Supplementary material

Contents

1 Basic vocabulary data

We recorded word forms and cognacy judgments across 207 meanings in 103 contemporary and ancient languages. Cognates are homologous words, related by common ancestry. To be diagnosed as cognate the words must have similar meaning and, most importantly, show systematic sound correspondences. For example, the English word '*five*' has cognates in German (*fünf*), Swedish (*fem*) and Dutch (*vijf*), reflecting descent from proto-Germanic (**fimf*). New cognate forms can be gained in a language when the language begins to use a new word for a given meaning. For example, at some point in the Indo-European lineage leading to the Germanic languages, Proto-Indo-European **ast(h)-*, meaning '*bone*', was replaced with Proto-Germanic **bainan*. Here the Proto-Indo-European form was lost, and a new form gained. The new form gave rise to a set of cognate words for bone in descendant Germanic languages. Cognate identification is non-trivial: other cognates of these words for '*five*' include Irish *cuig*, Italian *cinque*, Armenian *hing* and Polish *piec*. Known borrowings, such as English *mountain* acquired from French *montagne*, were not coded as cognate.

Contemporary language data were sourced from the Comparative Indo-European Lexical Database (*17*) and other published sources (Table S1). We excluded 11 speech varieties from ref. (*17*) identified by the authors of ref. (*17*) as duplicate dialects based on a less reliable source of cognacy judgments. We excluded another 14 duplicate sources for regional dialects ('doculects'), always favouring the doculect that had the least missing data. Ancient languages were sourced from ref. (*16*). These comprise the best attested ancient Indo-European languages, representing all of the major sub-groups. Using this set also allows us to avoid the charge of cherry-picking certain languages. Whilst our aim here is not to provide a complete history of every Indo-European lineage, future work could expand this set with additional ancient languages. For example, partial data on Sogdian, an ancient Eastern Iranian language, may provide more insight into the colonization of the Steppes. However, we note that Sogdian would sit neatly within our current Iranian clade and so is unlikely to have any effect on the inferred date or homeland of Proto-Indo-European.

Cognate data were coded as binary characters showing the presence or absence of a cognate set in a language. There were 5047 cognate sets in total, with most meanings represented by several different cognate sets. All cognate coding decisions were checked with published historical linguistic sources (Table S1). The database contained 25908 cognate coded lexemes. Of these, 67% came originally from ref. (*17*), 14% from ref. (*16*), and 19% were newly compiled from published sources. Ref. (*17*) required considerable correction, and changes were made to approximately 26% of coding decisions on individual lexemes. Ref. (*16*) required corrections to only 0.5% of lexemes.

The full dataset is available online here: `http://ielex.mpi.nl`.

2 Location data

Cognate data were combined with information on the geographic distribution of each language as recorded in the Ethnologue (*31*) based on digitized language maps (available from Global Mapping International `http://www.gmi.org`). This data allows us to assign each observed language an approximate range, rather than conditioning on a single point location. We excluded post-colonial ranges of languages. Some languages were nevertheless spoken in multiple unconnected ranges. It is not practical to assign tip locations to a number

of unconnected ranges because the Gibbs operator (see below) will rarely jump from one range to another. In order to overcome this problem, ranges that were close (most often the case) were joined into a single contiguous range. Where separate ranges were far apart or one population was much larger, the language was assigned to the range with the larger population. The ranges for ancient languages were based on the geographic distribution of source texts. Figure S4 shows the geographic range distribution for each of the 103 sampled languages, with the ancient languages highlighted in red.

3 Bayesian Phylogenetic Inference of Language Trees

Following previous computational approaches to the evolution of languages (*1–3*), we model language change as the gain and loss of cognates through time, using Bayesian phylogenetic inference to infer the likely set of plausible language trees, given our data and model. This model-based, Bayesian approach offers a number of advantages over previous approaches (*32*, *33*). First, any attempt to infer language ancestry from comparative data requires implicit or explicit simplifying assumptions about the processes of change involved and optimality criteria for evaluating trees. Computational modelling makes the optimality criteria and assumptions of the method explicit in the specification of the structure of the model and prior beliefs. Second, there are a vast number of possible language trees to evaluate - for the 103 languages in our dataset there are more possible trees than there are atoms in the universe (*34*). There is also stochastic uncertainty inherent in the system, such that the 'best' fitting tree may not in fact be the 'true' tree. Bayesian inference of phylogeny provides a principled framework with which we can efficiently explore the universe of possible language trees and quantify uncertainty in the inferred relationships and model parameters (*35*). This means conclusions incorporate uncertainty in the tree and model parameters and are not contingent on a specific tree topology. The method assumes a bifurcating tree (each lineage can only give rise to two daughter lineages), however, since the intervals between lineage splits can be arbitrarily small, we can effectively accommodate multifurcations (lineages splitting into many descendent lineages in a short time period) if the data

support this. Third, we can compare the fit of a range of models of language lineage evolution and spatial diffusion, including different cognate replacement models and relaxing the assumption of constant rates of change across the tree. We can also use simulated data to evaluate the accuracy of our findings and how robust results are to violations of model assumptions. We can therefore be confident that, for example, the binary coding of the cognate data allows accurate phylogenetic inference, and that inferences are not impaired by the presence of realistic rates of borrowing between lineages (*32*, *36*, *37*). Finally, we can explicitly test between competing origin hypotheses by quantifying their relative support, given our data and model (*1*, *2*, *32*).

Given a binary matrix representing the presence (1) or absence (0) of each cognate set in each language, we model the process of cognate gain and loss using transition rates defined as the probability of a cognate appearing ($0 \rightarrow 1$) or disappearing ($1 \rightarrow 0$) along each branch over a given amount of time. Starting at the root of the tree, an ancestral Indo-European language comprising some set of cognates thus evolves through descent with modification into the Indo-European languages we observe today.

Let $D = \{D_1, \ldots, D_m\}$ represent m columns of cognate presence/absence data, with each column spanning n languages. Data element $D_{j,k}$ ($1 \leq j \leq m, 1 \leq k \leq n$) indicates the presence or absence of cognate j in language k. The distribution of interest is the distribution of language trees given the cognate data, that is, $P(T|D)$ where T is the tree. Using Bayes theorem, this can be interpreted through

$$P(T|D) \propto P(T)P(D|T) \tag{1}$$

where $P(T)$ is the prior on the tree, $P(D|T)$ the likelihood and $P(T|D)$ the posterior. Considering individual cognates,

$$P(D|T) = \prod_{j=1}^{m} P(D_j|T)$$

The tree T has languages at its leaves. The branches of the tree are labeled with time, which allows us to write the likelihood for the jth cognate $P(D_j|T)$

as the probability of the tree marginalized over all the states of internal nodes,

$$P(D_j|T) = \sum_{x_{n+1},\ldots,x_{2n-1}} P(x_{2n-1}) \prod_{k=n+1}^{2n-2} P(x_k|x_{\pi(k)}, t_k) \prod_{k=1}^{n} P(D_{j,k}|x_{\pi(k)}, t_k)$$

where $x_1, \ldots, x_n$ are the leaf nodes containing cognate data, and $\pi(k)$ the index of parent node of node x_k, so $x_{\pi(k)}$ is the parent node of node x_k. By convention x_{2n-1} is the root node. Further, t_k is the length of the branch ending in node x_k. This looks like a formidable formula, but using the peeling algorithm (*38*) it can be calculated in linear time in the number of nodes, and quadratic in the cognate state space, making the computation tractable in practice. This allows us to calculate a likelihood for any tree, given a set of cognate data and model of cognate replacement. We combine this with Bayesian inference of phylogeny using Markov chain Monte Carlo methods as implemented in the BEAST software package (*20*). This approach allows us to efficiently sample trees and model parameters in proportion to their posterior probability, given our data, model and prior beliefs.

Cognate Substitution Models

We evaluated a series of models of cognate evolution and report results for the best fitting model, although the specific choice of model did not affect our main findings.

To obtain transition probabilities $P(x_k|x_{\pi(k)}, t_k)$, that is the probability that node x_k is of a particular binary cognate value given the value of its parent in the tree and the time elapsed t_k, we considered three different models of cognate substitution: reversible Continuous-time Markov chains (CTMC), covarion (*39*, *40*) and Stochastic Dollo processes (*19*, *41*).

The reversible continuous-time Markov chains are traditionally used for analysis of DNA data with many transition model structures existing, allowing uneven transition rates to and from certain states (e.g. the HKY model (*42*)). Under this model the change from every state to every other state and back is realizable in a finite amount of time. For a binary state-space (cognate absence or presence) the model can be parameterized with a single parameter γ, which captures the relative proportion of forward changes to backward changes. Equation (2) shows the infinitesimal time rate matrix of binary reversible CTMC where entry (i, j) represents the rate at which character i changes to character

j.

$$\begin{array}{c} 0: \\ 1: \end{array} \begin{pmatrix} - & \gamma \\ 1 & - \end{pmatrix} = Q \tag{2}$$

The finite-time transition probabilities for this CTMC model satisfy the Chapman-Kolmogorov equation

$$\dot{P}(t) = \Delta t P(t) Q \text{ with initial conditions } P(0) = I,$$

where Δt a small time step and I is the identity matrix. The solution is $P(t) = \exp(tQ)$. So, we calculate the transition probability of going from character i to character j over time span t as the exponent of t times Q, i.e. $P(x_k = i | x_{\pi(k)} = j, t) = \exp(tQ)_{i,j}$.

The covarion model extends the reversible Markov chain models to allow for cognates to transition from actively changing to non-changeable states (*39*, *40*), consistent with linguists' intuition that rates of change for certain meanings may differ at some points on the tree. The covarion has two additional parameters (δ and κ), which govern the transition between actively changing (variant) and non-changeable states (invariant). The infinitesimal rate matrix for the binary covarion model is summarised as

$$\begin{array}{rc} \text{variant} \left\{ \begin{array}{c} 0: \\ 1: \end{array} \right. \\ \text{invariant} \left\{ \begin{array}{c} 0: \\ 1: \end{array} \right. \end{array} \begin{pmatrix} - & \gamma & \delta & 0 \\ 1 & - & 0 & \delta \\ \kappa\delta & 0 & - & 0 \\ 0 & \kappa\delta & 0 & - \end{pmatrix} \tag{3}$$

The Stochastic Dollo process (*19*, *41*) applies what may be a more natural model of cognate evolution by postulating that a cognate can only arise once (with Poisson rate λ). In essence, this restriction ensures that each cognate is uniquely evolved and each cognate creation event generates a new cognate, unrelated to any other pre-existing cognate. Under the stochastic Dollo model, once in existence the cognate survives with a constant death rate μ. Upon a death event occurring in any lineage the cognate cannot re-emerge in that lineage via a backward mutation. That is, it forever remains in the absent state. The infinitesimal rate matrix of this process is

$$\begin{array}{c} 0: \\ 1: \end{array} \begin{pmatrix} 0 & 0 \\ \mu & - \end{pmatrix} \tag{4}$$

The invariant cognates were intentionally excluded from our analyses. This creates an ascertainment bias for the estimation of the parameters of the substitution models. To account for this bias we employ a commonly used technique

of ascertainment correction (recently described in (*19*)). In essence, this correction re-scales the finite-time transition probabilities, to take into account the fact that invariant sites have been removed and thus are unobservable. Gray et al. (*43*) have recently shown that this approach produces consistent and unbiased estimates of evolutionary rates, while still realizing the speed improvements from reducing the number of sites through exclusion of invariant cognates.

Calibrating rates of change

Branch lengths, t_k, are scaled in time units using an additional parameter that governs the rate of evolution (*44*). The rate of cognate replacement per unit time is calibrated by constraining lineage divergence times (at internal nodes) and ancient language ages (at the tips) based on historical sources after ref. (*1*, *16*) as listed in Table S2. We used probability distributions rather than point estimates for the age of each internal node and ancient language, in order to account for uncertainty in the historical data. Tip ages were numerically integrated via sampling as recently described by Shapiro et al. (*45*).

Since languages may not evolve at the same rate at every location through time, we compare the fit of a strict clock model (which assumes a constant rate of cognate replacement) to a relaxed clock model (*46*) that allows for rate heterogeneity among lineages. To calculate the transition probability of going to x_k from the parent of x_k in time t under an uncorrelated relaxed clock model, $P^{\text{relaxed}}(x_k|x_{\pi(k)}, t)$. The relaxed clock accommodates rate heterogeneity among branches with a rate distribution $P(r)$. By relaxing the clock assumption in this way, we can accommodate variation in rate of cognate replacement through time, estimating the degree to which rates vary from the data itself. Figure S1 summarizes how inferred rates of cognate replacement vary across branches in the tree.

MCMC Inference and Model Testing

We investigate the posterior distribution of each model using Markov Chain Monte Carlo (MCMC) integration in BEAST (*20*). We use a flexible tree prior based on a multiple change-point process implemented in the Bayesian skyline plot (BSP) (*47*). The BSP-based prior is based on coalescent theory, which describes the times two branches in a tree are expected to join into a single branch. This defines a distribution over the heights of the internal nodes in a tree that is influenced by an unknown population function of the BSP. Since this function

is very flexible and random, the BSP-based prior imparts little information on the inferred tree, allowing variation in the rate of lineage formation through time without the need to specify a parametric model. We use the Tracer tool in BEAST to examine the convergence of MCMC runs and the TreeAnnotator tool in BEAST to summarize trees in the form of maximum clade credibility (MCC) trees.

To compare different models, we employed an importance sampling estimator of the marginal likelihood that is frequently used to obtain (the natural logarithm of) Bayes factors for Bayesian phylogenetic and coalescent model comparison in an MCMC framework (*48*, *49*). Model comparison results are shown in Table S3.

The Stochastic Dollo substitution model showed the best fit to the data. This fits with the notion that its assumption that cognates are gained once and then differentially lost in descendant lineages is a more natural model of vocabulary evolution. The better fit of this model suggests that processes that allow a cognate to be gained multiple times on the tree (e.g. parallel semantic shift, the borrowing of basic vocabulary terms or sound changes that independently give rise to the same sound-meaning correspondence) are relatively rare in our data. This result is also consistent with previous work showing that rates of borrowing in basic vocabulary are much lower than in the wider vocabulary (*50*, *51*).

The uncorrelated log-normal relaxed clock was the preferred clock model, indicating that the rates of language change in our data vary across different lineages. This finding is also consistent with previous studies showing variation in rates across time in Indo-European (*52*) and other language families (*2*), and between different subsets of language features (*53*, *54*), presumably as an outcome of different divergence processes, technological advances, and usage patterns (*2*, *53*). The results presented in the paper are based on this best-fitting model although the specific choice of model did not affect our main findings.

Unlike previous studies, we did not assume a known outgroup. Instead this was inferred from the data under the assumption of a relaxed lexical clock. Our inferred outgroup (Anatolian) is consistent with the orthodox view in Indo-European linguistics (*55*). Figure S1 shows high posterior support for all of the established Indo-European sub-groups. Relationships between groups are

also well-supported, with uncertainty around the positioning of only the basal sub-groups - Tocharian, Albanian, Greek and Armenian. Some of these basal relationships and the generally high clade posterior probabilities differ slightly from (*1*). This likely reflects the improved model of cognate replacement, expanded data set and the inclusion of geographic information.

