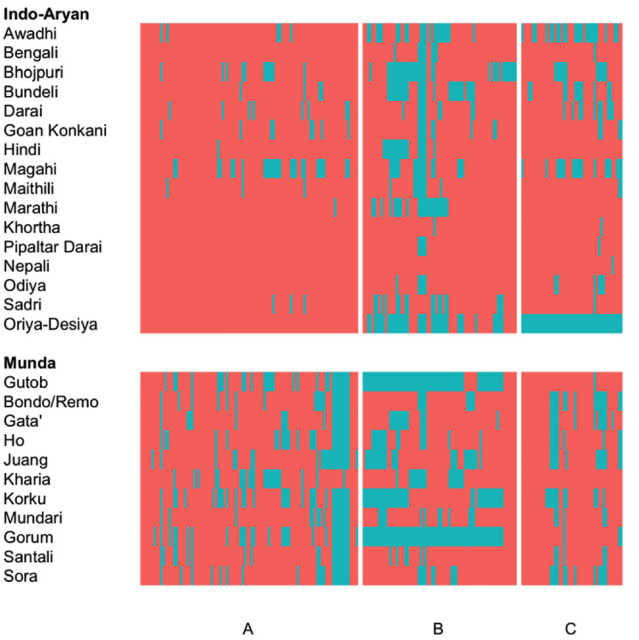


Figure 7: Data point coverage across languages.

3.1 Questionnaire A: broad typological features

Questionnaire A includes 101 morphosyntactic binary features (2,727 datapoints). In case the language source did not provide unambiguous information, the value was set as missing. The parameters belonging to this questionnaire cover a wide range of grammatical features. It is primarily inspired by the questionnaire used by

the Grambank project (Grambank 2019), which the South Asia project in Kiel has an ongoing collaboration with. The Grambank questionnaire offers a number of benefits: the questions are structured and organized by grammatical topics, it covers several different types of phenomena, and it has been tested and used on numerous languages. However, the questionnaire developed for the present analysis differs somewhat from that employed by the Grambank project. Questionnaire A does not contain information on features which are unattested in South Asia. Such features include, for example, the presence and expression of middle verbs or specific grammatical categories such as trial number.

The parameters in Questionnaire A are organized by domain. These include grammatical categories such as gender, number, case and definiteness within the nominal and verbal domains, general features such as comparative constructions, reciprocal structures and honorific forms. It also includes features that aim to describe specific lexical categories, such as pronouns. Syntax and grammatical relations are also covered, in addition to processes such as inflection and derivation (both nominal and verbal). The thematic subsections within the questionnaire allow us to explore groups of features about a specific category or domain in a straightforward way and to compare them with other domains. Also, having the questionnaire structured by modules enables us to identify at a glance the domains that show more or less variation across languages. The full dataset, which contains the list of languages, the parameters as well as the values, is openly accessible on the public access Zenodo repository.[10]

3.2 Questionnaire B: fine-grained features

Questionnaire B is built to explore in fine-grained detail the linguistic diversity of South Asia, without cherry-picking features. We developed this questionnaire by selecting and balancing typologically broad features, such as nominal number, possession, ergativity, and grammatical phenomena which are widespread in South Asia, such as non-nominative subjects, verb compounds and echo constructions. Questionnaire B is more detailed, with respect to individual features, compared to the other two questionnaires: each feature is described through multiple parameters that allow exploration at various levels of granularity. We describe not only the presence or absence of specific features, but also their

10 https://doi.org/10.5281/zenodo.3813195.

properties such as construction forms, syntactic constraints, language-specific syntactic distribution and semantic properties.

Questionnaire B includes 70 features for a total of 1,890 datapoints. The features are grouped under eight fine-grained domains: gender, number, classifiers, echo constructions, non-nominative subjects, possession, verb compounds and ergativity. The goal is to describe the presence, the formal structure, the distributional context and other relevant properties of each feature of interest. The full dataset, which contains the list of languages, the parameters as well as the values, is also openly accessible on Zenodo.[11]

3.3 Questionnaire C: specific South Asia features

The third set of features is an extension, in terms of the number of languages investigated, of the structural properties collected and explored in Peterson (2017a). That dataset focusses on the exploration of 28 features that are specific to South Asia. We adjusted the sample by further decomposing the features into 46 fine-grained variables, and by converting these into a binary format for both compatibility and comparability with the other two questionnaires. We "cleaned" the sample by removing the Dravidian languages and we added further Indo-Aryan and Munda languages to the original sample. The final sample contains 1,242 datapoints. The full dataset, which contains the list of languages, the parameters as well as the values, is openly accessible on Zenodo.[12]

3.4 Features subgroups

Features subgroups are those domains which are more extensively covered and are described through a richer and more fine-grained sets of variables. These features, which are contained and extrapolated from and across the questionnaires, when required, can be individually explored to identify deeper distributional trends across languages. In our questionnaires, the features that have received a more extensive description include ergativity, classifiers, noun and verb inflection, and number. A full list is available in Table 1.

11 https://doi.org/10.5281/zenodo.3813195.

12 https://doi.org/10.5281/zenodo.3813195.

Table 1: Overview of the results for each feature group.

Feature group	East–west divide	δ (similarity to Munda)		
		Left 2.5%	Median	Right 2.5%
The entire dataset	Yes	0.027	0.091	0.155
Questionnaire A	Yes	−0.304	0.035	0.570
Questionnaire B	Yes	0.034	0.124	0.212
Questionnaire C	Yes	0.121	0.217	0.316
Classifiers	Yes	0.127	0.353	0.578
Comparative constructions	Yes	0.312	0.463	0.630
Ergativity	Yes	0.996	1.001	1.005
Interclausal syntax	Yes	0.524	0.541	0.549
Noun inflection	Yes	0.030	0.245	0.450
Noun classes	Yes	−0.246	−0.023	0.246
Noun phrase	Yes	−0.080	0.096	0.252
Number	Yes	−0.100	−0.005	0.089
Case	No			
Echo constructions	No			
Negation	No			
Possession	No			
Pronouns	No			
Relative clause	No			
Syntax	No			
Verb inflection	No			

4 Methodology

The main goal of this paper is to test and explore in detail the linguistic divide claimed in Peterson (2017a) that suggests the presence of a clear split between eastern and western Indo-Aryan languages. To verify this claim, we utilize cluster analysis to obtain two clusters on the Indo-Aryan family from the data we have collected and visually inspect them. If the clusters follow the east–west divide, we interpret this as evidence for the claim. Furthermore, since our data is organized thematically, we can perform this analysis for different domains of grammar and examine the resulting differences. This allows us to study which morphosyntactic features (if any) follow the claim and which do not. Furthermore, by comparing results for different domains, we can identify the Indo-Aryan languages that constitute the core of the east and west sides of the divide (if any).

This first step enables us to test our second hypothesis, namely that the eastern Indo-Aryan languages are structurally closer to the Munda group, than the western Indo-Aryan varieties. This assumption was also made in Peterson (2017a), whose

overview was based on a small preliminary set of only 28 features. To test this, we compute the average similarity scores between each of the Indo-Aryan languages and the average score for the Munda languages in our sample. Our prediction is that the Indo-Aryan languages in the east will have a higher degree of similarity to Munda, on average, than the languages in the west. This prediction can be combined with the cluster analysis from the first step and validated statistically.

To verify our first prediction of the presence of a linguistic divide within the Indo-Aryan languages, we use cluster analysis as our data mining procedure. Cluster analysis can be defined as the task of partitioning data into meaningful groups, such that the members of the same group are maximally similar to each other and maximally different from the members of other groups. It is widely used in data exploration and has proven to be a useful tool for discovering and describing structures in complex datasets (For an extensive survey on generic data mining techniques see Berkhin 2006.)

To compute the dissimilarity matrix required for the cluster analysis, we use the Gower distance metric (Gower 1971) as implemented by the function *daisy* provided in the *cluster* R package.[13] Gower distance is a flexible hybrid dissimilarity measure that is well suited for complex datasets. It allows us to mix variables of different types (such as symmetric binary, asymmetric binary and categorical[14]) and it is also able to deal with missing data by ignoring the variable when comparing a pair of data-points with at least one of the values missing.

Before proceeding with data computation, we verified and cleaned the data to avoid miscomputations. Since most of our variables are asymmetric binary, contexts can arise where two languages have all variables set as false and thus are regarded as incomparable under the Gower dissimilarity metric. This can produce distortions when comparing languages in respect to a small group of selected variables. We detected such cases and treated the affected languages as being maximally similar instead (that is, for the case when all asymmetric binary variables are false and there are no other variables to make a comparison, we treated them as symmetric binary instead).