4 Spatial Diffusion Models

We connect the cognate evolution model with stochastic processes of spatial diffusion in a joint inference framework. To do this, we apply the same approach as for cognate evolution to infer internal node locations from the language ranges at the tips of the tree (see Figure S6). That is, we assume that languages disperse as they evolve through time such that we can model their dispersal through space along the branches of the language tree inferred together with cognate data. Note that the tree is informed both by cognate data and geographical data. For each of the extant languages in our sample, we know the ranges R where they are spoken (see Figure S6). So, we extend Equation (1) by assuming that, given the tree T, the cognates and ranges are independent via

$$P(T|D,R) \propto P(T)P(D,R|T) = P(T)P(D|T)P(R|T). \tag{5}$$

The terms $P(T)$ and $P(D|T)$ are obtained as before, such that we only need to specify the probability of language ranges given a tree $P(R|T)$. To get a handle on $P(R|T)$, we exploit data augmentation by introducing unknown locations $L = (l_1, \ldots, l_{2n-1})$ of the tree tips l_k for $k = 1, \ldots, n$, the internal nodes l_k for $k = n+1, \ldots, 2n-2$ and the root l_{2n-1} and consider

$$P(R|T) = \int P(R|L)P(L|T)\mathrm{d}L. \tag{6}$$

We set $P(R|L)$ equal to the indicator function that tip locations $(l_1, \ldots, l_n)$ all fall within their respective ranges and decompose $P(L|T)$ as the product of the transition probabilities over all internal nodes of dispersal to a node location from its parent node location in the tree over a time specified by the length of the branch. Taking the root node distribution into account gives us

$$P(L|T) = P(l_{2n-1}) \prod_{k=1}^{2n-2} P(l_k|l_{\pi(k)}, t_k), \tag{7}$$

where $l_{\pi(k)}$ is the location of the parent of l_k in the tree, connected by branch length t_k. Interest often lies in the inference of the posterior $P(l_{2n-1}|D,R)$ that arises during the data augmentation step in MCMC integration of $P(T|D,R)$.

We pursued two alternative models to infer the root location, $P(l_{2n-1}|T)$, a phylogeographic relaxed random walk (RRW) model in continuous space (*14*) (Section 4.1) and a discretized landscape based geographical model (Section 4.2). Despite some novel extensions introduced here, the RRW model assumes the same migration rates over land and over water. To verify that this assumption does not affect the outcome of the analysis, we developed a set of models that takes different migration rates for land and water into account. This was achieved by discretizing space so that we can run diffusion simulations on a relatively fine grid where the pixels represent either water or land. The difference between the RRW model and the landscape based model lies in the way transition probabilities $P(l_k|l_{\pi(k)},t_k)$ and the root prior distribution $P(l_{2n-1})$ are determined.

4.1 Relaxed Random Walk Geographical Model

We extend a Bayesian implementation of multivariate diffusion models recently developed for phylogeographic analysis of viruses (*14*). Rather than using a simple diffusion model, this approach takes advantage of the fact that we have information about the ancestral relationships between sampled languages. We consider spatial diffusion along the branches of an unknown yet estimable phylogeny as a generalized Brownian motion process, or 'random walk', in two-dimensional space, and exploit data augmentation of the unobserved locations (longitude and latitude) at the root and internal nodes of the tree. A random walk, based on movement via successive steps, each in a random direction, is routinely used to model the spread of organisms across a landscape (*56–58*). In the case of language, this does not mean that individual language migration events occur 'randomly', without social, cultural or ecological drivers. Rather, for a given time interval, the geographic distribution of languages expanding from some point of origin is assumed to be approximated by a Brownian random walk some languages will have moved far, some will not have moved at all, but most will have moved somewhere in between.

The Brownian random walk process we use here is governed by an infinites-

imal precision matrix **P** that scales arbitrarily in units-time. The Bayesian implementation specifies a Wishart prior on **P** and a bivariate normal prior on the root location. The most restrictive assumption of such random walk models – a constant diffusion rate throughout the phylogeny – can be relaxed by rescaling the precisions along each branch similar to relaxed clock models (*14*, *46*). Here, we further advance these relaxed random walk (RRW) models by accommodating arbitrarily shaped flat spatial prior distributions on the root, internal and external node locations. We use these geographically informed priors to exclude root and internal node realizations from areas over water and to accommodate a geographic distribution for the location of each sampled language, rather than conditioning on point locations. For estimation via Markov chain Monte Carlo under this improved framework, we modified the Gibbs samplers (*59*) for the root and internal node locations to take into account the geographical priors and also accommodate updates for the tip locations taking into account the geographical priors. We compared the fit of different distributions (Cauchy, gamma and lognormal) that relax the homogeneous Brownian random walk (BRW). A relaxed random walk provided a significantly better fit to the language diffusion process with the lognormal-RRW yielding the best fit for the data (See Table S4).

This approach means that rather than assuming a constant rate of diffusion, we can infer how regular the language expansion process is by estimating the degree of rate variation from the data. The best-fitting lognormal-RRW model indicates the rate of expansion is low on average (mean = 480 meters per year) but highly variable (coefficient of variation = 4.35) (Table S4). Crucially, despite relaxing the assumption of a constant rate of diffusion in this way, our results show that there is enough regularity in the expansion process to strongly support an Anatolian, over a Steppe origin. Figure 2 summarizes how inferred rates of geographic diffusion vary across branches in the tree.

4.2 Landscape Based Geographical Model

The RRW model does not distinguish between geographical features such as land and water for its migration rates. In order to be able to model different migration rates between such features, we introduce a feature-rich geographical model where different locations can have different migration patterns. We model

the process as inhomogeneous diffusion along the branches of a tree according to a continuous time Markov chain (CTMC) model. As we describe below in detail, this leads to an ordinary differential equation (ODE) which we can solve numerically. However, there are various obstacles to making such models computationally tractable. In particular computing the transition probabilities for a large number of locations to another large number of locations over a large number of time intervals requires a lot of memory.

Furthermore, calculating the root location can be done efficiently using the peeling algorithm, but only up to state spaces of size about a thousand different states. Therefore, we limit ourselves to approximately 600,000 locations and agglomerate these locations into a 32 by 32 grid. The choice of the grid needs to handle projection distortions due to the curvature of the earth. Instead of storing (or recalculating) transition probabilities for all time intervals determined by the tree, we store the probabilities only for a limited number of times and interpolate to get desired probabilities for a particular time of migration.

Below, we first describe details of the model and how it leads to an ODE. Then, we define the neighborhood and work out parameterization followed by a description of the details of the ODE solver and usage of the solution. Finally, we describe implementation details of relaxation for the landscape based model.

Model Details

First, we describe the mathematical details for the landscape based model before going into implementation details. The probability of the locations of the nodes in the tree, given the tree, is given by Equation (7). We will use $P(L|T)$ to determine the root locations given the locations of the tips in the tree, that is, $P(l_{2n-1}|l_1, \ldots, l_n, T)$ where $l_1, \ldots, l_n$ are the tip locations and the root location l_{2n-1}. To obtain transition probabilities, $P(l_k|l_{\pi(k)}, t)$, we assume a fixed discretization of space with orthogonal geometry. We define a neighborhood $\phi(i)$ for the set of locations $i = 1 \ldots N$ containing the locations to the left, right, top and bottom of location i for all locations that do not lie on the boundary. Hence, the cardinality of such neighborhood $|\phi(i)|$ is 4.

Let $T(i) \in \{1, \ldots, K\}$ specify a geographical type of location i. Here we differentiate between water and land location types, however, the approach we have developed could in principle be extended to incorporate other geographic

features such as mountains or deserts. Consider a $K \times K$ matrix $R = \{r_{k\ell}\}$, where $r_{k\ell}$ is the infinitesimal transition rate (proportional to the finite-time transition probability) for moving from location type k to type ℓ given that two locations are adjacent. Then infinitesimal-time rate matrix $\Lambda = \{\lambda_{ij}\}$, can be defined as

$$\begin{aligned} \lambda_{ij} &= r_{T(i)T(j)} \text{ for } j \in \phi(i) \\ \lambda_{ii} &= -\sum_{j \neq i} \lambda_{ij}, \\ \lambda_{ij} &= 0 \text{ elsewhere} \end{aligned}$$

The finite-time transition probabilities for this CTMC model satisfy the Chapman-Kolmogorov equation

$$\dot{P}(t) = \Delta t P(t) \Lambda \text{ with initial conditions } P(0) = I, \tag{8}$$

The solution of this ODE is

$$P(t) = \exp(t\Lambda). \tag{9}$$

Computing the matrix exponential in (9) is best approached through numeric approximation. For very small Δt, a first-order Taylor expansion

$$\exp(\Delta t \Lambda) \approx I + \Delta t \Lambda = P$$

furnishes approximate equality between a discrete time Markov chain (DTMC) and CTMC representations. The corresponding DTMC unfolds as

$$\begin{aligned} p_{ij} &= r_{T(i)T(j)} \times \Delta t \text{ for } j \in \phi(i), \\ p_{ii} &= 1 - \sum_{j \neq i} p_{ij}, \\ p_{ij} &= 0 \text{ elsewhere} \end{aligned} \tag{10}$$

where Δt is a small time increment. This is known as the Euler method for solving the ODE from Equation (8).

Neighborhood Definition and Parameterization

Now the ODE is derived we are ready to describe the method of discretization and parameterization. With the Euler method it is computationally feasible

to deal with up to 600,000 locations, which we will refer to as pixels in the remainder, to calculate transition probabilities (that is p_{ij} in Equation (10)). To determine the distribution over the root location, a variant of the peeling algorithm can be used. The peeling algorithm will not be able to handle a state space of 600,000 states but can fairly conveniently deal with state spaces of around a thousand. Therefore, we aggregate individual segments into a 32 by 32 rectangular grid such that the blocks in the grid cover the whole area in the discretized landscape.

One of the issues in discretizing space over the area covered by the Indo-European languages is that the curvature of the earth has considerable influence on the sizes of the grid if the grid is drawn in standard Mercator projection. This influences the ODEs as detailed below. To minimize distortions of the grid block sizes, we used a Mercator projection that first rotates the earth by 0.07π radians, since the languages are on an axis that is slightly slanted. Then, an 'equator' is drawn through the center of the language ranges and then a standard Mercator projection is applied. In other words, the earth is slightly rotated, then turned and then a Mercator like projection applied. The resulting grid covers an area that is 9655 km wide at the center of the longer axis and 4885 km high at the center of the smaller axis. Figure S7 shows the grid in this projection. The top image shows the grid in the rotated and translated Mercator projection, and the bottom image shows the grid drawn in a standard Mercator projection, which shows significant distortions of the grid.

The language areas can cover a number of grid blocks, and for the case of Russian a large number of grid blocks. So, to compensate for this fact we average over the areas covered by the locations (similar to using ambiguities in standard phylogenetic analysis). Let L_k denote the set of locations covered (or partially covered) by language k and $P_k(l)$ the amount of coverage for language k. This amount of coverage is determined by calculating the area of land inside a grid block covered by the language range. Equation (7) can be refined to take tip location distributions into account like this,

$$P(L|T) = P(l_{2n-1}) \prod_{k=n+1}^{2n-2} P(l_k|l_{\pi(k)}, t_k) \prod_{k=1}^{n} \sum_{l \in L_k} P(l|l_{\pi(k)}, t_k) P_k(l)$$

The transition probabilities over the aggregated locations $P(l_k|l_{\pi(k)}, t_k)$ are determined by the ODE over the underlying map as follows; underneath the grid

block lies a pixel map of 1104 by 560 pixels. Thus, a pixel covers on average 8.7 by 8.7 km. A pixel is either land or water. To determine $P(l_k|l_{\pi(k)}, t_k)$ a probability mass of 1 is distributed over all land pixels in a grid block. If there are no such pixels because the grid block covers water only, we assume no migration out of that grid block occurs and encode the distribution of $P(l_k|l_{\pi(k)}, t_k)$ as 1 when $x_{\pi(k)} = x_k$ and zero otherwise for all evolution times t_k. If there is any land in the grid block, we perform the Euler method to standard diffusion by allowing a fraction of the probability mass to move for each time step. The move is either to the left, right, above or below pixel and the rate is determined by the contents of the pixels. To specify the model, we need to define the transition probabilities of the DTMC of Equation (10). Define $\rho_{T(i),T(j)} = r_{T(i)T(j)}\Delta t$. We distinguish three parameters;

- $\rho_{l,l}$ the proportion going from a land pixel to a land pixel,
- $\rho_{l,w}$ the proportion going from a land pixel to a water pixel,
- $\rho_w = \rho_{w,l} = \rho_{w,w}$ the proportion going out of a water to either land or water pixel

We use these parameters to define four different models of geographic diffusion. First, when all three proportions are equal, we obtain a discretized approximation of the continuous space diffusion model - water and land are indistinguishable. Since settlement on water is impractical, a more realistic parameterization is to set ρ_w to one, which implies that whenever one is in the water, all of the probability mass moves out in the next time step. To allow sufficient expedient movement over water, $\rho_{l,l}$ and $\rho_{l,w}$ should be relatively small. Furthermore, the ratio $\rho_{l,w}/\rho_{l,l}$ determines how slowly probability mass moves into water; the lower the ratio the slower the movement - i.e. the less likely is migration across water. In our computations, we set $\rho_{l,l}$ to 1% for all models (including the diffusion model). We used the ratio $\rho_{l,w}/\rho_{l,l}$ to define three additional models representation varying probabilities of movement into water. A ratio of 0.1 was used for the model labeled '10 times less likely to go into water'. A ratio of 0.01 was used for '100 times less likely to go into water'. Finally, we used a ratio of 1 to define a 'sailor' model, in which movement from land into water is as likely as from land to land. Results under each model are summarized in Table 1.

After a number of steps in the Euler method for grid block k, for each of the grid blocks we record how much of the probability mass moved to the pixels underlying that particular grid block and use this as an estimate for $P(l_k|l_{\pi(k)}, t)$ where we normalize t so that at the end of the ODE solver we have $t = 1$. Since there are 1024 locations, for every time stamp we need to record 1024x1024 probabilities. We record a total of 100 time stamps, taking 1024x1024x100x8 is approximately 0.8 GB. This explains why it is impractical to use much finer grids; the memory requirement for storing the transition probability distributions would exceed computer memory very quickly.

When we perform the calculations of Equation (7), and we need $P(l_k|l_{\pi(k)}, t_k)$ for a branch time t_k, first we need to determine time step Δt in terms of years. Instead of committing to a fixed value for Δt and in order to remain flexible with respect to the number of time steps, we normalize the duration of the ODE solver to span one unit of time and determine a scale factor for the tree to match. To this effect, a linear scaling is applied by multiplying t_k by w, where scale factor w is called wanderlust. Intuitively, the higher the wanderlust, the more eager the migrant is to move.

ODE Result Storage

Here, we describe which solutions of the ODE solver to store and how to use them. Since it would be impractical to store transition probabilities for all time intervals, we only store them for a limited number of steps during the run of the Euler algorithm. Then, to find a transition probability $P(l_k|l_{\pi(k)}, t_k)$ for a branch time t_k, if $t_k \times w$ is stored, we can look up the transition probability, but if $t_k \times w$ is not stored, we interpolate linearly between stored values. Figure S8 shows a typical set of transition probability tables, which shows that the curves are rather smooth, so interpolation can be expected to lead to very small approximation errors.

The number of the Euler method steps between storing samples was determined empirically as follows. First, an ODE solver for a grid block covering Scotland was performed where 10,000 equally distant transition probabilities were stored. The grid block was selected to generate a varied number of conditions (sea boundary, land boundary, close to grid boundary) so that a large variety of transition probability functions would be generated. Then, one by

one samples were eliminated such that the maximum distortion due to linear interpolation from its nearest remaining neighbors in the sample was minimized. The process was repeated to leave 100 samples from the original 10,000. Figure S9 shows the sample number on the y-axis for the remaining 100 samples. Also shown is the empirical derivative. Clearly, the derivative is smooth at the first 50 samples but after that shows some discontinuities. From this empirical derivative, we derived a smoothed derivative and by integration the smoothed sample numbers (shown in green and yellow respectively in Figure S9). Interestingly, the intervals are very closely spaced at the start and increase further apart towards the end. This coincides with the intuition we get from considering Figure S8, where it is clear that most of the non-linear behavior is at the start of the ODE solver. The further towards the end, the smoother the curves become and the better linear approximation will become over larger intervals.

Figure S10 shows a typical example of a number of transition probabilities with the sample numbers linearly displayed on the x-axis as opposed to Figure S8 where sample numbers are shown proportional to number of steps on the x-axis. The slight discontinuity at x=18 is caused by the step increase from 1 to 2. The evidently observable smoothness of the transition probability functions justifies the linear interpolation.

The number of time steps of the ODE solver is determined empirically such that all desired transition probabilities can be calculated from the stored transition probability matrices. For the $\rho_{l,l}$, $\rho_{l,w}$, $\rho_{w,l}$ and $\rho_{w,w}$ defined above, after 5 million time steps the average distance traveled was around 1300 km which suffices for this purpose. Since ODE solver runs are very time intensive, an implementation was developed to run on a graphics processing unit (GPU) using the CUDA library. Software was developed on a Linux system using an Intel i7 920 CPU and two GTX 295 NVidia cards containing two GPUs each. Utilizing a GPU gave an 8 fold increase in speed over a CPU. By utilizing two computers both with 4 GPUs, a total speed up over a CPU of about 56 was achieved. ODE solver run time for the high $\rho_{l,l}$ case was well over a day, and for the low $\rho_{l,l}$ case three days. The same calculations would take almost half a year on a CPU.

Figure S11 gives an impression of how the difference in land and water rate affects the boundary of the migration region; light blue means high concentration of probability mass, dark blue low probability mass and fuzzy background

color even less probability mass. The sharp boundary between foreground and background shows the boundary at low threshold level. The ODE solver starts at a grid block covering Scotland and flows out very fast over water. However, on land the flow has a lower speed but a much higher density.

The root distribution $P(l_{2n-1})$ is determined for each grid block by calculating the land area covered by the grid block. This is done by counting the number of land pixels in the grid block. So, our prior on the root is uniformly spread over the land pixels and the prior on the root originating in water is zero.

Relaxation of the landscape-based diffusion rates

Like the RRW analysis, the landscape-based analysis relaxed the diffusion rate from branch to branch. To calculate the transition probability of going to grid x_i from the parent of x_i in time t, that is $P^{\text{relaxed}}(x_i|x_{\pi(i)}, t)$ the relaxed clock averages the rate with a rate distribution $P(r)$, calculated as

$$\int_r P(x_i|x_{\pi(i)}, tr)P(r)dr \tag{11}$$

For the branch rate distribution, we use a distribution with mean of 1 as for the RRW model. To apply relaxed diffusion to the landscape based model, straightforward component-wise linear approximation as suggested by Drummond et al.(*46*) can be used. So, instead of the Equation (11), $P^{\text{relaxed}}(x_i|x_{\pi(i)}, t)$ is approximated by $\sum_{k=1}^{m} \frac{1}{m} P(x_i|x_{\pi(i)}, t\rho(k))$ where the rate distribution $P(r)$ is split in m equal probability intervals with means $\rho(k)$. Since we have to interpolate between sample points to obtain $P(x_i|x_{\pi(i)}, t\rho(k))$, an equally convenient approximation is $\sum_{\tau \in T} P(x_i|x_{\pi(i)}, \tau)P(\tau)P^{\text{relaxed}}(\tau/t|\sigma)$ where T is the set of sample times and $P(\tau)$ the probability the rate comes from interval τ to the next sample time. $P^{\text{relaxed}}(\tau/t|\sigma)$ is the probability according to the log-normal distribution of being in the interval τ/t and the next value of τ.

We followed the continuous model in using a log-normal distribution for relaxing rates of diffusion along branches. The log normal distribution requires the standard deviation σ to be specified. Furthermore, we need to specify the wanderlust parameter w. We found suitable values by performing a grid search over both σ and w under the assumption of preferring a smaller wanderlust over a higher by putting a $1/w$ prior on the wanderlust and likewise for σ. This gives an optimal posterior for $\sigma = 2$ which coincides with the $\sigma = 2.2$ found for the continuous model. The optimal value for w thus found was 0.0002. Root

densities were calculated by averaging over the root densities for the posterior sample of trees obtained by running an MCMC chain without geographic data as described by Pagel et al. (*60*).

5 Phylogeographic Hypothesis Testing

Data were fitted to the language evolution and spatial diffusion models using BEAST (*14*, *19*, *20*). BEAST XML files for the various analyses are available as supporting on-line material. The output from these analyses is a posterior distribution of trees with geographic locations at the root and internal nodes drawn in proportion to their posterior probability. This full probabilistic approach accounts for uncertainty in the phylogeny, age constraints, models of cognate replacement and spatial diffusion process.

Our phylogeographic model allows us to infer the location of ancestral langauge divergence events corresponding to the root and internal nodes of the Indo-European family tree. Since we model internal node locations as points in space, our posterior estimate for the location of divergence events can be interpreted as a composite of the range over which the ancestral language was spoken and stochastic uncertainty inherent in the model. The relative importance of each does not impact our ability to test whether one hypothesized origin range is more likely than another. Nevertheless, the picture of language expansion that emerges must be interpreted with the caveat that we can only trace the expansion of language divergence events (not the rapid expansion of a single language), and only between those languages that are in our sample. Nodes associated with branches not represented in our sample will not be captured. For example, our "Celtic" group contains only Insular Celtic languages and so cannot tell us about the timing or location of separation from Continental Celtic. With further methodological development, it may also be possible to incorporate indirect evidence about the range of a language prior to attestation, but this inevitably requires assumptions about how to map putative ancestral ranges onto the tree. In this sense, our approach is conservative, relying as it does on the location in time and space of well-attested modern and ancient languages. We note that our full analysis can be viewed as an extension of the contemporary language analysis, with the addition of information about lineage locations at ancestral nodes (in the form of the ancient languages). The fact

that both analyses support an Anatolian origin suggests that additional internal node information is unlikely to affect our conclusions.

We use these spatiotemporal reconstructions of the evolution of languages to test two theories of Indo-European origin: the 'Kurgan Steppe' and the 'Anatolian Farming' hypotheses. We formalize the alternative spatial hypotheses in the form of three areas (Figure 1). One area represents the Anatolian hypothesis covering the earliest Anatolian Neolithic sites (*11*). We represent the Kurgan Steppe hypothesis by two areas in the Pontic Steppe, an initial proposed origin (*5*) and a later refined (*7*) area.

The areas of the hypotheses are approximately $92,000$ km^2 for the Anatolian hypothesis, $421,000$ km^2 for the narrow Steppe hypothesis, and $1,760,000$ km^2 for the wider Steppe hypothesis. So, these areas show a bias toward the Steppe hypothesis; the area covered by the narrow Steppe hypothesis is more than four times larger than that of the Anatolian hypothesis. Likewise, the area covered by the wider Steppe hypothesis is more then 19 times larger than that of the Anatolian hypothesis.

We evaluate the support for these different hypotheses using Bayes factors calculated as

$$\frac{Posterior(H1)/Prior(H1)}{Posterior(H2)/Prior(H2)}$$

where the prior represents the probability that the origin will fall within a hypothesised area before observing the data. For the RRW model, these prior probabilities are obtained by a Monte Carlo simulation procedure that draws from the diffuse multivariate prior on the root (with zero means and covariance, and precisions of 0.001 for root latitude and longitude) and summarizes the frequency at which these draws are in the areas representing the different hypotheses. For the landscape based geographical model, which has a uniform prior over all pixels, these priors are calculated as the ratio of pixels within the hypothesis areas over the total number of pixels.

Kernel Density Estimation Method for Visualization

We followed Snyder (*61*) in using bivariate kernel density estimation to plot the highest posterior density contours in Figure 2 of the main text. Our density estimator used bivariate normal kernels with a diagonal bandwidth. We selected each dimension's bandwidth based on Silverman's "rule-of-thumb" plug-in value

(*62*) truncated to a maximum value of 0.5. This truncation controls for potential over-smoothing that may occur since the spacing between kernel locations is adjusted to match the range of each posterior sample. The dots shown in Figure 1b in the main text and Figure S5 are sampled from the land area in a grid location with probability proportional to the root distribution over grid locations.

6 Constrained Analysis

Bayesian inference is ideal for incorporating prior knowledge into the analysis. Phonological and morphological innovations provide an alternative source of comparative data to the lexical cognate data we analyze here. We repeated our analysis, constraining the tree topology to a topology based on phonological and morphological data (Figure S12) (*16*). The phonological and morphological dataset comprised 24 Indo-European languages, including the 20 ancient languages in our dataset. Table S5 shows how languages in the phonological and morphological dataset were mapped to our larger dataset to produce a set of topological constraints on the tree topology. Each clade was required to be monophyletic and the relationship between clades was used to constrain the internal nodes in the tree relating these clades. After adding the constraints and running the analysis under the RRW model we obtained a Bayes factor of 215.68 and 226.87 in favor of the Anatolian hypothesis when compared to the earlier Steppe and later refined Steppe hypothesis respectively. These results are comparable to the other less constrained models.

References

31. P. M. Lewis, *Ethnologue: Languages of the World. 16 edn,* (SIL International, 2009).

32. Q. D. Atkinson, R. D. Gray, in *Phylogenetic methods and the prehistory of languages* (MacDonald Institute for Archaeological Research, Cambridge, 2006), pp. 91–109.

33. S. J. Greenhill, R. D. Gray, in *Austronesian historical linguistics and culture history: a festschrift for Robert Blust* (Pacific Linguistics, Canberra, 2009), pp. 375–397.

34. J. Felsenstein, *Systematic Zoology* **27**, 27–33 (1978).

35. J. P. Huelsenbeck, F. Ronquist, R. Nielsen, J. P. Bollback, *Science* **294**, 2310–2314 (2001).

36. Q. D. Atkinson, G. K. Nicholls, D. Welch, R. D. Gray, *Transactions of the Philological Society* **103**, 193–219 (2005).

37. S. J. Greenhill, T. E. Currie, R. D. Gray, *Proceedings of the Royal Society B: Biological Sciences* **276**, 2299–2306 (2009).

38. J. Felsenstein, *Journal of Molecular Evolution* **17**, 368–376 (1981).

39. C. Tuffley, M. A. Steel, *Mathematical Biosciences* **147**, 63–91 (1998).

40. D. Penny, B. J. Mccomish, M. A. Charleston, M. D. Hendy, *Journal of Molecular Evolution* **53**, 711–723 (2001).

41. G. Nicholls, R. Gray, *J. Roy. Stat. Soc. B* **70**, 545–566 (2008).

42. M. Hasegawa, H. Kishino, T. Yano, *Journal of Molecular Evolution* **22**, 160–174 (1985).

43. R. R. Gray *et al.*, *Mol Biol Evol* **28**, 1593–603 (2011).

44. A Rambaut, *Bioinformatics* **16**, 395–9 (2000).

45. B. Shapiro *et al.*, *Mol Biol Evol* **28**, 879–87 (2011).

46. A. J. Drummond, S. Y. W. Ho, M. J. Phillips, A. Rambaut, *PLoS Biol* **4**, e88 (Mar. 2006).

47. A. J. Drummond, A. Rambaut, B. Shapiro, O. G. Pybus, *Molecular Biology and Evolution* **22**, 1185–1192, ISSN: 0737-4038 (2005).

48. M. A. Suchard, C. M. R. Kitchen, J. S. Sinsheimer, R. E. Weiss, *Syst Biol* **52**, 649–64 (2003).

49. B. D. Redelings, M. A. Suchard, *Syst Biol* **54**, 401–18 (2005).

50. S. Nelson-Sathi *et al.*, *Proceedings of the Royal Society B: Biological Sciences* **278**, 1794–803, ISSN: 1471-2954 (June 2011).

51. U. Tadmor, M. Haspelmath, B. Taylor, *Diachronica* **2**, 226–246 (2010).

52. Q. D. Atkinson, A. Meade, C. Venditti, S. J. Greenhill, M. Pagel, *Science* **319**, 588, ISSN: 1095-9203 (Mar. 2008).

53. M. Pagel, Q. D. Atkinson, A. Meade, *Nature* **449**, 717–720, ISSN: 1476-4687 (2007).

54. S. J. Greenhill, Q. D. Atkinson, A. Meade, R. D. Gray, *Proceedings of the Royal Society B: Biological Sciences* **277**, 2443–50, ISSN: 1471-2954 (2010).

55. B. W. Fortson, *Indo-European Language and Culture: an introduction* (Wiley-Blackwell, Malden, 2010).

56. F. Bartumeus, M. G. E. D. Luz, *Ecology* **86**, 3078–3087 (2005).

57. P. Turchin, *Quantitative Analysis of Movement: measuring and modeling population redistribution in plants and animals* (Sinauer Associates, Sunderland, MA, 1998).

58. H.-i. Wu, B.-L. Li, T. A. Springer, W. H. Neill, *Ecological Modelling* **132**, 115–124 (2000).

59. A. Gelfand, S. Hills, A. Racine-Poon, A. Smith, *Journal of the American Statistical Association* **85**, 972–985 (1990).