For a specific clustering method for our purpose and data, we chose robust k-means clustering around medoids, as implemented by the function *pam()* in the R package *cluster* (Kaufman and Rousseeuw 2009). This algorithm is well suited to our research question, as we are interested in finding two groups that are maximally different from each other. The k-means algorithm requires us to select a

13 https://CRAN.R-project.org/package=cluster.

14 A binary variable is symmetric if both of its states are equally valuable and carry the same weight. A binary variable is asymmetric if the outcome of the results is not equally important. A categorical variable is one with a limited number of distinct value or categories.

desired number of clusters to be computed. While we are primarily interested in two clusters only, we need to take precaution against the possibility that our clusters might not be well-behaved or stable. To guard against this, we repeat the analysis for a greater number of clusters (from two to five) and compare the average silhouette width[15] for each result as a measure of cluster validity (Rousseeuw 1987). In each case we discuss, two clusters are either the best fit overall or very close to it. In addition, we studied silhouettes for each specific cluster in order to identify languages with high and low degree of fit within the cluster.

As a next step, we visually inspected the obtained clusters for each feature set and evaluated whether their geographical distribution reflects the presence of an east–west divide within the Indo-Aryan language group. In order to support our visual inspection and to avoid the potential risks of over-interpreting the patterns, we set up a specific visual inference process, followed by a "Line-Up Protocol" procedure (Kerman et al. 2008).

Visual statistical inference is a fairly recent visual approach developed in statistics and information studies (Buja et al. 2009; Wickham et al. 2010). The main idea behind visual inference is to bring rigourous statistical testing to data visualization by treating visual data representations as "test statistics" and comparing them with a set of randomly generated plots that support the null hypothesis, that usually posits no relevant structure in data. If the true data plot does not look significantly different from the others, the null hypothesis would be supported. Conversely, if the true data plot clearly stands out from the rest, one could consider this result as a rejection of the null hypothesis of no structure in data. This procedure is known as "Line-up", and its protocol is straightforward: we generate and plot a set of decoys (null datasets that are random permutations of the real data), and we randomly position the true data plot among the decoys. We then show the plots to an impartial observer, asking whether s/he can spot the true data plot, which, in our case, corresponds to the plot that shows an east–west divide. At the same time, features that do not show any significant clustering would be undistinguishable from the randomly generated null plots, confirming the null hypothesis for the given feature set. We performed this test in double-blind fashion to ensure the reliability of the experiment (which means that neither the person showing the plots nor the person viewing them should be aware of which one

15 The silhouette width is a measure of how well an object (in our case, a given language and its features) fits into its assigned cluster. This value ranges between –1 and +1. The higher the value, the better the object fits into its assigned cluster. Objects with negative silhouette width are considered as outliers in their assigned cluster. The presence of many datapoints with a low silhouette width might indicate that the clustering configuration has too many or too few clusters, and hence the clustering is a poor fit for its data. See Rousseeuw (1987) for a detailed account.

corresponds to the true one) for the entire dataset, each of the questionnaires, and selected features sets.

Figure 8 illustrates the graphic representation for the "Line-up Protocol" on the whole dataset: the plot representing the true values is positioned among a set of randomly generated decoy plots. We asked a group of impartial observers to identify, among the plots, which one(s) (given a choice of maximum 3 plots, ranked in the order of best acceptance) are more representative of an east–west divide. In case the answer corresponded to the true plot, we accepted it as confirmation of the divide. Without knowing which plot represents real data and which one was a randomly permuted decoy, impartial observers tend to select J and Q as plots that show an east–west divide. Incidentally, the true data plot in Figure 8 is J.

For the second part of our hypothesis, we wanted to test the prediction that the eastern Indo-Aryan languages show a higher degree of similarity to the overall value for the languages of the Munda group, as compared to the western Indo-Aryan varieties. To do this we computed the average dissimilarity score for each of the Indo-Aryan languages in the two clusters to the languages in the Munda group. We then studied the distribution of these scores. We performed a Bayesian t-test (Moser and Stevens 1992; Welch 1937), as implemented in the R package *BEST*

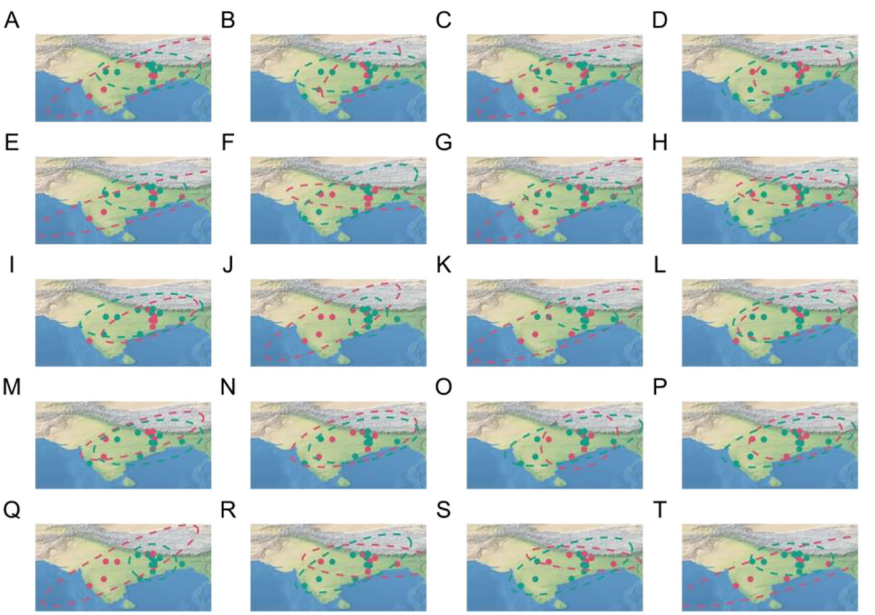


Figure 8: Visual inference test for the whole feature set.

(Kruschke 2013).[16] The Bayesian t-test offers several advantages compared to a basic t-test: it provides richer information, such as the distribution of a credibility interval[17] and group means, and it allows us to estimate at a glance the mean difference between groups rather than a basic *p*-value. For each group analysis, we computed and reported the posterior distribution of the differences in mean dissimilarity to Munda between the eastern and western Indo-Aryan cluster.

As an example, Figure 9 shows the result of the Bayesian t-test for the entire dataset (discussed in more detail in Section 5). The histogram graph displays the relative probability that the difference in mean dissimilarity to Munda between the two clusters has a specific value, given the evidence. The possible values for mean differences in dissimilarity to Munda are displayed on the *x* axis, the relative probability that a concrete value is supported by the data is displayed on the *y* axis. Positive values on the *x* axis thus represent the case when the eastern cluster languages are more similar to Munda on average, while negative values on the *x* axis represent the case where the western cluster languages are more similar to Munda on average. By comparing the area (cumulative relative probability) under the graph for intervals on x we can evaluate which case is more likely. For instance, looking at Figure 9 we can infer that the relative probability of difference in mean

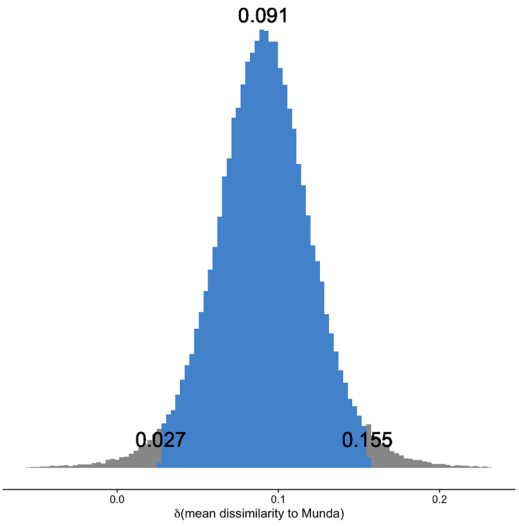


Figure 9: Bayesian t-test posterior distribution histogram.

16 https://CRAN.R-project.org/package=BEST.

17 A credible interval is the interval in which an (unobserved) parameter has a given probability. It differs from a frequentist confidence interval because, unlike a confidence interval, the credibility interval is dependent on the prior distribution. In addition, in a credible interval, the parameter is treated as a random variable, while the bounds are considered fixed.

dissimilarity of Munda being close to 0 (that is, that neither the eastern nor the western groups are more similar to Munda) is much lower than the relative probability that the difference in mean dissimilarity of Munda is larger than zero (that is, that languages in the eastern cluster are more similar to Munda on average). Moreover, the graph suggests that the highest relative probability (highest peak on the *y* axis) is around 0.1 – which is the median value of the distribution. This suggests that the languages in the eastern cluster are on average 0.1 points more similar to Munda than the languages in the western cluster.