60. M. Pagel, A. Meade, D. Barker, *Systematic Biology* **53**, 673–684, ISSN: 10635157 (2004).

61. W. V. Snyder, *ACM Trans. Math. Softw.* **4**, 290–294, ISSN: 0098-3500 (3 1978).

62. B. W. Silverman, *Density Estimation* (Chapman and Hall, London, 1986).

7 Supplementary Figures

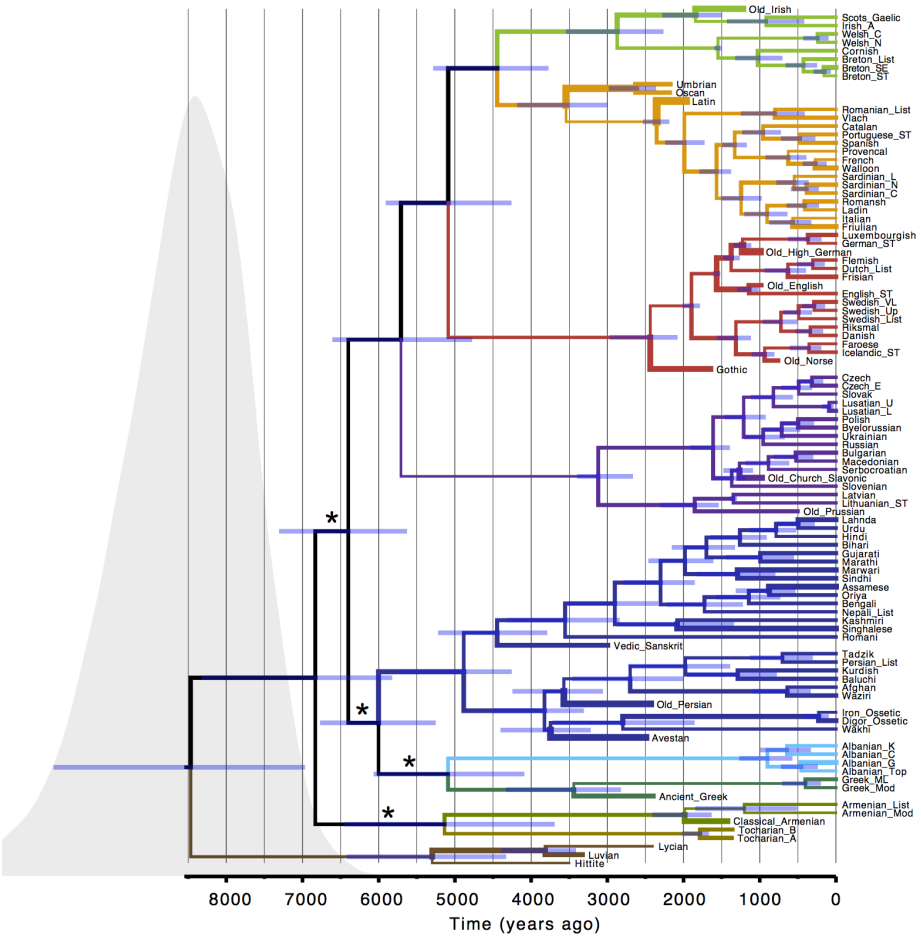

Figure S 1: **Maximum clade credibility tree depicting the variation in rates of cognate replacement along branches for the 103 Indo-European languages in our sample.** Maximum clade credibility tree for the 103 Indo-European languages in our sample. Branches are colored to indicate the main sub-families following the scheme used in Figure 2. The thickness of the branches reflects the relative rate of cognate replacement along branches; these rates vary on average within 47% of the mean rate. Actual values are recorded in Supplementary Data File "1219669IndoEuropean_2MCCtrees_annotated.tre". The gray density represents the marginal posterior probability estimate for the root age. Blue bars represent the 95% HPD intervals for the node ages. All major nodes were supported by a posterior probability > 0.95 except those indicated with a '*'.

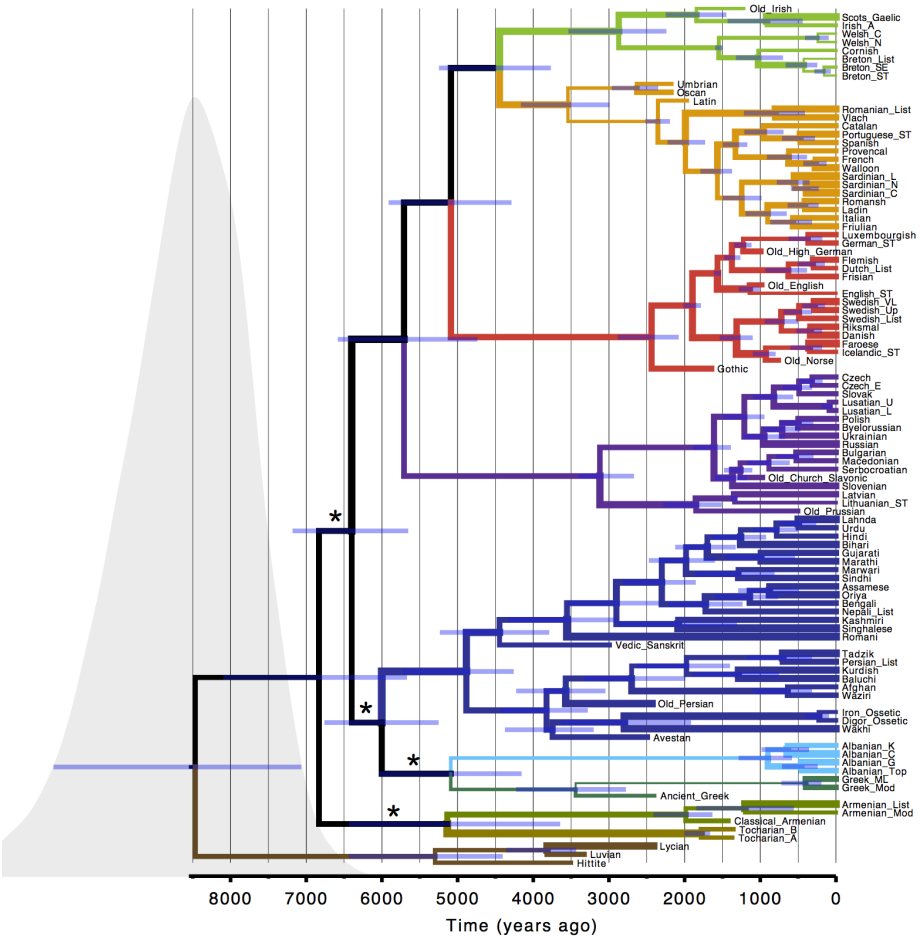

Figure S2: **Maximum clade credibility tree depicting the variation in rates of spatial diffusion along branches for the 103 Indo-European languages in our sample.** Maximum clade credibility tree for the 103 Indo-European languages in our sample. Branches are colored to indicate the main sub-families following the scheme used in Figure 2. The thickness of the branches reflects the relative rate of spatial diffusion along branches; these rates vary on average within 435% of the mean rate (mean rate=0.48 km/yr, 95%HPDs=0.42-0.55). Actual values are recorded in Supplementary Data File "1219669IndoEuropean_2MCCtrees_annotated.tre". The gray density represents the marginal posterior probability estimate for the root age. Blue bars represent the 95% HPD intervals for the node ages. All major nodes were supported by a posterior probability > 0.95 except those indicated with a '*'.

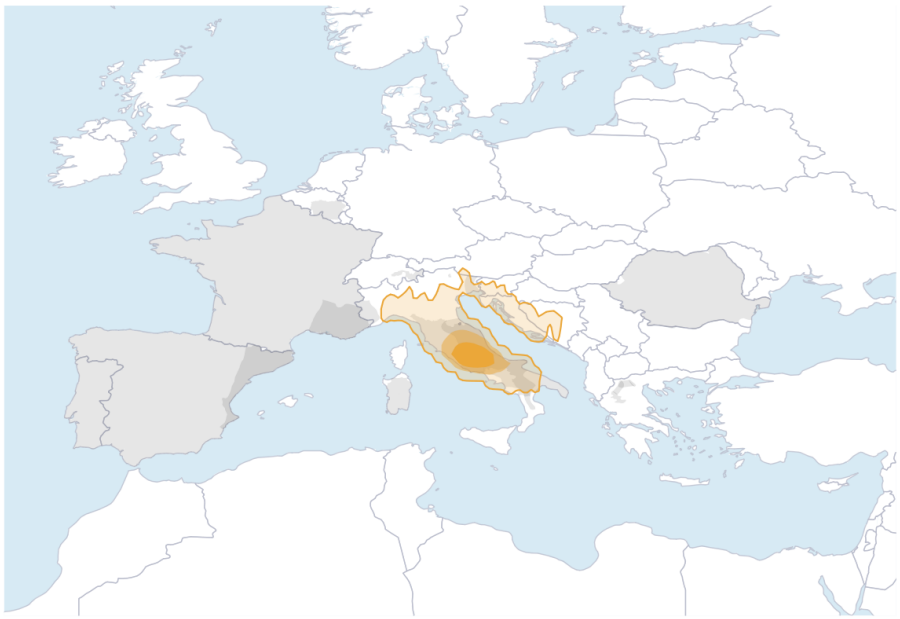

Figure S3: **Inferred location of the most recent common ancestor of the Romance group of languages** (excluding Latin). Contours represent the 95% (largest), 75% and 50% HPD regions, based on kernel density estimates.

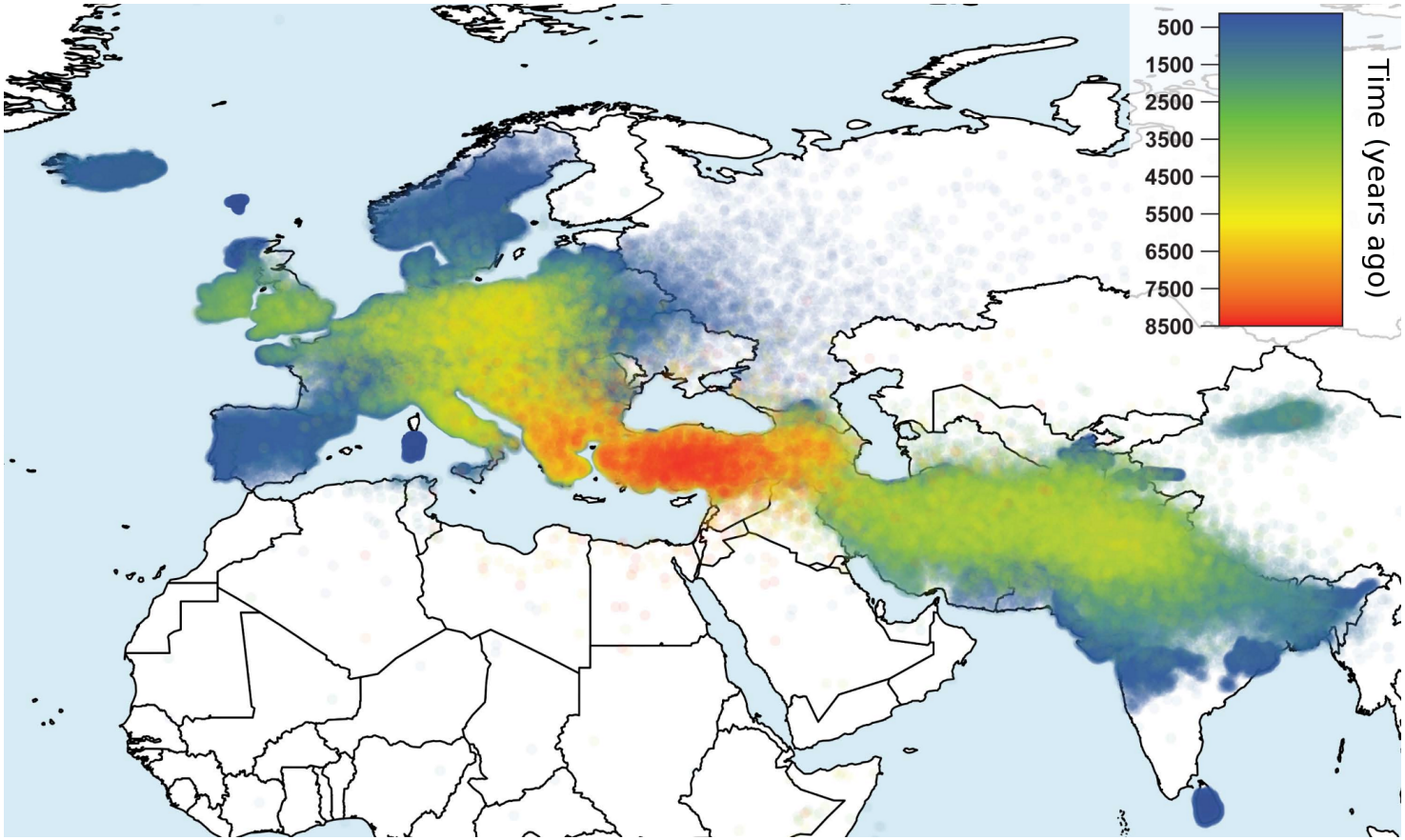

Figure S4: **Spatial reconstruction of the expansion of the sampled Indo-European languages in continuous space.** The posterior distribution of node location estimates through time are plotted as opaque points with a color that indicates their corresponding age estimate. Older nodes are shown on the foreground to clearly depict the temporal diffusion pattern. This figure needs to be interpreted with the caveat that we can only represent the geographic extent corresponding to langauge divergence events, and only between those languages that are in our sample. The rapid expansion of a single langauge and nodes associated with branches not represented in our sample will not be reflected in this figure. For example, the lack of Continental Celtic variants in our sample means we miss the Celtic incursion into Iberia and instead infer a later arrival into the Iberian peninsular associated with the break-up of the Romance languages (and not the initial rapid expansion of Latin). The chronology represented here therefore offers a minimum age for expansion into a given area.

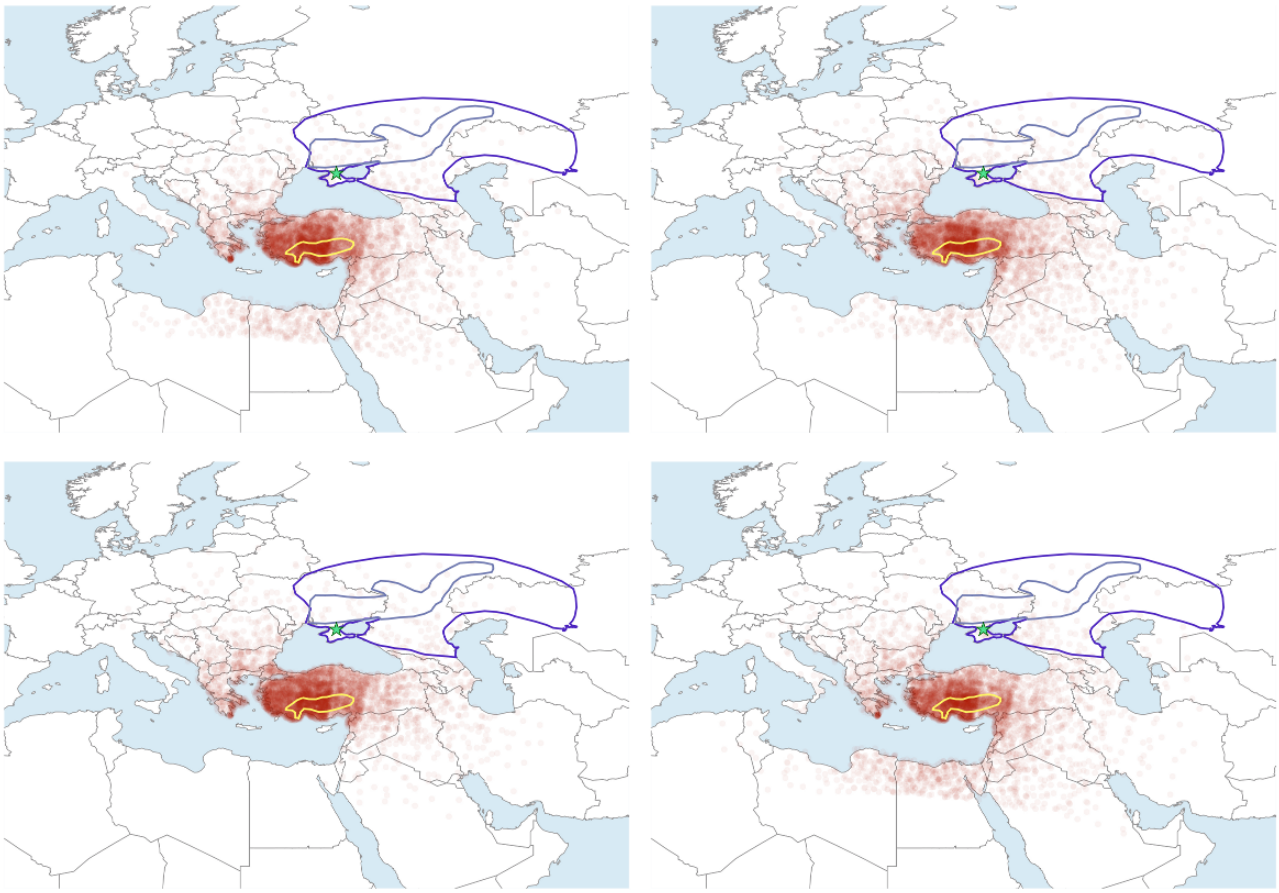

Figure S5: **Inferred geographic origin of the Indo-European language family under four landscape based models.** Sampled locations are plotted in translucent red such that darker areas correspond to increased probability mass. Diffusion model at top left, 10 times less likely into water top right, 100 times less likely into water model bottom left, and the sailor model bottom right. The green star represents the centroid location of the languages.

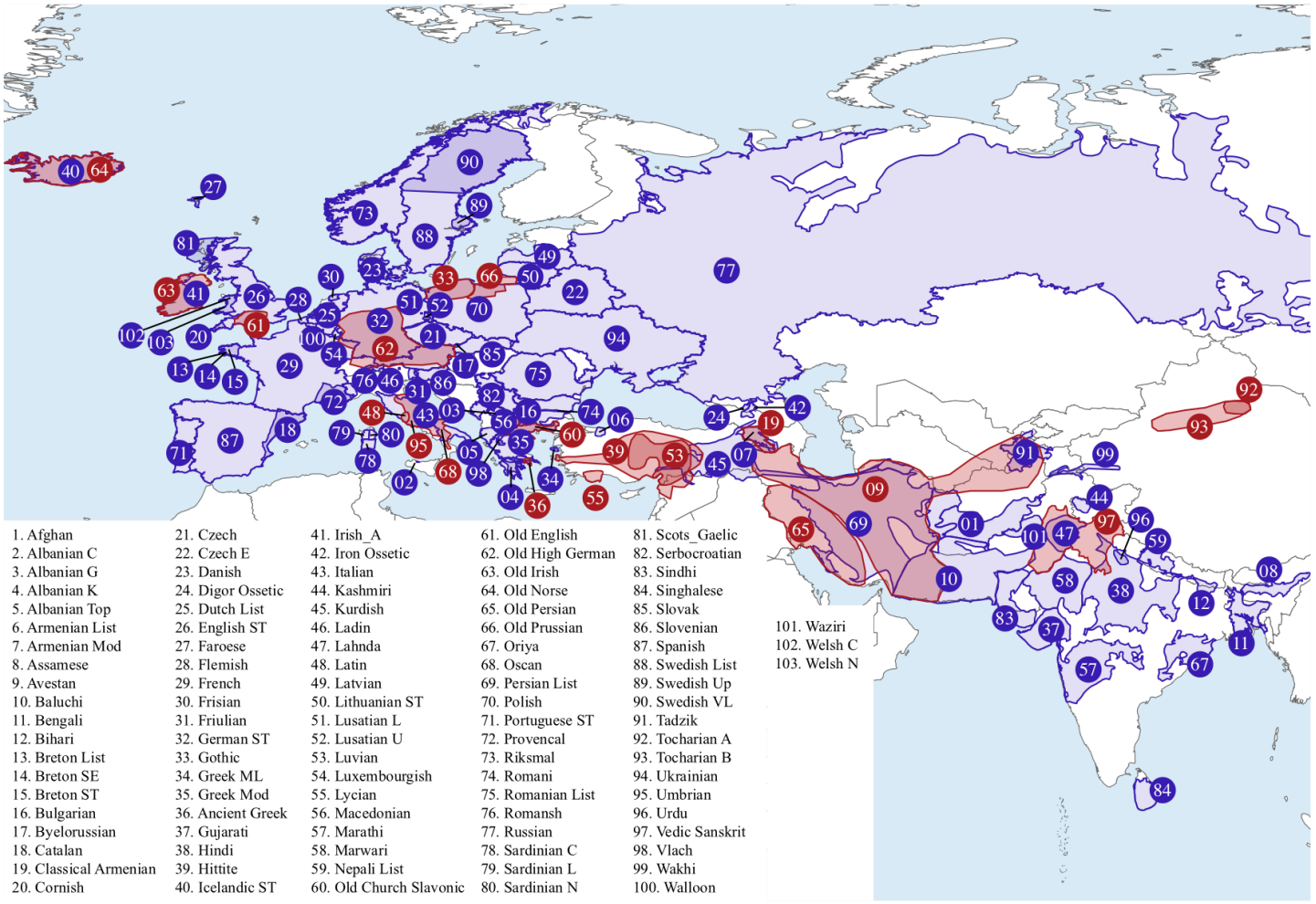

Figure S6: **Locations of the 103 languages in our sample.** Colored polygons represent the geographic area assigned to each language based on Ethnologue(*31*). Red areas indicate ancient languages and blue areas indicate modern languages.

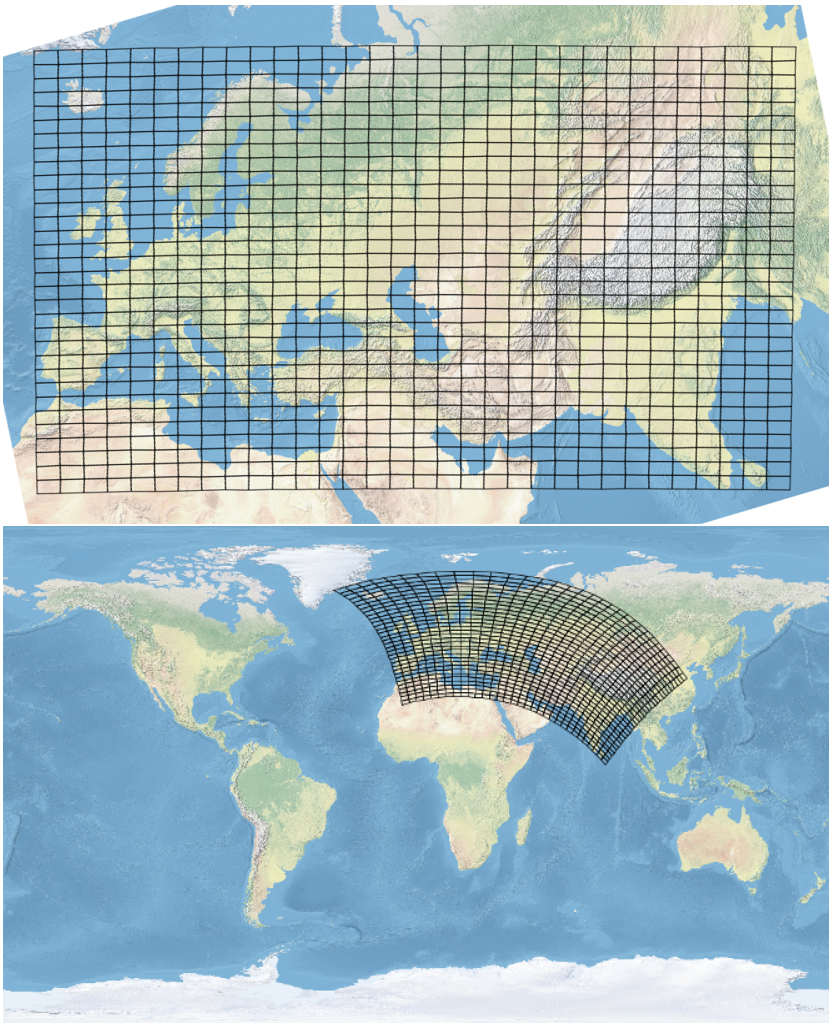

Figure S7: 32x32 grid locations. In rotated Mercator projection (top) and drawn on a standard Mercator projection (bottom).

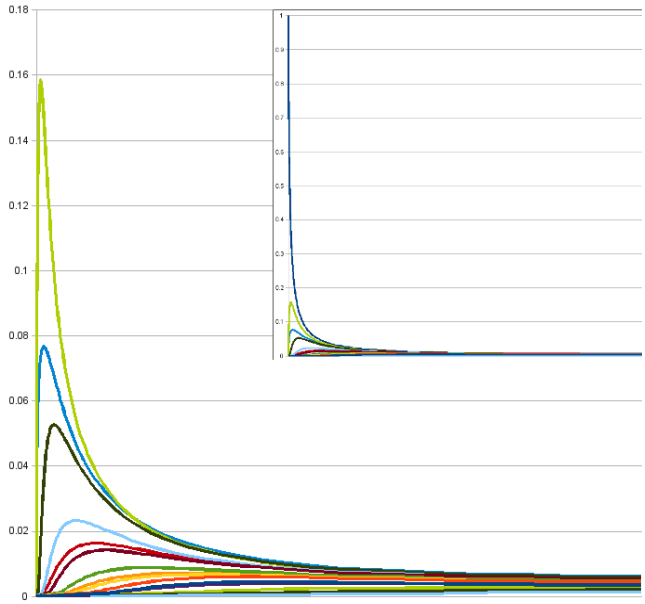

Figure S 8: Typical set of transition probability curves. X-axis shows time from $t = 0$ up to $t = 1$ and Y-axis shows transition probabilities and their interpolations. Inset shows same but with transition from source location to itself (blue line, starting at 1 when t=0) as well.

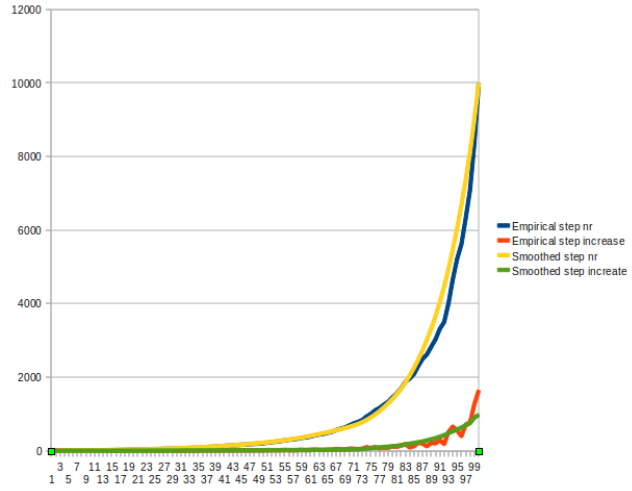

Figure S9: Sample step numbers; Blue line is empirically determined from 10.000 samples with equal intervals by minimizing distortion. X-axis the number of the sample step, Y-axis the size of the step.

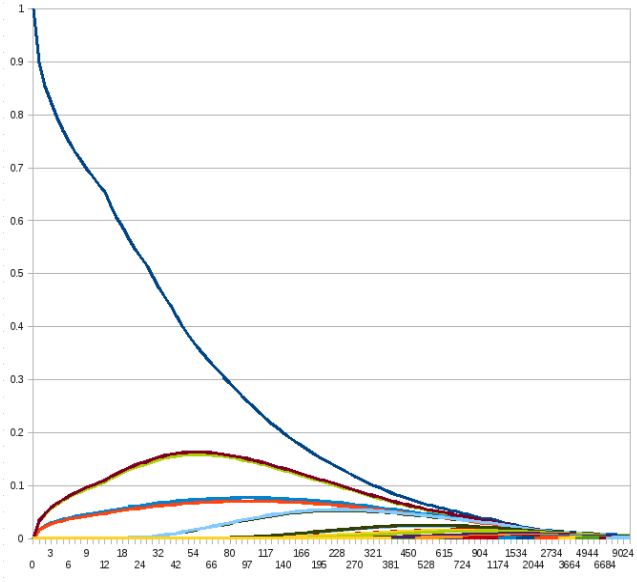

Figure S10: Typical ODE solver based transition probabilities with X-axis the sample numbers and Y-axis the transition probability.

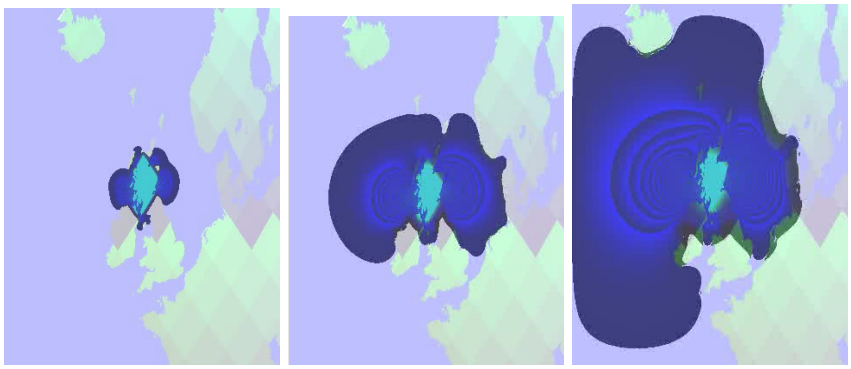

Figure S11: Three stages in an ODE solver run starting in the grid block covering the north of Scotland in standard Mercator projection.