To evaluate whether the posterior distribution of the Bayesian t-test supports our hypothesis, we proceed as follows. First, we report the 95% Highest Density Interval (HDI) – that is, the interval on the *x* axis that contains 95% of the most credible values for the difference in mean dissimilarity to Munda given the evidence. The choice of 95% mimics the standard choice of 5% significance threshold, as willingness to dismiss the least probable 5% of outcomes as unlikely. On a posterior graph, the HDI interval is shaded and its boundaries (the left 2.5% boundary and the right 2.5% boundary) as well as its middle (median) are reported as numbers. The Highest Density Interval for Figure 9 is thus reported as HDI [0.027; 0.155]. If all values within this interval are larger than zero, we conclude that our hypothesis is supported by the data for the given feature group (that is, languages in the eastern cluster are more similar to Munda on average). Otherwise, we conclude that our hypothesis is not supported by the data for the given feature group. In addition, we can use the HDI to compare the results across different feature groups. This allows us to detect feature groups for which the split between the two clusters is more evident (such groups will have the HDI positioned closer to 1.0 on the *x* axis).

5 Results

In Table 1 we show an overview of the results. We explored the presence of an east–west divide in the Indo-Aryan languages of our sample and, in the presence of a divide, we computed the similarity score of the eastern cluster to the languages of the Munda group. We undertook this procedure for the entire dataset, for each questionnaire and each feature subgroup. Each feature subgroup described above is included in one of the main questionnaires; they have been selected and analyzed among the various domains collected in the questionnaires for their exhaustivity: each of them is described through a rich set of variables. The table summarizes the results: for each feature group, we reported the presence of a geographical east–west divide (first column) and the similarity to Munda based on the Highest Density Interval of the posterior distribution that we obtained from the

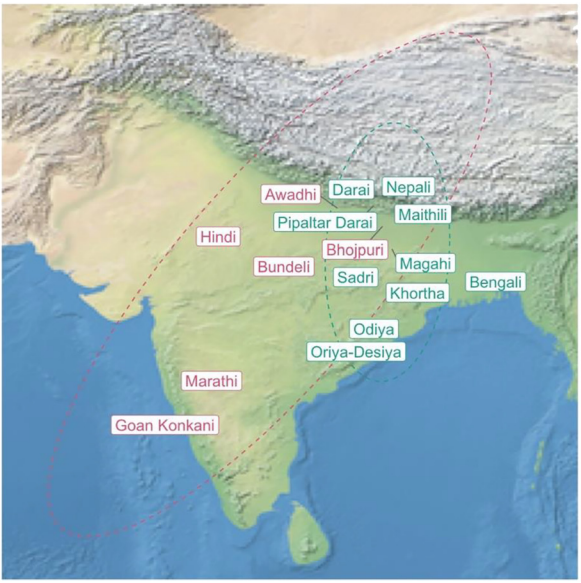
Bayesian t-test (second column). The meaning of these values and their interpretation are explained in detail in Section 4.

The results reported in Table 1 illustrate the presence of an east–west divide in the entire dataset, in each of the questionnaires and in some specific grammatical domains: classifiers, comparative constructions, ergativity, interclausal syntax, noun inflection, noun classes, noun phrase and number. Other features do not reveal any significant east–west divide. Rather, results from case, echo constructions, negation, possession, pronouns, relative clause, syntax and verbal inflection seems to reflect a more homogeneous pattern across the Indo-Aryan languages of the sample.

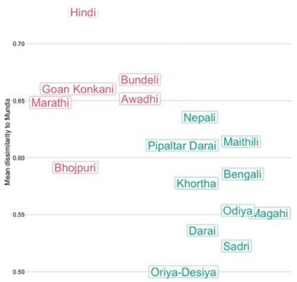
For all the cases where we attested the presence of an east–west divide, we computed the difference of similarity score to Munda between eastern and western clusters. The results suggest that the languages belonging to the eastern clusters are more similar to Munda, compared to the languages in the western group, for most of the feature groups. These groups include the entire dataset, each of the questionnaires, and feature subgroups such as classifiers, comparative constructions, ergativity, interclausal syntax and noun inflection. Cluster analysis on feature subgroups such as noun classes, noun phrase and grammatical number reveals the presence of an east–west divide; for these features, however, there is no support for the languages in the respective eastern clusters being more similar to the languages of the Munda group.

In what follows, we illustrate in more detail the results pertaining to some selected feature groups. We will focus on the results pertaining to the entire dataset and each of the questionnaires. We will then move onto the discussion of the results from the selected features ergativity and noun inflection. We will touch upon on the noun phrase, the results of which indicate an east–west divide with weak support of the similarity to Munda hypothesis, and we will briefly illustrate results from negation and echo formation, features that do not show any significant clustering.

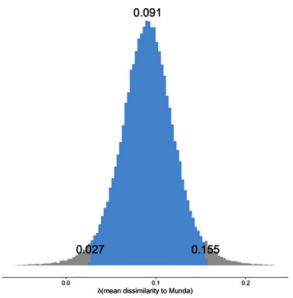
The results for the entire dataset take into account all of the 217 variables coded for each Indo-Aryan language of the sample. Cluster analysis produces two groups: one includes Hindi, Awadhi, Bundeli, Marathi, Bhojpuri and Goan Konkani; the other contains Oriya-Desiya, Bengali, Sadri, Maithili, Darai, Khortha, Magahi and Nepali. The geographical distribution (Figure 10a) clearly reflects an east–west divide within the Indo-Aryan languages surveyed, with a group in the West and a group in the East, as also confirmed by the visual inference protocol which we performed on the plot. The mean dissimilarity to Munda for each Indo-Aryan language is shown in Figure 10b, and follows our prediction that the eastern group is more similar to Munda. The posterior probability distribution of the difference of means (Figure 10c) further supports our prediction: the median of the posterior difference in mean dissimilarity is 0.091 and the 95% HDI [0.27; 0.155] does not include zero.



(a) Geographical distribution



(b) Mean dissimilarity to Munda



(c) Posterior distribution

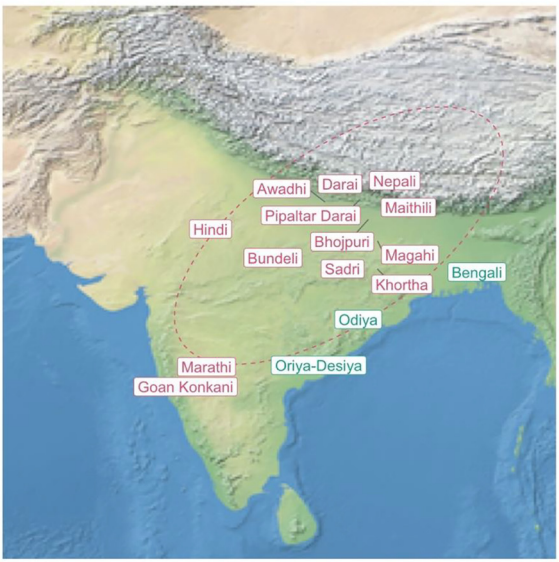
Figure 10: Indo-Aryan clusters (all variables).

The exploration of each individual main feature group (Questionnaires A, B and C) shows similar results: although we observe a slight variation in terms of the languages attested for each cluster, the east–west Indo-Aryan divide is nevertheless preserved across the different parameter sets, along with the eastern group being more similar to Munda.

The analysis performed on Questionnaire A (broad typological features), taken in isolation, does not give overall strong support to our hypotheses. Cluster analysis identifies one group of Indo-Aryan languages, located geographically in the West, and another small group of languages spoken in the eastern part of the region, as confirmed by the visual inference protocol. Languages belonging to the western group are Bhojpuri, Hindi, Awadhi, Bundeli, Marathi, Maithili, Darai, Pipaltar Darai, Goan Konkani, Magahi and Nepali, as seen in Figure 11a. The languages in the East group include Odiya, Bengali and Oriya-Desiya. We find weak support related to the similarity of the languages of the eastern cluster to Munda as illustrated in Figure 11b, as confirmed by the posterior distribution of the difference of means, which displays a median of 0.035, 95% HDI [−0.304; 0.570] (see Figure 11c), when compared to the previous histogram that illustrates the posterior distribution for the whole features set. This is not surprising, since Questionnaire A is made up of many heterogeneous features. However, when we observe the whole set (as in Figure 10a), or the individual feature groups (discussed below), we find stronger support for our hypothesis.