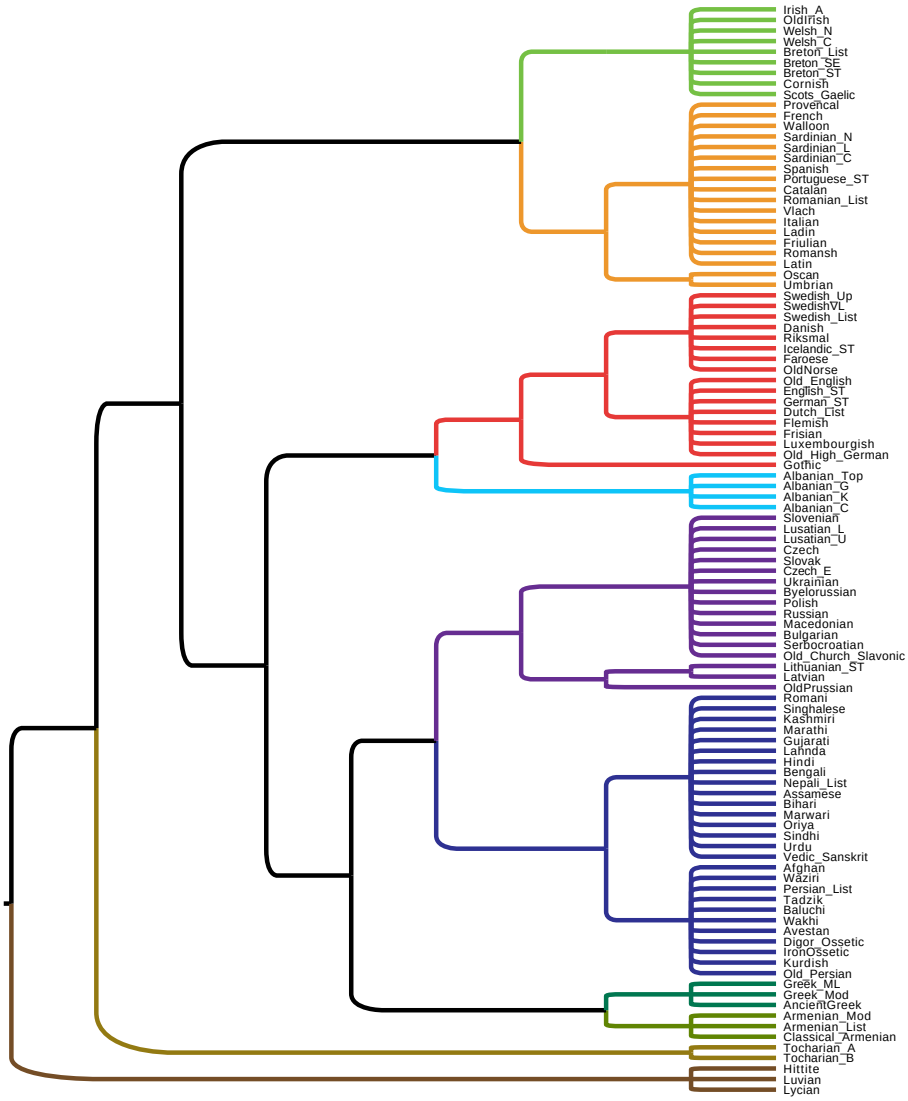

Figure S12: Tree diagram showing the topology used in our 'constrained' analysis to restrict the pattern of Indo-European diversification to that advocated in Ringe et al. (*16*) using data that included weighted phonological and morphological language characteristics.

8 Supplementary Tables

Table S 1: **Cognate source by language.** Numbers in square brackets are reference numbers of sources listed at the and of this table, and numbers in brackets are the number of cognates obtained from the source indicated by the reference number.

Language (iso code)	Lexeme source (count)	Cognate judgement source (count)
Afghan (prs)	[1] (217), [2] (1)	[1] (159), [3] (17), [4] (5), [5] (4), [6] (1), [7] (1), [8] (1)
Albanian C	[1] (191), [3] (1), [7] (1), [9] (1)	[1] (141), [3] (15), [7] (4), [10] (3), [6] (2), [5] (2), [2] (1), [11] (1)
Albanian G (aln)	[1] (210), [2] (1), [7] (1), [11] (1), [3] (1), [12] (1)	[1] (171), [3] (16), [7] (5), [10] (2), [8] (2), [11] (2), [6] (1), [2] (1), [5] (1)
Albanian K	[1] (202), [7] (3)	[1] (147), [3] (18), [7] (6), [10] (3), [6] (2), [2] (1), [11] (1), [5] (1)
Albanian Top	[1] (202), [13] (1), [11] (1)	[1] (171), [3] (14), [7] (4), [10] (3), [6] (2), [2] (1), [11] (1), [13] (1)
Ancient Greek (grc)	[14] (374), [15] (137), [6] (19), [11] (8), [7] (3), [16] (2), [9] (2), [17] (2), [18] (1)	[3] (90), [11] (74), [5] (44), [6] (20), [7] (12), [17] (7), [8] (5), [1] (3), [9] (2), [13] (2), [19] (2), [20] (2), [14] (1), [18] (1), [21] (1), [4] (1)
Armenian List	[1] (193), [19] (3), [9] (2), [22] (1), [7] (1), [23] (1), [4] (1), [17] (1)	[1] (127), [5] (19), [19] (10), [3] (10), [7] (2), [11] (2), [23] (1), [4] (1), [17] (1)
Armenian Mod (hye)	[24] (218), [1] (216), [19] (13), [7] (6), [11] (3), [20] (2), [6] (1), [8] (1)	[1] (126), [5] (41), [19] (18), [3] (12), [7] (7), [11] (6), [25] (4), [6] (2), [8] (2), [9] (1), [4] (1)
Assamese (asm)	[26] (209), [27] (125), [28] (8), [29] (5)	[3] (106), [5] (50), [29] (27), [11] (16), [1] (2), [8] (1), [26] (1), [17] (1)
Avestan (ave)	[20] (202), [11] (1)	[5] (89), [20] (35), [11] (11), [4] (1)
Baluchi (bgp)	[1] (209), [3] (1), [11] (1)	[1] (126), [3] (13), [4] (4), [5] (4), [11] (3), [8] (1)
Bengali (ben)	[1] (220), [30] (40), [29] (6)	[1] (142), [3] (32), [29] (15), [5] (7), [11] (4), [6] (1), [26] (1)
Bihari (mai)	[31] (204), [32] (118), [33] (19), [34] (9), [29] (1), [27] (1)	[3] (124), [5] (65), [29] (19), [11] (12), [17] (1)

Continued on next page

Table S1 continued – **Cognate source by language.**

Language [iso code]	Lexeme source [count]	Cognate judgement source [count]
Breton List	[1] (255), [3] (2), [11] (2)	[1] (174), [11] (21), [3] (17), [23] (7), [7] (5), [5] (2)
Breton ST (bre)	[1] (215), [35] (178), [36] (38), [11] (5), [7] (1), [3] (1)	[1] (176), [11] (19), [3] (15), [23] (6), [7] (2), [5] (1)
Breton SE	[1] (204), [3] (1)	[1] (167), [11] (12), [3] (11), [23] (5), [7] (3), [5] (1)
Bulgarian (bul)	[1] (199), [37] (194)	[1] (159), [3] (13), [11] (9), [37] (5), [25] (4), [38] (4), [8] (1), [5] (1)
Byelorussian (bel)	[1] (208), [3] (2), [29] (1), [38] (1)	[1] (164), [3] (13), [11] (12), [38] (6), [7] (2), [8] (1), [5] (1)
Catalan (cat)	[1] (269), [39] (181), [40] (77), [23] (1)	[1] (161), [3] (39), [11] (17), [25] (15), [40] (14), [7] (2), [41] (2), [5] (2), [42] (1), [13] (1)
Classical Armenian (xcl)	[20] (238)	[5] (139), [20] (41), [19] (1)
Cornish (cor)	[43] (220), [44] (162), [45] (86), [46] (57), [47] (14), [23] (3), [11] (1)	[3] (112), [5] (54), [11] (46), [23] (17), [7] (10), [1] (5), [17] (4), [8] (2), [9] (1)
Czech (ces)	[1] (218), [37] (208), [11] (6), [38] (2), [29] (1)	[1] (184), [3] (21), [37] (14), [11] (10), [38] (5), [25] (1), [7] (1), [8] (1)
Czech E (ces)	[1] (211), [3] (1), [11] (1)	[1] (176), [3] (19), [11] (5), [38] (4), [5] (3), [7] (2), [8] (1)
Danish (dan)	[48] (205), [1] (201), [11] (4)	[1] (177), [11] (14), [3] (10), [7] (2), [42] (1), [5] (1)
Digor Ossetic (oss)	[49] (201)	[5] (176), [4] (2)
Dutch List (nld)	[50] (230), [1] (225), [11] (6), [5] (1)	[1] (178), [11] (22), [3] (16), [7] (3), [51] (2), [52] (2), [5] (2), [29] (1), [42] (1)
English (eng)	[53] (203), [1] (200), [11] (6)	[1] (152), [11] (24), [3] (12), [8] (3), [7] (2), [5] (2), [54] (2)
Faroese (fao)	[1] (224), [55] (2)	[1] (175), [3] (21), [11] (16), [7] (5), [8] (1), [42] (1)
Flemish (vls)	[1] (216), [3] (1)	[1] (177), [3] (18), [11] (14), [7] (3), [42] (2), [51] (1), [8] (1), [4] (1), [5] (1), [52] (1)
French (fra)	[56] (212), [1] (201), [11] (10), [57] (1)	[1] (166), [11] (25), [3] (18), [41] (5), [7] (1), [5] (1)

Continued on next page

Table S1 continued – **Cognate source by language.**

Language [iso code]	Lexeme source [count]	Cognate judgement source [count]
Frisian (frs)	[1] (216), [58] (152), [55] (34), [42] (2), [7] (1)	[1] (150), [3] (22), [11] (11), [55] (4), [7] (3), [42] (2), [25] (1), [29] (1), [8] (1)
Friulian (fur)	[59] (228)	[5] (193), [11] (2), [1] (1), [54] (1)
German (deu)	[60] (216), [1] (208), [11] (7), [3] (3), [61] (1)	[1] (171), [11] (25), [3] (11), [7] (8), [42] (1), [23] (1), [5] (1)
Gothic (got)	[20] (186), [62] (33), [63] (13), [64] (10), [7] (4), [11] (3), [1] (2), [65] (2), [22] (1), [9] (1), [52] (1)	[3] (84), [11] (57), [5] (40), [7] (8), [20] (4), [8] (3), [17] (3), [66] (1), [38] (1), [9] (1), [42] (1)
Greek Ml	[1] (199), [6] (1)	[1] (165), [3] (16), [11] (12), [6] (3), [7] (3)
Greek Mod (ell)	[1] (217), [67] (209), [11] (4), [6] (2), [17] (1)	[1] (159), [3] (21), [11] (17), [5] (7), [25] (5), [7] (5), [6] (4), [8] (1), [9] (1)
Gujarati (guj)	[1] (213), [68] (11), [3] (2), [29] (2)	[1] (143), [3] (28), [29] (13), [5] (5), [11] (3), [25] (1), [7] (1)
Hindi (hin)	[1] (222), [30] (79), [29] (3), [69] (2), [3] (1), [17] (1)	[1] (166), [3] (27), [29] (13), [5] (7), [11] (6), [30] (3), [40] (2), [25] (1), [7] (1), [17] (1)
Hittite (hit)	[20] (188), [2] (16), [70] (11), [66] (5), [9] (3), [11] (2), [17] (2), [7] (1)	[2] (25), [3] (22), [66] (18), [11] (14), [20] (13), [17] (9), [8] (5), [9] (4), [5] (2), [7] (1), [18] (1), [23] (1), [13] (1), [19] (1)
Icelandic ST (isl)	[1] (205)	[1] (166), [11] (18), [3] (14), [7] (3), [5] (2), [8] (1), [42] (1)
Irish A	[1] (215), [23] (3), [11] (3), [3] (1), [6] (1)	[1] (148), [11] (26), [3] (19), [23] (5), [5] (2)
Iron Ossetic (oss)	[49] (200)	[5] (188), [4] (2)
Italian (ita)	[71] (228), [1] (220), [11] (11)	[1] (165), [11] (37), [3] (20), [5] (6), [41] (3), [7] (1)
Kashmiri (kas)	[1] (240), [29] (4), [3] (2), [18] (1)	[1] (120), [3] (32), [29] (12), [5] (9), [11] (3), [6] (1)
Kurdish (kmr)	[72] (136)	[3] (40), [5] (34), [11] (7), [8] (2), [10] (1)
Ladin (lld)	[1] (221), [54] (1), [7] (1)	[1] (154), [3] (35), [11] (12), [5] (12), [41] (3), [6] (1)

Continued on next page

Table S1 continued – **Cognate source by language.**

Language [iso code]	Lexeme source [count]	Cognate judgement source [count]
Lahnda (pnb)	[1] (199), [29] (3)	[1] (155), [3] (16), [29] (9), [5] (5), [11] (2), [7] (1)
Latin	[20] (207), [73] (199), [74] (196), [75] (94), [13] (18), [11] (10), [17] (3), [7] (1)	[11] (84), [3] (73), [5] (41), [17] (8), [13] (5), [20] (5), [8] (4), [1] (3), [7] (3), [41] (2), [9] (2), [2] (1), [18] (1), [66] (1), [76] (1), [54] (1), [75] (1)
Latvian (lav)	[20] (231), [1] (215), [11] (6), [77] (2), [3] (2), [7] (1), [38] (1), [17] (1)	[1] (122), [11] (122), [3] (112), [5] (40), [7] (19), [8] (5), [38] (5), [66] (3), [20] (3), [29] (2), [17] (2), [51] (1)
Lithuanian ST (lit)	[1] (218), [78] (79), [11] (8), [17] (4), [3] (1), [23] (1), [13] (1)	[1] (165), [11] (16), [3] (11), [7] (10), [17] (5), [20] (5), [38] (2), [2] (1), [66] (1), [8] (1), [13] (1), [5] (1)
Lusatian L (dsb)	[1] (192)	[1] (172), [3] (12), [11] (3), [7] (1), [8] (1), [38] (1)
Lusatian U (hsb)	[1] (192)	[1] (176), [3] (12), [11] (2), [38] (2), [7] (1), [8] (1)
Luvian	[20] (107), [11] (1)	[20] (36), [5] (1)
Luxembourgish (ltz)		[3] (95), [5] (59), [11] (43), [7] (3), [42] (2), [8] (1), [38] (1), [23] (1), [52] (1)
Lycian	[20] (40)	[20] (18)
Macedonian (mkd)	[1] (233), [21] (66)	[1] (167), [3] (26), [21] (14), [11] (10), [38] (8), [7] (1), [8] (1), [5] (1)
Marathi (mar)	[79] (239), [1] (220), [29] (2), [3] (1)	[1] (138), [5] (45), [3] (20), [29] (11), [25] (9), [7] (2), [11] (2), [6] (1)
Marwari (rwr)	[80] (128), [81] (69), [82] (51), [83] (38), [84] (2)	[3] (97), [5] (56), [11] (8), [29] (6), [1] (1), [17] (1)
Nepali (nep)	[1] (257), [85] (16), [29] (13), [3] (1), [5] (1)	[1] (167), [3] (38), [29] (35), [5] (8), [25] (4), [6] (1), [7] (1), [11] (1), [4] (1)
Old Church Slavonic (chu)	[20] (222), [21] (185), [86] (7), [87] (2), [11] (2), [1] (1), [7] (1), [88] (1)	[3] (96), [11] (68), [5] (54), [7] (6), [38] (6), [8] (4), [17] (3), [1] (1), [51] (1), [18] (1), [9] (1)
Old English (ang)	[20] (242), [11] (3), [89] (1), [23] (1), [13] (1)	[3] (98), [11] (75), [5] (52), [7] (5), [8] (4), [1] (1), [51] (1), [52] (1), [13] (1), [54] (1), [20] (1)
Old High German (goh)	[20] (255)	[5] (217), [20] (15)
Old Irish (sga)	[20] (241)	[5] (153), [20] (14), [11] (2)