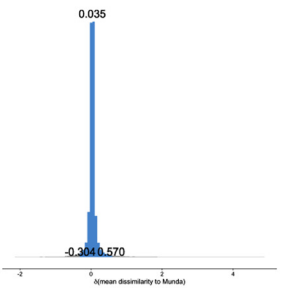
The feature set dedicated specifically to the fine-grained exploration of features (Questionnaire B) exhibits analogous results in terms of the presence of an east–west divide in the Indo-Aryan languages, and its geographical divide is also confirmed by the visual inference "Line-up" protocol. Cluster analysis on the parameters of Questionnaire B shows a western cluster that includes Nepali, Darai, Pipaltar Darai, Hindi, Bhojpuri, Bundeli, Marathi, Goan Konkani and Oriya-Desiya. The second cluster includes five languages: Maithili, Magahi, Bengali, Sadri and Odiya (Figure 12a), all spoken in the East. Average dissimilarity to Munda for each Indo-Aryan language is shown in Figure 12b, where we observe that all the languages of the eastern cluster have a higher similarity to Munda languages, when compared to the western languages. In addition, Oriya-Desiya and Bhojpuri, which cluster to the western group, but are geographically located in the East, show higher similarity values to Munda, similar to the languages of the eastern group. The posterior distribution of the difference of the means of the two clusters (shown in Figure 12c) confirms the trend, with a median of 0.124 and the 95% HDI excluding zero [0.034; 0.316].

The last main feature set (Questionnaire C), which expands the data in Peterson (2017a), produces the results in Figure 13a. The east–west divide appears robust, and is supported by the visual inference protocol. The western group comprises Awadhi,



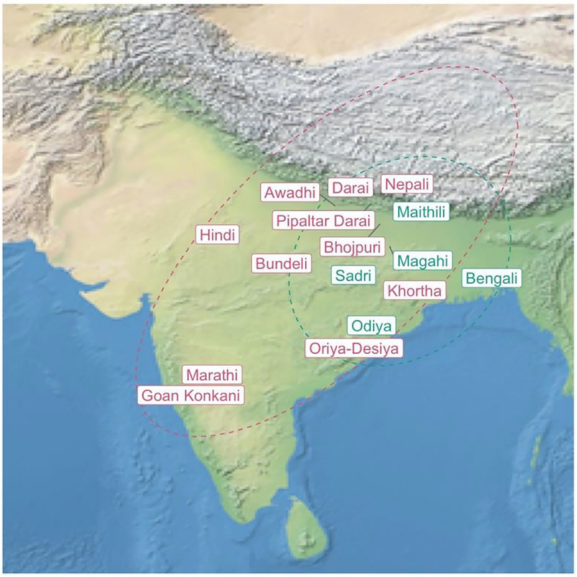
(a) Geographical distribution

(b) Mean dissimilarity to Munda

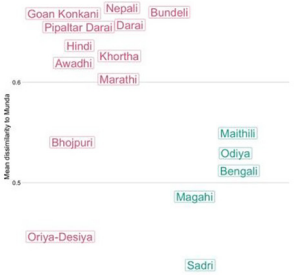


(c) Posterior distribution

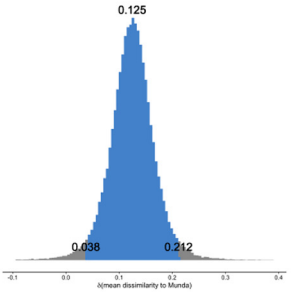
Figure 11: Indo-Aryan clusters (Questionnaire A).



(a) Geographical distribution

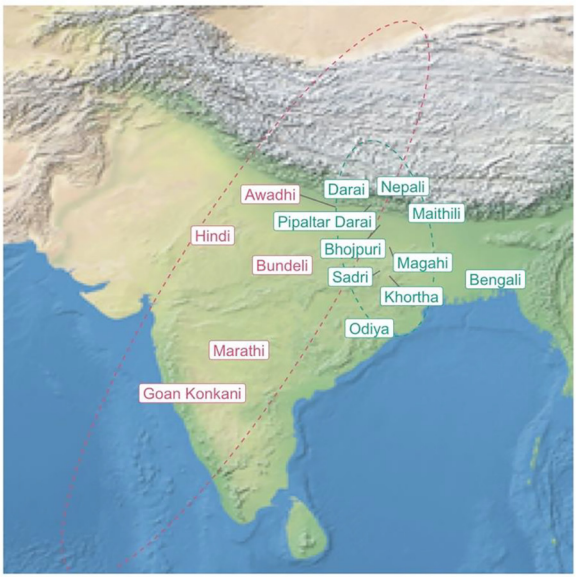


(b) Mean dissimilarity to Munda

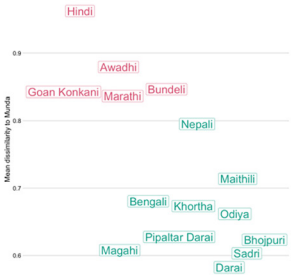


(c) Posterior distribution

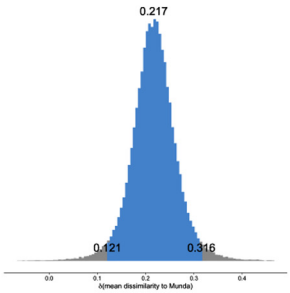
Figure 12: Indo-Aryan clusters (Questionnaire B).



(a) Geographical distribution



(b) Mean dissimilarity to Munda



(c) Posterior distribution

Figure 13: Indo-Aryan clusters (Questionnaire C).

Bundeli, Hindi, Marathi and Goan Konkani; the eastern group includes Darai, Maithili, Nepali, Bhojpuri, Sadri, Pipaltar Darai, Bengali, Magahi, Khortha and Odiya.[18] The eastern cluster group is again more similar to Munda (Figure 13b). The posterior distribution in Figure 13c further confirms our hypotheses, with a mean value of 0.217 and the 95% HDI [0.121; 0.316], thus excluding zero.

The exploration of parameters that describe specific grammatical features confirms the trend seen above for a number of features. In what follows we discuss the results for ergativity, noun inflection and noun phrase features.

Many Indo-Aryan languages are characterized by morphological ergativity, with different ergativity patterns across languages and mostly conditioned by verb morphology (as perfect and non-perfect past tenses). Mostly western Indo-Aryan languages, such as Hindi and Marathi, have retained split ergativity patterns. (1) illustrates the presence of split ergativity along the dimension of aspect in Hindi.

(1) Hindi

a.	*ram*	*gaɽi*	*cʌla-ta*	*(hai)*
	Ram.M.SG.NOM	car.F.SG.NOM	drive-IPFV.M.SG	be.PRES.3.SG
	‘Ram drives a car.’			
b.	*ram=ne*	*gaɽi*	*cʌla-yi*	*(hai)*
	Ram.M.SG=ERG	car.F.SG.NOM	drive-PERF.F.SG	be.PRES.3.SG
	‘Ram has driven a/the car’			
	(from Butt in Coon et al. 2017)			

On the other hand, many languages of the eastern subgroup are known to have lost ergative marking, such as Oriya-Desiya, Bengali and Sadri, the latter shown in (2).

(2) Sadri

a.	*hamre=mʌn=ke*	*bʌɽaik=kʌr*	*upadhi*	*hɪya=kʌr*	*raja=mʌn*	
	1PL=PL=OBL	Baraik=GEN	title	here=GEN	king=PL	
	de-l-ʌ̃					
	give-PST-PL					
	‘The kings from here (=Jharkhand) gave us the title of “Baraik”’					
b.	*i*	*sadi=kʌr*	*badme*	*mo=ke*	*nebhi*	*bʌʈ-e*
	3SG	wedding=GEN	after	1SG=OBL	navy	road-LOC
	le-i	*ge-l-ʌ̃*				
	take-LNK	V2:TEL-PST-3HON				
	‘and after the wedding he took me off to the Navy’					
	(Peterson, field notes)					

18 Oriya-Desiya is not present, since there are too few datapoints for Questionnaire C, as seen in Figure 7.

Exceptions include varieties such as Nepali and Assamese, both of which are in close contact with Tibeto-Burman languages (Masica 2001: 247). We also find ergativity patterns in eastern Indian regions such as Jharkhand, albeit very limited: Khortha (Indo-Aryan) displays ergative marking on nouns and first person pronouns only in the perfective aspect, as shown in (3).

(3) Khortha

a. *ham-ẽ* *pechu* *mʌhina* *e=go* *kʌmij* *sil-wa-l-ie*
1SG-ERG last month one=CLF shirt stitch-CAUS-PST-1
'I had someone stitch a shirt last month'

b. *ham* *agu* *mʌhina* *e=go* *ghar* *bʌn-wa-bʌ-i*
1SG next month one=CLF house build-CAUS-FUT-1
'I will have someone build a house next month'

c. *ham-ẽ* *piji* *tin* *sal-ẽ* *pura* *kʌr-l-i=o*
1SG-ERG PG three year-LOC finish do-PST-1=ADDR
'I finished (my) post graduate in three years'

d. *budhna-ẽ* *aij* *bihane* *sap=ke* *(ʈheŋga-ẽ)*
Budhna-ERG today morning snake=OBL stick-INS
ɖheɽaw-l-ʌi
beat-PST-3SG
'Budhna killed the snake this morning (with a stick)'
(Paudyal, field notes)

By contrast, Munda languages do not display ergativity as shown in Kharia (example 4).

(4) Kharia

a. *moɲ* *dinu* *aba=ɖom* *sou$^{?}$b=te* *remakh=o?*
one day father=3POSS all=OBL call=ACT.PST
ro *gam=o?* *[…]*
and say=ACT.PST
'One day their father called them all and said "…"'

b. *la?* *sou$^{?}$b=ga* *bhai=ki* *juda* *juda*
then all=FOC brother=PL separate REP
mu?=ki=may, *kinir=te*
emerge=MID.PST=3PL forest=OBL
'Then all the brothers set out separately, into the forest'
(Peterson 2010a: 441–442)

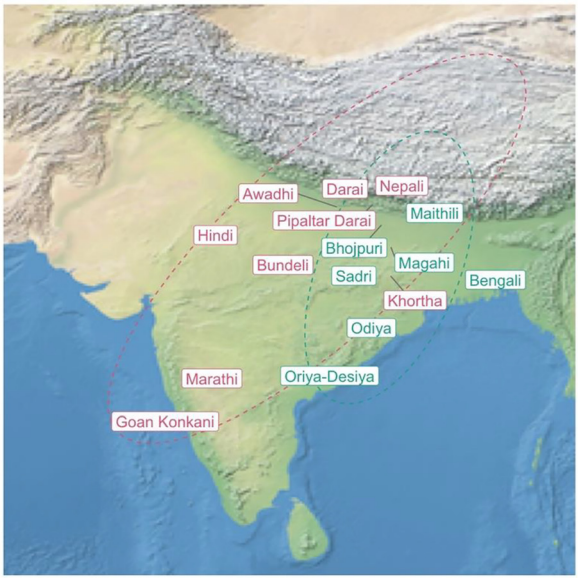
In the cluster analysis for ergativity, we again obtain a geographical divide that follows the east–west distinction, as shown in Figure 14a and confirmed by the visual inference test. Ergativity offers the strongest support to our second hypothesis when compared to the other feature groups: all the languages grouped in the eastern cluster display maximal similarity to Munda, as this is clearly visible in the distance to Munda plot (cf. Figure 14b). This similarity is reflected in the posterior distribution in Figure 14c, where the difference of means between the two groups is 1.

Noun inflection also supports our hypotheses, while the noun phrase partially supports it: while we still detect the presence of an east–west divide, the latter case is not supported by a higher similarity to Munda for the respective eastern cluster. The presence of a geographical east–west divide after cluster analysis for both feature groups, with some differences in the composition of the obtained clusters, is illustrated for noun inflection in Figures 15a and 16a for noun phrase, while the results pertaining to the average similarity to Munda of each eastern cluster are illustrated in Figures 15b and 16b.

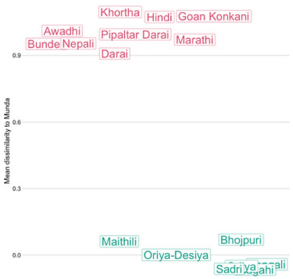
The posterior distributions for both feature groups favor our hypotheses for the feature noun inflection (95% HDI [0.030; 0.450], zero excluded), while it is less favorable over the noun phrase domain (95% HDI [–0.080; 0.252], zero included). The posterior distribution histograms are presented in Figures 15c and 16c.

We conclude our overview by illustrating feature subgroups that do not show any relevant clustering. These features include negation and echo constructions. Both domains tend to be stable across Indo-Aryan languages, with negation being expressed through a standard negation particle (/nahi/) and echo constructions being a South Asian areal feature, characterized by a striking homogeneity in both form and functions across the Indo-Aryan language family (Abbi 1992). Geographical distributions of the obtained clusters are illustrated for both negation and echo formation domains in Figure 17a and b, respectively.

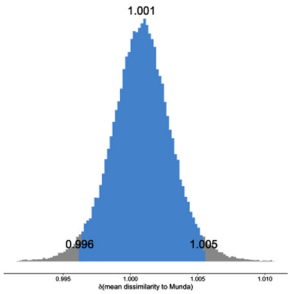
In conclusion, we found evidence of the presence of an east–west divide in the Indo-Aryan languages of our sample. The presence of the geographical divide is clear from the exploration of the entire dataset (217 traits for 27 Indo-Aryan languages) and also by investigating in detail each of the main feature sets (Questionnaires A, B, and C), each of them treating different domains and with a specific depth. We obtained similar results by observing individual feature groups, where we have found clear presence of an east–west divide in most of them. Not surprisingly, not all the features under investigation show an east–west divide (such as negation and echo formations), and this different behavior offers significant insights in attesting the depth of this divide, and which grammatical domains are involved in language contact. All the languages included in the respective eastern clusters for the entire set, and each main feature group and feature subgroups that



(a) Geographical distribution

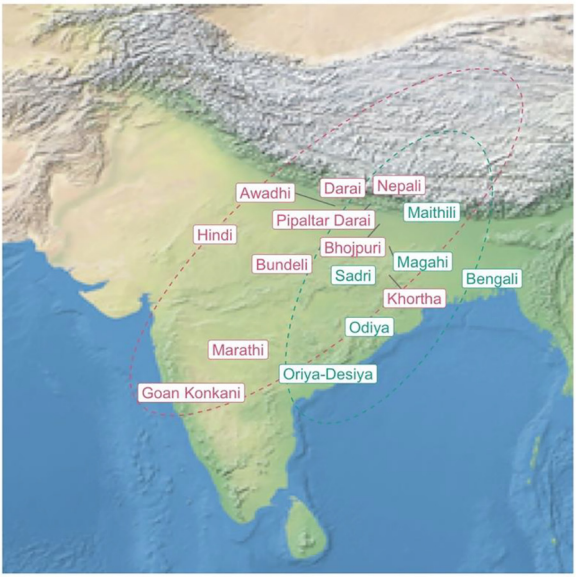


(b) Mean dissimilarity to Munda



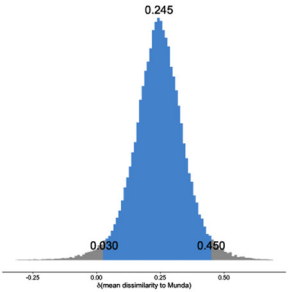
(c) Posterior distribution

Figure 14: Indo-Aryan clusters (ergativity).



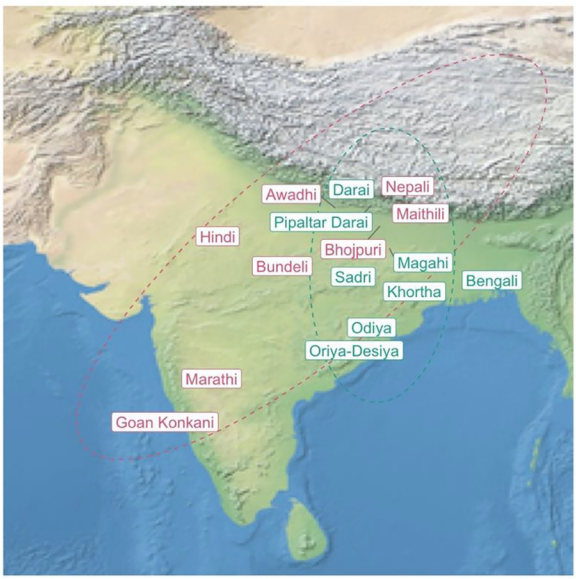
(a) Geographical distribution

(b) Mean dissimilarity to Munda



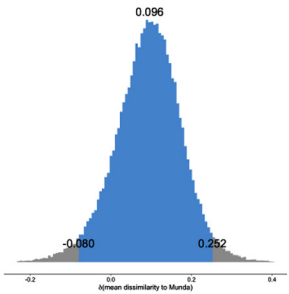
(c) Posterior distribution

Figure 15: Indo-Aryan clusters (noun inflection).



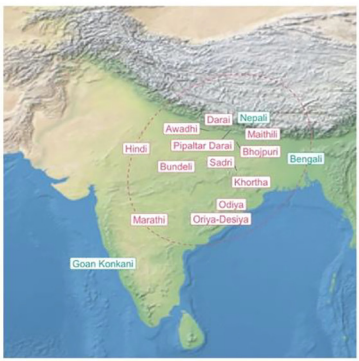
(a) Geographical distribution

(b) Mean dissimilarity to Munda



(c) Posterior distribution

Figure 16: Indo-Aryan clusters (noun phrase).