Continued on next page

Table S1 continued – **Cognate source by language.**

Language [iso code]	Lexeme source [count]	Cognate judgement source [count]
Old Norse (non)	[22] (252), [20] (243), [11] (3), [1] (1), [7] (1)	[3] (116), [11] (86), [5] (55), [7] (9), [20] (7), [17] (3), [8] (2), [1] (1), [51] (1), [42] (1), [52] (1), [13] (1)
Old Persian	[20] (81), [11] (1)	[20] (32), [5] (13), [11] (3), [4] (1)
Old Prussian (prg)	[20] (158), [90] (27), [7] (1), [38] (1), [91] (1), [11] (1), [9] (1), [17] (1)	[3] (45), [5] (34), [11] (30), [7] (18), [17] (4), [20] (4), [38] (3), [9] (3), [2] (1)
Oriya (ori)	[26] (221), [92] (201), [29] (3), [93] (2)	[3] (113), [5] (68), [29] (37), [11] (15), [1] (2), [6] (1), [26] (1), [17] (1)
Oscan	[20] (53)	[20] (28)
Persian (pes)	[1] (202), [30] (54), [17] (4), [11] (2), [7] (1), [13] (1)	[1] (152), [3] (17), [5] (10), [4] (5), [11] (4), [8] (2), [40] (2), [17] (2), [29] (1), [7] (1), [13] (1)
Polish (pol)	[94] (211), [1] (200), [11] (6), [17] (1)	[1] (174), [11] (13), [25] (12), [3] (12), [38] (4), [7] (1), [8] (1), [5] (1), [17] (1)
Portuguese ST (por)	[1] (242), [95] (223)	[1] (177), [3] (38), [25] (17), [11] (13), [41] (3), [7] (2), [5] (1)
Provencal (prv)	[1] (251), [40] (35), [2] (1), [41] (1), [11] (1), [3] (1)	[1] (171), [3] (38), [11] (22), [40] (8), [5] (7), [41] (5), [7] (1)
Riksmal (nob)	[1] (200)	[1] (172), [3] (12), [11] (12), [42] (1), [7] (1)
Romani (rmy)	[1] (182), [29] (5)	[1] (80), [3] (27), [29] (11), [5] (5), [11] (4), [7] (2), [8] (1)
Romanian List (ron)	[1] (229), [96] (125), [11] (14), [76] (2), [2] (1)	[1] (148), [11] (40), [3] (23), [25] (5), [76] (4), [5] (3), [41] (2), [13] (2)
Romansh (roh)	[40] (219)	[5] (197), [60] (1), [1] (1)
Russian (rus)	[5] (210), [1] (201), [97] (200), [11] (9), [38] (1), [17] (1)	[1] (166), [11] (16), [3] (12), [38] (6), [7] (3), [8] (1), [5] (1)
Sardinian C (sro)	[1] (199)	[1] (152), [3] (22), [11] (11), [41] (3), [7] (1), [13] (1), [5] (1)
Sardinian L	[1] (200)	[1] (156), [3] (17), [11] (15), [41] (3), [5] (2), [7] (1), [13] (1)
Sardinian N	[1] (198)	[1] (143), [3] (25), [11] (11), [41] (3), [7] (1), [13] (1)

Continued on next page

Table S1 continued – **Cognate source by language.**

Language [iso code]	Lexeme source [count]	Cognate judgement source [count]
Scots Gaelic (gla)	[98] (231), [36] (75), [11] (2)	[3] (79), [5] (52), [11] (49), [23] (11), [7] (4), [8] (2), [99] (1), [36] (1), [4] (1)
Serbocroatian (bos)	[1] (200), [11] (4)	[1] (167), [3] (15), [11] (12), [38] (4), [7] (3), [8] (1), [5] (1)
Sindhi (snd)	[100] (160), [101] (49), [34] (48), [29] (8), [7] (1)	[3] (90), [5] (52), [29] (19), [11] (12), [1] (2), [17] (2), [6] (1), [7] (1)
Singhalese (sin)	[1] (216), [102] (21), [29] (8)	[1] (81), [3] (15), [29] (12), [25] (3), [5] (2), [11] (1)
Slovak (slk)	[1] (222)	[1] (183), [3] (20), [11] (10), [38] (4), [5] (2), [7] (1), [8] (1)
Slovenian (slv)	[1] (203)	[1] (160), [3] (13), [11] (2), [7] (1), [8] (1), [38] (1)
Spanish (spa)	[103] (219), [1] (215), [11] (7), [23] (1), [13] (1)	[1] (165), [3] (31), [11] (25), [103] (12), [41] (3), [7] (2)
Swedish (swe)	[1] (237), [104] (160), [11] (5), [55] (4), [3] (2)	[1] (178), [3] (28), [11] (25), [55] (3), [5] (3), [25] (1), [7] (1), [42] (1)
Swedish Up	[1] (225), [29] (2)	[1] (178), [3] (22), [11] (17), [5] (3), [7] (2), [8] (1), [42] (1)
Swedish Vl	[1] (216), [7] (1)	[1] (169), [3] (21), [11] (18), [5] (3)
Tadzik (tgk)	[1] (248), [30] (78), [3] (1), [4] (1)	[1] (173), [3] (19), [5] (13), [4] (4), [30] (3), [11] (2), [25] (1), [8] (1)
Tocharian A (xto)	[20] (174), [105] (10), [11] (3), [9] (2), [4] (1), [17] (1)	[3] (51), [105] (28), [11] (19), [5] (18), [17] (9), [20] (6), [8] (3), [2] (2), [9] (2), [13] (2), [99] (1), [66] (1), [106] (1), [23] (1), [4] (1)
Tocharian B (txb)	[20] (208), [105] (7), [11] (1), [23] (1), [4] (1), [13] (1), [17] (1)	[3] (55), [105] (31), [5] (21), [11] (18), [20] (10), [17] (9), [8] (4), [2] (2), [9] (2), [23] (2), [13] (2), [99] (1), [66] (1), [106] (1), [4] (1)
Ukrainian (ukr)	[1] (260), [3] (1), [21] (1), [38] (1)	[1] (175), [3] (30), [38] (9), [11] (8), [5] (3), [7] (2), [107] (1), [8] (1)
Umbrian	[20] (65)	[20] (37), [13] (1)
Urdu (urd)	[108] (213), [1] (6), [92] (1)	[3] (113), [5] (72), [11] (16), [29] (9), [7] (1)

Continued on next page

Table S1 continued – **Cognate source by language.**

Language [iso code]	Lexeme source [count]	Cognate judgement source [count]
Vedic Sanskrit (san)	[20] (229), [109] (40), [110] (24), [11] (17), [111] (8), [7] (5), [29] (5), [17] (4), [9] (3), [1] (1), [112] (1)	[11] (84), [3] (69), [5] (27), [29] (11), [8] (10), [7] (9), [20] (6), [9] (5), [17] (5), [66] (3), [105] (3), [112] (3), [4] (3), [1] (1), [51] (1), [18] (1), [13] (1)
Vlach (rup)	[1] (186)	[1] (125), [3] (17), [11] (11), [41] (3), [5] (2), [76] (1), [13] (1)
Wakhi (wbl)	[1] (221), [4] (2), [3] (1)	[1] (110), [3] (26), [4] (5), [11] (2), [23] (1), [5] (1)
Walloon (wln)	[1] (212), [11] (1)	[1] (160), [3] (20), [11] (12), [41] (3), [5] (2), [7] (1)
Waziri (pst)	[1] (222)	[1] (156), [3] (12), [5] (7), [4] (3), [6] (1), [8] (1), [105] (1), [11] (1)
Welsh C	[1] (204), [23] (1), [11] (1)	[1] (166), [11] (14), [3] (8), [23] (7), [7] (1), [9] (1), [5] (1)
Welsh N (cym)	[1] (219), [36] (28), [11] (6), [23] (2), [6] (1), [7] (1), [17] (1), [113] (1), [20] (1)	[1] (169), [11] (22), [3] (9), [23] (7), [7] (3), [20] (3), [25] (1), [5] (1), [17] (1)

References

[1] J. Dyen, Isidore Kruskal, P. Black, *Transactions of the American Philosophical Society* **82** (1992).

[2] D. M. Weeks, Hittite vocabulary: An Anatolian appenix to Buck's "Dictionary of selected synonyms in the prinicipal Indo-European languages", Ph.D. thesis, University of California, Los Angeles (1985).

[3] J. Ludewig. University of Freiburg.

[4] J. Cheung, *Etymological Dictionary of the Iranian Verb* (Leiden: Brill, 2007).

[5] M. Dunn Max Planck Institute for Psycholinguistics.

[6] R. Beekes, *Etymological Dictionary of Ancient Greek* (Leiden: Brill, 2010).

[7] A. Walde, *Vergleichendes Woerterbuch der Indogermanischen Sprachen* (de Gruyter, 1930).

[8] J. Mallory, D. Adams, *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World* (Oxford, 2006).

[9] J. Pokorny, *Indogermanisches etymologisches Wörterbuch* (Tuebingen: Francke, 1994).

[10] V. Orel, *Albanian Etymological Dictionary* (Leiden: Brill, 1998).

[11] C. D. Buck, *A dictionary of selected synonyms in the principal Indo-European languages* (Chicago: University of Chicago Press, 1949).

[12] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Albanian_Swadesh_list.

[13] M. d. Vaan, *Etymological Dictionary of Latin and the other Italic Languages* (Leiden: Brill, 2008).

[14] M. Woodhouse, *English-Greek Dictionary - A Vocabulary of the Attic Language* (London: George Routledge and Sons, 1910).

[15] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Ancient_Greek_Swadesh_list.

[16] D. J. Mastronarde, Ancient greek tutorials http://socrates.berkeley.edu/~ancgreek/ancient_greek_start.html (1999–2005).

[17] Wiktionary .

[18] R. Beekes, *Comparative Indo-European Linguistics: An Introduction* (Amsterdam: John Benjamins, 1995).

[19] H. Martirosyan, *Etymological Dictionary of the Armenian Inherited Lexicon* (Brill, 2010).

[20] D. Ringe, T. Warnow, A. Taylor, *Transactions of the Philological Society* **100**, 59 (2002).

[21] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Slavic_languages.

[22] K. Bergsland, H. Vogt, *Current Anthropology* **3**, 115 (1962).

[23] R. Matasovic, *Etymological Dictionary of Proto-Celtic* (Leiden: Brill, 2009).

[24] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Armenian_Swadesh_list.

[25] K. Bellamy-Dworak. Max Planck Institute for Psycholinguistics.

[26] D. P. Pattanayak, *A Controlled Historical Reconstruction of Oriya, Assamese, Bengali and Hindi* (The Hague: Mouton and Co, 1966).

[27] G. Grierson, *Linguistic Survey of India. Vol. V. Indo-Aryan Family. Eastern Group, pt.1. Specimen of the Bengali and Assamese Languages* (Delhi: Motilal Banarsidass, 1903 [1968]).

[28] X. Community, Assamese dictionary http://www.xobdo.net/dic/index.php.

[29] R. L. Turner, *A Comparative and Etymological Dictionary of the Nepali* (London: Kegan Paul, Trench, Trubner and Co, 1931).

[30] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Indo-Iranian_Swadesh_lists_(extended).

[31] R. Trail, *Patterns in Clause, Sentence and Discourse in Selected Languages of India and Nepal* (SIL, 1973).

[32] A. Davis, *Basic Colloquial Maithili. A Maithili-Nepali-English Vocabulary with Some Structural Notes* (Delhi: Motilal, 1984).

[33] G. Grierson, *Linguistic Survey of India. Vol. V. Indo-Aryan Family. Eastern Group, pt.2. Specimen of the Bihari and Oriya Languages* (Delhi: Motilal Banarsidass, 1903 [1968]).

[34] C. G., D. Jain, *The Indo-Aryan Languages* (London: Routledge, 2003).

[35] R. Delaporte, *Elementary Breton-English and English-Breton Dictionary, Second Edition* (Lesneven: Mouladurioù Hor Yezh, 2006 [1979]).

[36] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Celtic_languages.

[37] Wiktionary, http://en.wikipedia.org/wiki/Swadesh_list_of_Slavic_languages.

[38] R. Derksen, *Etymological Dictionary of the Slavic Inherited Lexicon* (Leiden: Brill, 2008).

[39] R. B. Dictionaries, *Catalan Dictionary: English-Catalan, Catalan-English* (London: Routledge, 1993).

[40] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_further_Romance_languages.

[41] F. Diez, *An Etymological Dictionary of the Romance Languages* (Truebner, 1864).

[42] D. Boutkan, S. M. Siebinga, *Old Frisian Etymological Dictionary* (Leiden: Brill, 2005).

[43] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Celtic_Swadesh_lists.

[44] F. Jago, *An English-Cornish Dictionary. Compiled from the best sources* (London: Simpkin, Marshall and Co, 1887).

[45] I. Wmffre, *Late Cornish* (Munich: Lincom Europa, 1998).

[46] G. T. Gambill, http://www.freelang.net/online/cornish.php?lg=gb.

[47] M. e. Ball, *The Celtic Languages* (London: Routledge, 1993).

[48] D. D. S. og Litteraturselskab, Den Danske Ordbog. http://ordnet.dk/ddo/ (2011).

[49] D. Erschler, Field notes (2011).

[50] N. Osselton, R. Hempelman, *The New Routledge Dutch Dictionary: Dutch-English, English-Dutch* (Utrecht-Antwerpen: Van Dale Lexicografie, 2003).

[51] E. Klein, *A Comprehensive Etymological Dictionary of the English Language* (Amsterdam: Elsevier, 1986).

[52] F. Kluge, *Etymologisches Woerterbuch der deutschen Sprache* (Berlin: de Gruyter, 1975).

[53] M. Makins, A. Isaacs, D. Adams, *Collins English Dictionary* (Glasgow: HarperCollins, 1994 [1979]).

[54] O. online, http://dictionary.oed.com/.

[55] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Germanic_languages.

[56] M.-H. Corréard, V. Grundy, *The Oxford-Hachette French Dictionary, French-English, English-French* (Oxford: Oxford University Press, 1994).

[57] C. N. de Ressources Textuelles et Lexicales, http://www.cnrtl.fr/etymologie/.

[58] J. Zantema, *Frysk Wurdboek, frysk-nederlânsk* (Leeuwarden/Ljouwert: A.J. Osinga Uitgeverij, 1984).

[59] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Friulian_Swadesh_list.

[60] H. Cox, W. Martin, W. Pijnenburg, P. van Sterkenburg, G. Tops, *Groot woordenboek Duits-Nederlands* (Utrecht/Antwerpen: Van Dale Lexicografie, 1983).

[61] L. GmbH, Online german dictionary http://dict.leo.org/.

[62] G. Köbler, *Gotisches Wörterbuch* (Leiden: Brill, 1989).

[63] O. Priese, *Deutsch-gotisches Wörterbuch* (Halle: Niemeyer, 1933).

[64] W. Krause, *Handbuch des Gotischen* (Muenchen: C.H.Beck, 1953).