(a) Geographical distribution of negation clusters

(b) Geographical distribution of echo formations

Figure 17: Indo-Aryan clusters (negation and echo formation).

show a geographical divide, have a higher similarity to Munda. We take up this point again in the following section.

6 Conclusions and implications for earlier contact in the eastern Indo-Gangetic Basin

In this article we have shown that there are systematic structural differences between eastern and western Indo-Aryan and that the eastern Indo-Aryan languages are consistently closer to Munda than western Indo-Aryan languages. Our results thus confirm the east–west divide found in Peterson (2017a). This is

especially significant since the present study is based on a considerably larger data base than that earlier study: although the number of languages investigated here does not differ significantly from that in Peterson (2017a), the present study involves up to 217 features per language, whereas the previous study involved only up to 28 features per language.[19]

In the previous section we suggested that the east–west divide confirmed in the present study is likely due to contact between eastern Indo-Aryan languages and Austro-Asiatic languages. The evidence for this can be summarized as follows: there is a clear east–west divide within Indo-Aryan with respect to all of the features investigated in the present study as well the questionnaires and most of the individual feature groups. Not all feature groups show such a cluster, e.g. negation and echo constructions, which are much more homogeneous throughout Indo-Aryan, with only very minor deviations.[20] However – and most importantly – where we do find evidence for an east–west divide within Indo-Aryan, the eastern group is almost always significantly more similar to the Munda languages than the western group.

Assuming that the east–west divide within Indo-Aryan does in fact result from an earlier contact in the eastern Indo-Gangetic Basin, as we believe it does, then this earlier contact later likely led to an Austro-Asiatic substrate in eastern Indo-Aryan. There are several reasons for assuming this.

To begin with, as argued in Peterson (2017a), a typological split of the type we observed within Indo-Aryan is entirely unexpected in a spread zone, unless some natural boundary is involved which divides eastern and western Indo-Aryan into two separate geographical areas. However, as Figure 3 shows, this is clearly not the case for the Indo-Gangetic Basin. Furthermore, the relevant similarities which we observed between eastern Indo-Aryan and Munda (i.e., Austro-Asiatic) point to influence from Austro-Asiatic on eastern Indo-Aryan and not the other way around. Although Indo-Aryan influence might seem likely at first glance, given the much larger number of speakers of these languages when compared with Munda, such an explanation would not account for the clear schism between eastern and

19 The present article contains data from 27 languages while Peterson (2017a) is based on data from 29 languages. However, that earlier study also contains data from seven Dravidian languages which are excluded from the present article, so that the present article contains five more languages from Indo-Aryan and Munda than the earlier study.

20 This suggests that the "South Asian language area" best fits what Campbell (2017: 27) refers to as a "trait-sprawl area" or "TSA". This is a type of contact area in which some features are found "crisscrossing some languages while others crisscross other languages, with some extending in one direction, others in another direction, with some partially overlapping others in part of their distribution but also not coinciding in other parts of their geographical distribution". This is in contrast to the "linguistic area *sensu stricto*" or "LASS", in which traits are shared across the languages of a clearly definite geographical area (Campbell 2017: 28).

western Indo-Aryan which the present study has confirmed. On the other hand, assuming Austro-Asiatic influence on eastern Indo-Aryan accounts both for the east–west divide within Indo-Aryan as well as the smaller typological distance between eastern Indo-Aryan and modern-day Munda languages. The question then remains: what type of contact phenomena are we dealing with here?

It seems that the most likely candidate is that of a substrate, resulting from the gradual wholesale shift from Austro-Asiatic to Indo-Aryan. As was just noted, there is clear evidence that Austro-Asiatic influenced eastern Indo-Aryan, otherwise we would not expect to find such clear evidence of an east–west divide there. However, as Austro-Asiatic is now largely confined to peripheral regions of eastern India and has all but vanished from the Indo-Gangetic Basin, we conclude that these languages have disappeared from these regions as entire ethnic groups which once spoke them have given up their traditional languages in favor of Indo-Aryan, a process which is still taking place in this region today. Thus, an eastern Austro-Asiatic substrate would easily account for:

1. the east–west divide in Indo-Aryan,
2. the typological similarities between eastern Indo-Aryan and Munda today (as well as the corresponding distance between western Indo-Aryan and Munda), as well as
3. the fact that these Austro-Asiatic languages are no longer spoken throughout most of eastern India.

Other languages were also surely spoken in this region at that time, such as perhaps the predecessors of the modern-day isolates Nihali and Kusunda, and earlier Dravidian languages. There were also certainly other languages spoken in this region which have only left indirect traces of their existence, such as the unknown source of many common words in Kurmali (see Paudyal and Peterson, this issue). Nevertheless, as Austro-Asiatic languages are by far the most widespread non-Indo-Aryan languages of this region, we assume that it was speakers of these languages who contributed most to this substrate in eastern Indo-Aryan.

Conventional wisdom has it that speakers of Indo-Aryan languages first entered the Indian subcontinent through the northwest and gradually migrated eastwards. In such a model, these speakers will have initially constituted a very small minority in eastern India. With respect to explaining how the language of an initially small number of Indo-Aryan speakers came to serve as *lingua franca* of the entire eastern Indo-Gangetic Basin of the time, and later went on to displace almost all of the indigenous languages in this region, we present here some tentative thoughts with respect to the likely dynamics involved.

We can assume that the Indo-Aryan speakers possessed a certain military advantage over the peoples of the eastern Gangetic Basin. However, this alone cannot explain the eventual predominance of the language of the newcomers, as

they were certainly vastly outnumbered by indigenous groups at the time. Clearly, factors other than mere military might were also at play.

Unfortunately, our task is hampered by the fact that we cannot be sure how many and which indigenous ethnic groups of this region at that time were hunter-gatherers, small farmers, pastoralists, or perhaps practiced a combination of these different forms of subsistence. Also, it is not entirely clear how many and what type of Indo-Aryan speakers first made their way eastwards, e.g. did this group consist primarily of young men (so-called "founders") or was it a larger migration involving whole family units? These questions are important since the respective lifestyles of the various ethnic groups will have played an important role in the power relationship between the indigenous and the Indo-Aryan speaking ethnic groups and their languages. For example, hunter-gatherer groups will presumably have lived in smaller communities with fewer speakers per language[21] and would likely have been multilingual in the languages of their neighbors (cf. e.g. Barnard 2016). Such groups who did not flee to higher, less accessible areas will likely have quickly assimilated to the *lingua franca* and culture of the newcomers.[22]

With sedentary farming groups or pastoralists, however, the possible scenarios become more numerous and complex. While it is often assumed that sedentary farming groups' languages will prevail over those of semi-nomadic groups, numerous cases are also documented where a relatively small, mobile minority was eventually able to install its language as the majority language.[23] In fact, the main assets of farming peoples, i.e. their crops and homes, can actually become a liability, subject to raids by more mobile groups (Pereltsvaig and Lewis 2015: 210). As such, a detailed account of the mechanisms of language shift to Indo-Aryan in the eastern Indo-Gangetic Basin must await future research into the status of these different groups. However, contact scenarios of this type involving asymmetrical power structures are now quite well understood,[24] so that we believe

21 E.g. Hammarström (2016: 22) estimates that such languages will have an average size of 1,000 speakers or less.

22 Cf. e.g. Heggarty (2015: 620): "… the hunter-gatherers' languages, if they survive at all, invariably end up cantoned into inhospitable areas of little value to agriculturalists."

23 Cf. e.g. Anthony and Ringe (2015: 209), who discuss such a scenario with respect to the spread of Indo-European languages to agricultural communities "in the absence of empire or conquest."

24 There is considerable literature on possible outcomes involving asymmetrical power constellations in situations of language contact. For example, Hill (1996:1) bases her theory of anthropological dialectology on social (in-)equality. She writes: "People with secure primary claims on essential resources are more likely to favor localist stances, while people who lack adequate primary claims and draw instead on a diverse range of secondary or indirect claims are more likely to favor distributed stances. Distributed stances encourage the spread of sociolinguistic variables, while localist stances inhibit spread." This topic will be dealt in more detail in a future study.

that we can provide at least a general outline here of how this language shift in eastern India may have come about.

As noted above, we can safely assume that the early Indo-Aryan speakers in the eastern Indo-Gangetic Basin possessed a certain military advantage over the indigenous groups of this region, and this will likely also have held with respect to technology in general. These advantages also included writing, although as Heggarty (2015: 619) notes, this is in and of itself not decisive but probably more indicative of the overall socio-cultural differences between different ethnic groups in such encounters. We believe that there was one further, decisive advantage in favor of Indo-Aryan in this region, namely their control of long-distance trade and the vast networks necessary for it, which afforded many opportunities and prestige to those who could speak the language of the newcomers. As Heggarty (2015: 622) notes, this plays a key role in such contact situations: "In linguistic terms, a new, wider speech community has been forged, within which the incomers' language is unique in enjoying wide currency, prestige and utility right across it."

If this assumption is correct for the eastern Indo-Gangetic Basin in the mid-first millennium BCE, this could help to explain how a presumably small group of Indo-Aryan speakers eventually succeeded in making its own language the *lingua franca* of the entire region, later forms of which would then replace most indigenous languages of the region. These latter languages disappeared almost entirely, with a small number surviving in peripheral, less hospitable mountainous and hilly regions, while most of the indigenous languages of that time have left only indirect proof of their existence in the form of substrate effects in the new languages of the land.

7 Outlook – the broader perspective

Recent works on linguistic prehistory in South Asia stress the need for an interdisciplinary approach to the linguistic history of South Asia, such as Masica (2001: 258–259) or Peterson (2017a: 246). In the present section we briefly note two areas in which this interdisciplinary work has already led to interesting results.

One notable study from religious history, noted already in Peterson (2017a), is Bronkhorst (2007). In his work, Bronkhorst notes that the culture of the eastern Indo-Gangetic Plain, which he refers to as "Greater Magadha", differed considerably in the first millennium BCE from the Brahmanical culture of the west. He also cites archeological evidence for this divide. While his work is based primarily on religious–historical argumentation and the thrust of his argumentation is quite different from ours, it is notable that the two lines of research have come to rather similar results with respect to determining two geographically and culturally/linguistically distinct regions in northern India. Future work comparing the results reached in

these two different lines of research will undoubtedly provide us with important information about the linguistic and cultural situation in this region at that time.

Perhaps the most dynamic research at present on the prehistory of this region is within the field of genetics, research which is also in line with the results of the present study. For example, a number of publications by David Reich and his associates (e.g. Reich 2018; Reich et al. 2009) have identified two major genetic groups in South Asia, which they label "ANI" for "Ancestral North Indians" and "ASI" for "Ancestral South Indians". The first group, ANI, is ultimately related to western Eurasians whose presence in South Asia goes back to migrations into the subcontinent from the Eurasian steppe and who likely brought Indo-European languages to South Asia. The second group, ASI, is hypothesized to be deeply related to the Andaman Islanders and to represent a very old genetic stock which has been in South Asia for several millennia.[25] Reich and his colleagues speak here of the "Indian Cline", which primarily refers to the different proportions of Western Eurasian related ancestry in the genetic make-up of an individual or ethnic group. This is illustrated in Figure 18, from Reich (2018: 131), further annotated by the authors of this study according to the data in Reich et al. (2009: 492). For the larger perspective, "Europeans" are given at the far upper left in Figure 18, and "East Asians" and "Sino-Tibetan speakers" at the upper right of the figure.

The lower part of Figure 18 presents a number of individual members of modern South Asian ethnic groups along the Indian Cline, with Kashmiri Pundits at the upper end, with more ANI ancestry, and several Dravidian-speaking groups such as the Kurumba at the lower end, with a considerably smaller percentage of ANI ancestry. A line running through the Figure from the top to the bottom in the middle separates individuals (i.e., the individual points) with more eastern ancestry (to the right) from those with more western ancestry (to the left). Similarly, a line running horizontally through the lower half of the figure separates individuals with more northern (upper) from those with more southern (lower) ancestry.

In the lower right-hand quarter, i.e., for individuals with a more eastern and southern ancestry, we find members of two Austro-Asiatic–speaking tribal groups, the Kharia (South Munda) and Santali (North Munda), outside of the Indian Cline. Note that there are also a number of individuals from Indo-Aryan-speaking groups on the "eastern" side of the diagram, close to these two Austro-Asiatic-speaking

25 The development is actually considerably more complex than this brief description suggests. Cf. e.g. the discussion in Narasimhan et al. (2016) and Shinde et al. (2019) on the role of Indus-Periphery–related ancestry in South Asia in the formation of ANI and ASI. This adds a further level of complexity to the model, namely the "Indus Periphery Cline" which will not be dealt with here for reasons of space.

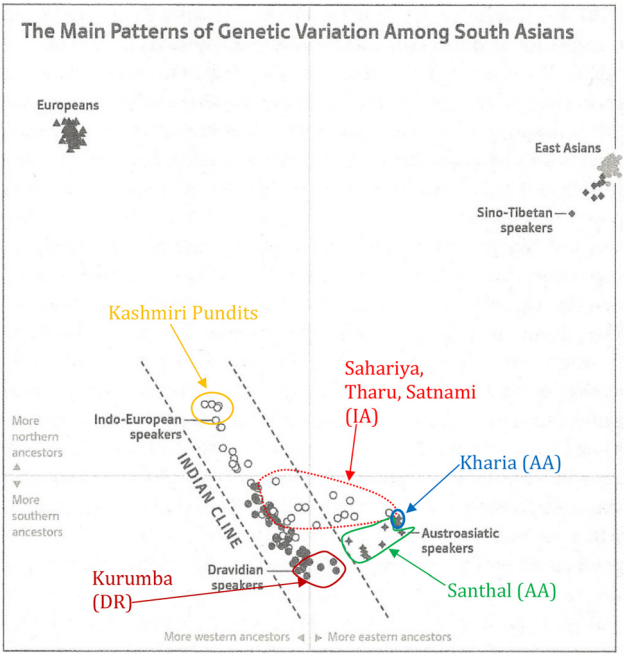


Figure 18: The main patterns of genetic variation among South Asians.[26]

groups. This includes e.g. members of the Sahariya (the outer four of the five dots directly to the right of the Indian Cline), a "low-caste" Indo-Aryan speaking group whose members for this study are from Uttar Pradesh in central northern India. Members of the Satnami ethnic group are also found scattered throughout the red circle in Figure 18, another "low-caste" Indo-Aryan–speaking group, whose members for this study are from Chhattisgarh in eastern central India, and the Tharu, an Indo-Aryan–speaking tribal group found primarily in the Nepalese lowlands. The easternmost Indo-Aryan speaker within this red circle is a member of the Tharu ethnic group, and the "inner" individual in the group of five Indo-Aryan speakers just to the right of the Indian Cline is a member of the Satnami group. The Indo-Aryan speakers within the area contained simultaneously within both the red circle and the Indian Cline belong to both the Tharu and the Satnami groups.

In our present age of linguistic mass extinction, an increasing number of speakers of Austro-Asiatic languages are choosing not to pass their traditional

26 Reproduced from Figure 17b in Reich (2018: 131) and annotated with information on various ethnic groups according to the discussion in Reich et al. (2009: 492). Reproduced here with the kind permission of David Reich, which we gratefully acknowledge.

languages on to their children, favoring instead the regional and supra-regional Indo-Aryan languages, such as Sadri, Hindi, etc. Figure 18 suggests that this process has been going on for quite some time, as it shows that some ethnic groups with a more eastern ancestry, historically associated with speakers of Austro-Asiatic languages, must have switched to Indo-Aryan at some earlier stage in their history. This provides a historical backdrop for the Austro-Asiatic substratum in eastern Indo-Aryan argued for in the present study, which has left an indelible mark on the linguistic structures of eastern Indo-Aryan languages.

While this or similar developments have long been assumed in linguistic studies, genetic studies are now providing further convincing proof for these large-scale prehistoric language shifts, and interdisciplinary work of this type promises to yield many more new and exciting insights into the prehistory of South Asia.

List of abbreviations

1, 2, 3	Person
ACT	active
ADDR	addressee
AMB	ambulative
CAUS	causative
CLF	classifier
ERG	ergative
F	feminine
FOC	focus
FUT	future
GEN	genitive
HON	honorific
INS	instrumental
IPFV	imperfective
LNK	linker
LOC	locative
M	masculine
MID	middle voice
NOM	nominative
OBL	oblique
PL	plural
PRS	present
PST	past
REP	repetition
SG	singular
TEL	telic
V2	vector verb

References

Abbi, Anvita. 1992. *Reduplication in South Asian languages: An areal, typological, and historical study*. New Delhi: Allied Publishers.

Abbi, Anvita. 1997. Languages in contact in Jharkhand. In Anvita Abbi (ed.), *Languages of tribal and indigenous peoples of India. The ethnic space*, 131–148. Delhi: Motilal Banarsidass.

Abbi, Anvita. 2009. Is Great Andamanese genealogically and typologically distinct from Onge and Jarawa? *Language Sciences* 31(6). 791–812.

Anderson, Gregory D. S. 2008. Introduction to the Munda languages. In Gregory D. S. Anderson (ed.), *The Munda languages (Routledge Language Family Series 3)*, 1–10. London & New York: Routledge.

Anthony, David W. & Don Ringe. 2015. The Indo-European homeland from linguistic and archaeological perspectives. *The Annual Review of Linguistics* 1(1). 199–219.

Barnard, Alan. 2016. *Language in prehistory*. Cambridge: Cambridge University Press.

Berkhin, Pavel. 2006. A survey of clustering data mining techniques. In Kogan Jacob, Charles Nicholas & Marc Teboulle (eds.), *Grouping multidimensional data. Recent advances in clustering*, 25–71. Berlin: Springer.

Bloch, Jules. 1934. *L'indo-aryen du Veda aux temps modernes*. Paris: Adrien-Maisonneuve.

Bronkhorst, Johannes. 2007. *Greater Magadha: Studies in the culture of early India*. Leiden: Brill.

Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne & Hadley Wickham. 2009. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences* 367(1906). 4361–4383.

Campbell, Lyle. 2017. Why is it so hard to define a linguistic area. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 19–39. Cambridge: Cambridge University Press.

Coon, Jessica, Massam, Diane & Lisa deMena, Travis (eds.). 2017. *The Oxford handbook of ergativity*. Oxford & New York: Oxford University Press.

Eberhard, David M., Simons, Gary F. & Fennig, Charles D. (eds.). 2019. *Ethnologue: Languages of the world*. Dallas, TX: SIL International. Available at: http://www.ethnologue.com (accessed 2 April 2021).

Ebert, Karen H. 1993. Kiranti subordination in the South Asian areal context. In Karen H. Ebert (ed.), *Studies in clause linkage: Papers from the First Köln-Zürich Workshop*, 83–110. Zurich: Seminar für Allgemeine Sprachwissenschaft, Universität Zürich.

Ebert, Karen H. 1999. Nonfinite verbs in Kiranti languages – an areal perspective. In Yogendra P. Yadava (ed.), *Topics in Nepalese linguistics*, 371–400. Kathmandu: Royal Nepal Academy.

Emeneau, Murray B. 1956. India as a linguistic area. *Language* 32(1). 3–16.

Enfield, Nick J. 2005. Areal linguistics and mainland Southeast Asia. *Annual Review of Anthropology* 34. 181–206.

Gower, John C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27(4). 857–871.

Grambank Consortium. 2019. *Grambank*. Jena: Max Planck Institute for the Science of Human History.

Gumperz, John J. & Robert Wilson. 1971. Convergence and creolization: A case from the Indo-Aryan/Dravidian border in India. In Dell H. Hymes (ed.), *Pidginization and creolization of languages*, 151–167. Cambridge: Cambridge University Press.

Hammarström, Harald. 2016. Linguistic diversity and language evolution. *Journal of Language Evolution* 1(1). 19–29.
Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019. *Glottolog 4.1*. Jena: Max Planck Institute for the Science of Human History.
Heggarty, Paul. 2015. Prehistory through language and archaeology. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 616–644. London & New York, NY: Routledge. https://doi.org/10.4324/9781315794013-42.
Hill, Jane H. 1996. *Languages on the land: Toward an anthropological dialectology (David Skomp Distinguished Lectures in Anthropology series)*. Bloomington: Indiana University, Department of Anthropology.
Kaufman, Leonard & Peter J. Rousseeuw. 2009. *Finding groups in data: An introduction to cluster analysis* (Wiley Series in Probability and Statistics 344). New York: John Wiley & Sons .
Kerman, Jouni, Andrew Gelman, Tian Zheng & Yuejing Ding. 2008. Visualization in Bayesian data analysis. In Chun-houh Chen (ed.), *Handbook of data visualization*, 709–724. Berlin: Springer. https://doi.org/10.1007/978-3-540-33037-0_27.
Kruschke, John K. 2013. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General* 142(2). 573–603.
Masica, Colin P. 2001. The definition and significance of linguistic areas: Methods, pitfalls, and possibilities (with special reference to the validity of South Asia as a linguistic area). In Peri Bhaskararao & Karumuri Venkata Subbarao (eds.), *The yearbook of South Asian languages and linguistics 2001*, 205–267. London: SAGE. https://doi.org/10.1515/9783110245264.205.
Masica, Colin P. 1976. *Defining a linguistic area (South Asia)*. Chicago: The University of Chicago Press.
Moser, Barry K. & Gary R. Stevens. 1992. Homogeneity of variance in the two-sample means test. *The American Statistician* 46(1). 19–21.
Narasimhan, Vagheesh M., Karen A. Hunt, Dan Mason, Christopher L. Baker, Konrad J. Karczewski, Michael R. Barnes, Anthony H. Barnett, Bates Chris, Srikanth Bellary, Nicholas A. Bockett, Kristina Giorda, Christopher J. Griffiths, Harry Hemingway, Zhilong Jia, M. Ann Kelly, Hajrah A. Khawaja, Lek Monkol, Shane McCarthy, Rosie McEachan, Anne O'Donnell-Luria, Kenneth Paigen, Constantinos A. Parisinos, Eamonn Sheridan, Laura Southgate, Louise Tee, Mark Thomas, Yali Xue, Michael Schnall-Levin, Petko M. Petkov, Chris Tyler-Smith, Eamonn R. Maher, Richard C. Trembath, Daniel G. MacArthur, John Wright, Richard Durbin & David A. van Heel. 2016. Health and population effects of rare gene knock-outs in adult humans with related parents. *Science* 352(6284). 474–477.
Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press.
Nichols, Johanna. 1997. Modeling ancient population structures and movement in linguistics. *Annual Review of Anthropology* 26(1). 359–384.
Osada, Toshiki. 1991. Linguistic convergence in the Chotanagpur area. In S. Bosu Mullick (ed.), *Cultural Chotanagpur: Unity in diversity*, 99–119. New Delhi: Uppal Publishing House.
Pereltsvaig, Asya & Martin W. Lewis. 2015. *The Indo-European controversy*. Cambridge: Cambridge University Press.
Peterson, John. 2010a. *A grammar of Kharia: A South Munda language*. Leiden: Brill.
Peterson, John. 2010b. Language contact in Jharkhand: Linguistic convergence between Munda and Indo-Aryan in eastern-central India. *Himalayan Linguistics* 9(2). 56–86.

Peterson, John. 2015a. From "finite" to "narrative" – The enclitic marker *= a* in Kherwarian (North Munda) and Sadri (Indo-Aryan). *Journal of South Asian Languages and Linguistics* 2(2). 185–214.

Peterson, John. 2015b. Introduction – advances in the study of Munda languages. *Journal of South Asian Languages and Linguistics* 2(2). 149–162.

Peterson, John. 2017a. Fitting the pieces together – towards a linguistic prehistory of eastern-central South Asia (and beyond). *Journal of South Asian Languages and Linguistics* 4(2). 211–257.

Peterson, John. 2017b. Jharkhand as a "linguistic area" – language contact between Indo-Aryan and Munda in eastern-central South Asia. In Raymond Hickey (ed.), *The Cambridge handbook of areal linguistics*, 551–574. Cambridge: Cambridge University Press.

Reich, David. 2018. *Who we are and how we got here: Ancient DNA and the new science of the human past*. Oxford: Oxford University Press.

Reich, David, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price & Lalji Singh. 2009. Reconstructing Indian population history. *Nature* 461(7263). 489–494.

Rousseeuw, Peter J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20. 53–65.

Shinde, Vasant, Vagheesh M. Narasimhan, Nadin Rohland, Swapan Mallick, Matthew Mah, Mark Lipson, Nakatsuka Nathan, Nicole Adamski, Broomandkhoshbacht Nasreen, Ferry Matthew, Lawson Ann Marie, Michel Megan, Oppenheimer Jonas, Stewardson Kristin, Jadhav Nilesh, Kim Yong Jun, Chatterjee Malavika, Munshi Avradeep, Panyam Amrithavalli, Waghmare Pranjali, Yadav Yogesh, Patel Himani, Kaushik Amit, Thangaraj Kumarasamy, Meyer Matthias, Patterson Nick, Rai Niraj & Reich. David. 2019. An ancient Harappan genome lacks ancestry from steppe pastoralists or Iranian farmers. *Cell* 179(3). 729–735.

Welch, Bernard L. 1937. On the z-test in randomized blocks and Latin squares. *Biometrika* 29(1/2). 21–52.

Wickham, Hadley, Dianne Cook, Heike Hofmann & Andreas Buja. 2010. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics* 16(6). 973–979.

Zide, Norman H. 1969. Munda and non-Munda Austroasiatic languages. In Thomas Sebeok (ed.), *Current trends in linguistics*, vol. 5, 411–430. The Hague: Mouton.