[65] J. Heinsius, *Nederlandsch-Gotische Woordenlijst* (Groningen: Noordhoff, 1893).

[66] A. Kloekhorst, *Etymological Dictionary on the Hittite Inherited Lexicon* (Leiden: Brill, 2008).

[67] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Greek_Swadesh_list.

[68] G. Cardona, B. Suthar, *The Indo-Aryan Languages* (London: Routledge, 2003), chap. Gujarati.

[69] C. Shapiro, Michael, *The Indo-Aryan Languages* (London: Routledge, 2003), chap. Hindi.

[70] J. Puhvel, *Hittite Etymological Dictionary* (Berlin: Mouton de Gruyter, 1984).

[71] E. Love, Catherine, *Collins Giunti Marzocco Italian-English, English-Italian Dictionary* (Collins, 1993 [1985]).

[72] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Indo-Iranian_Swadesh_lists.

[73] J. A. Rea, *Current Trends in Linguistics* **11**, 355 (1973).

[74] B. Kessler, *The Significance of Word Lists* (Stanford: CSLI Publications, 2001).

[75] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Swadesh_lists_for_Iberian_languages.

[76] A. Cioranescu, *Dictionarul etimologic al limbii romane* (Bucarest: Saeculum, 2002).

[77] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Latvian_Swadesh_list.

[78] Wiktionary, http://en.wiktionary.org/wiki/Appendix:Swadesh_list_for_Baltic_languages.

[79] R. V. Pandharipande, *Marathi* (London: Routledge, 1997).

[80] L. Gusain, *Marwari* (Munich: Lincom Europa, 2004).

[81] G. Grierson (1908 [1968]).

[82] G. Macalister, A dictionary of the dialects spoken in the state of Jeypore. http://dsal.uchicago.edu/dictionaries/macalister/.

[83] JatLand.com, http://www.jatland.com/home/Glossary_of_Rajasthani_Language.

[84] H. Singh, http://www.freewebs.com/hanvant/rajasthani_dictionary.htm.

[85] T. Riccardi, *The Indo-Aryan Languages* (London: Routledge, 2003), chap. Nepali.

[86] R. Auty, *Handbook of Old Church Slavonic. Part II. Text and Glossary* (London: Athlone Press, 1960).

[87] J. Deschler, *Kleines Wörterbuch der kirchenslavischen Sprache: Wortschatz der gebräuchlichsten liturgischen Texte mit deutscher übersetzung* (Munich: Otto Sagner, 1987).

[88] G. Nandris, *Handbook of Old Church Slavonic. Part I. Old Church Slavonic Grammar* (London: Athlone Press, 1959).

[89] G. Jember, *English-Old English, Old English-English Dictionary* (Boulder, Colo.: Westview Press, 1975).

[90] V. Mažiulis, The etymological dictionary of prussian `http://donelaitis.vdu.lt/prussian/databan.htm`.

[91] M. Klussis, Dictionary of revived Prussian: Prussian-English, English-Prussian `http://donelaitis.vdu.lt/prussian/Engl.pdf` (2005–2006).

[92] L. Neukom, M. Patnaik, *A Grammar of Oriya. (=Arbeiten des Seminars fuer Allgemeine Sprachwissenschaft 17)* (Zurich: University of Zurich, 2003).

[93] F. Rau, personal communication.

[94] Wiktionary, `http://en.wiktionary.org/wiki/Appendix:Polish_Swadesh_list`.

[95] Wiktionary, `http://en.wiktionary.org/wiki/Appendix:Portuguese_Swadesh_list`.

[96] Wiktionary, `http://en.wiktionary.org/wiki/Appendix:Romanian_Swadesh_list`.

[97] C. Howlett, *The Pocket Oxford Russian Dictionary* (Oxford University Press, 1994).

[98] Wiktionary, `http://en.wiktionary.org/wiki/Appendix:Scottish_Gaelic_Swadesh_list`.

[99] J. Gvozdanović, *Indo-European Numerals* (Mouton de Gruyter, 1992).

[100] E. Trumpp, *Grammar of the Sindhi Language* (Delhi: Jetley, 1986).

[101] P. Mewaram, Sindhi-English dictionary. `http://dsal.uchicago.edu/dictionaries/mewaram/`.

[102] D. Chandralal (2010).

[103] Wiktionary, `http://en.wiktionary.org/wiki/Appendix:Spanish_Swadesh_list`.

[104] H. Kornitzky, E. Engbrant-Heider, *Langenscheidts Taschenwörterbuch, Schwedisch, Schwedisch-Deutsch, Deutsch-Schwedisch* (Berlin/München: Langenscheidt, 1984).

[105] A. J. v. Windekens, *Le Tokharien confronté les autres langues indo-européennes. Volume 1. La phonétique et le vocabulaire* (Louvain: Centre International De Dialectologie Générale, 1976).

[106] J. H. Jasanoff, *Tocharian and Indo-European Studies* **3**, 125 (1989).

[107] E. Seebold, *Etymologisches Wörterbuch der deutschen Sprache* (Walter de Gruyter, 1989).

[108] I. of Applied Linguistics Trust, `http://ialtrust.com/swadeshlanguagelist/`.

[109] S. Sanskrit, `http://spokensanskrit.de/`.

[110] M. Williams, Sanskrit-english dictionary (2008 revision) `http://www.sanskrit-lexicon.uni-koeln.de/mwquery/` (2008).

[111] A. Borooah, *English-Sanskrit Dictionary* (Assam: Publication board, 1887 [1971]).

[112] M. Monier-Williams, *A Sanskrit-English Dictionary* (Oxford: Clarendon Press, 1899).

[113] W. O. Pughe, *A dictionary of the Welsh language. [Preceded by] A Welsh grammar* (Denbigh: Thomas Gee., 1832).

Table S2: **Age Calibration Information.** Prior distributions for the ages of ancient languages. We incorporate prior knowledge on both the age of the ancient languages (tips) and the age of common ancestors of language groups (internal nodes) based on historical sources to calibrate the time-scale of our language trees. The first three entries list the prior distributions used as calibrations for the tips, and the remainder for internal nodes.

Language	Calibration	Historical Information
Hittite	Normally distributed with mean 3450.0 and 125.0 standard deviation	Oldest Hittite text of King Anittas from the 18th century BCE. Latest texts from the 14th-13th centuries BCE.
Luvian	Normally distributed with mean 3350.0 and 75.0 standard deviation	Cuneiform Luvian texts 16th to 13th century BCE.
Lycian	Normally distributed with mean 2400.0 and 50.0 standard deviation	Lycian inscriptions from 500BCE to 300BCE
Vedic Sanskrit	Normally distributed with mean 3000.0 and 100.0 standard deviation	Vedic period 1200BCE to 800BCE.
Avestan	Normally distributed with mean 2500.0 and 50.0 standard deviation	'Younger' Avestan attested around 500BCE
Old Persian	Normally distributed with mean 2450.0 and 75.0 standard deviation	Cuneiform inscriptions of the Achaemenid era (600BCE-300BCE)
Ancient Greek	Normally distributed with mean 2400.0 and 50.0 standard deviation	Classical Attic
Umbrian	Normally distributed with mean 2200.0 and 100.0 standard deviation	Umbrian inscriptions 300BCE to 100BCE
Oscan	Normally distributed with mean 2200.0 and 100.0 standard deviation	Oscan inscriptions 300BCE to 100BCE
Latin	Normally distributed with mean 2050.0 and 75.0 standard deviation	Classical Latin
Gothic	Normally distributed with mean 1650.0 and 25.0 standard deviation	Visigothic source texts
Old High German	Normally distributed with mean 1050.0 and 50.0 standard deviation	9th century East Franconian source texts

Continued on next page

Table S2 continued – **Age Calibration Information.**

Language	Calibration	Historical Information
Old English	Normally distributed with mean 1000.0 and 50.0 standard deviation	Late West Saxon Old English
Old Norse	Normally distributed with mean 775.0 and 40.0 standard deviation	Old Icelandic source texts
Classical Armenian	Normally distributed with mean 1450.0 and 75.0 standard deviation	Classical Armenian source texts from the 5th to 7th century
Old Irish	Normally distributed with mean 1250.0 and 75.0 standard deviation	Old Irish source texts
Tocharian A	Normally distributed with mean 1375.0 and 75.0 standard deviation	Earliest texts from later half of 1st millennia CE. No texts after 750CE by which time Tocharians are thought to have been assimilated with Turkish invaders.
Tocharian B	Normally distributed with mean 1350.0 and 75.0 standard deviation	Earliest texts from later half of 1st millennia CE. No texts after 750CE by which time Tocharians are thought to have been assimilated with Turkish invaders.
Old Church Slavonic	Normally distributed with mean 1000.0 and 50.0 standard deviation	Source texts from 10th and 11th century CE
Old Prussian	Normally distributed with mean 500.0 and 50.0 standard deviation	15th and 16th century Old Prussian source texts
Lithuanian/Latvian	Normally distributed with mean 1350.0 and 25.0 standard deviation	Historical sources indicate differentiation of Lithuanian and Latvian (Latgalian) in the 7th century as Proto-Latvian and Proto-Lithuanian tribes emerged,
Balto-Slavic	Normally distributed with mean 3100.0 and 600.0 standard deviation (truncated from 2000 to 3400)	Distinct Slavic culture and language known to pre-date 100CE on the basis of Tacitus's "Germany". Archaeological evidence suggests the split may have occurred as early as 1,400BCE.
Northwest Germanic	Normally distributed with mean 1875.0 and 67.0 standard deviation	Earliest attested North Germanic inscriptions date from 3rd century CE.

Continued on next page

Table S2 continued – **Age Calibration Information.**

Language	Calibration	Historical Information
Indo-Iranian	Truncated between 3000.0 and 10000.0	Rgveda, an identifiably Indic collection of sacred texts, is thought to date from between 1450-1000BCE. The Avesta, a similar Iranian collection of sacred texts, has been recorded in oral tradition since before 800BCE.
Iranian	Log normally distributed with mean 400.0 and 0.8 standard deviation with offset of 2600.0	By 500BCE Old Persian was distinct from the Eastern Iranian dialects.
Tocharic	Log normally distributed with mean 200.0 and 0.9 standard deviation with offset of 1650.0	Tocharian languages are thought to have diverged shortly after the fall of Bactria (135BCE) and no later than 100 years before the first known inscriptions of Tocharian B.
West Germanic	Normally distributed with mean 1550.0 and 25.0 standard deviation	Anglo-saxons began to settle Britain from around 400CE.
French/Iberian	Normally distributed with mean 1400.0 and 100.0 standard deviation	Beginning of repetition of Latin liturgical formulas without comprehension in sixth to eighth centuries CE. Strasburg Oaths, 842CE.
Indic	Log normally distributed with mean 1000.0 and 1.0 standard deviation with offset of 2150.0	Singhalese records dating from as early as 2nd century BCE indicate that Indic languages had begun to diverge by this time.
Celtic	Log normally distributed with mean 2000.0 and 0.6 standard deviation with offset of 1200.0	Archaic Irish inscriptions date back to the 5th century CE - divergence must have occurred well before this time.
Latin/Romance	Normally distributed with mean 2000.0 and 135.0 standard deviation	Last Roman troops withdrawn to south of Danube, 270CE. Dacia conquered by Rome, 112CE.
Slavic	Log normally distributed with mean 300.0 and 0.6 standard deviation with offset of 1200.0	Old Church Slavonic and East Slavic texts date to beginning of 9th century and indicate significant divergence by this time. The split must have occurred after the Balto-Slavic divergence.
Brythonic	Normally distributed with mean 1550.0 and 25.0 standard deviation	Migrants from Britain colonize Brittany in 5th century CE.

Continued on next page

Table S2 continued – **Age Calibration Information.**

Language	Calibration	Historical Information
Greek split	before 1,500BC	Earliest form of an ancient Greek dialect is Mycenaean, attested in Linear B texts dating from 15th century BCE.

Table S3: Log probability of the CTMC, covarion and stochastic Dollo substitution models with and without relaxed clock. Logarithm of Bayes factors comparing the models pair-wise.

	$\ln(P(model\|data))$			CTMC		Covarion		S.Dollo	
				relaxed	strict	relaxed	strict	relaxed	strict
CTMC relaxed clock	-53017	±	0.9	0	-147	415	263	646	462
CTMC strict clock	-53355	±	0.2	147	0	562	410	793	608
Covarion relaxed clock	-52061	±	0.2	-415	-562	0	-152	231	46
Covarion strict clock	-52411	±	0.2	-263	-410	152	0	383	198
S.Dollo relaxed clock	-51530	±	0.5	-646	-793	-231	-383	0	-184
S.Dollo strict clock	-51954	±	0.2	-462	-608	-46	-198	184	0

Table S4: BRW and RRW model comparison Numbers in brackets are 95% HPD intervals.

	Brownian RW	Cauchy-RRW
Log Posterior	-52566.52	-52293.97
Marginal LnL	-51207.26	-50958.20
Hyperparameters	NA	$\nu/2 = 0.5$
Coefficient of variation	NA	1.38 (1.21,1.55)
Dispersal rate (km/year)	0.60 (0.52,0.69)	0.48 (0.42,0.55)
	gamma-RRW	lognormal-RRW
Log Posterior	-52355.45	-52240.96
Marginal LnL	-51032.26	-50939.34
Hyperparameters	$\nu/2 = 0.75(0.45, 1.08)$	$\sigma = 2.43(1.92, 2.90)$
Coefficient of variation	1.14 (0.89,1.43)	4.35 (2.33,6.51)
Dispersal rate (km/year)	0.49 (0.43,0.56)	0.48 (0.42,0.55)

Table S5: Correspondence between languages in the Ringe et al.(*16*) data set and the major sub-groups in our language data. These correspondences were used to produce the constraints illustrated in Figure S12.

Constraint name	Ringe Language	Language set
Anatolian	Hittite, Lycian, Luvian	Hittite, Lycian, Luvian
Tocharian	Tocharian A, Tocharian B	Tocharian A, Tocharian B
OscanUmbrian	Oscan, Umbrian	Oscan, Umbrian
Romance	Latin	Provencal, French, Walloon, Sardinian N, Sardinian L, Sardinian C, Spanish, Portuguese ST, Catalan, Romanian List, Vlach, Italian, Ladin, Friulian, Romansh, Latin
Celtic	Old Irish, Welsh	Irish A, Old Irish, Welsh N, Welsh C, Breton List, Breton SE, Breton ST, Cornish, Scots Gaelic
Albanian	Albanian	Albanian Top, Albanian G, Albanian K, Albanian C
EastGermanic	Gothic	Gothic
NorthGermanic	Old Norse	Swedish Up, Swedish VL, Swedish List, Danish, Riksmal, Icelandic ST, Faroese, Old Norse
WestGermanic	Old English, Old High German	Old English, English ST, German ST, Dutch List, Flemish, Frisian, Luxembourgish, Old High German
Greek	Greek	Greek ML, Greek Mod, Ancient Greek
Armenian	Armenian	Armenian Mod, Armenian List, Classical Armenian
LithuanianLatvian	Lithuanian, Latvian	Lithuanian ST, Latvian
OldPrussian	Old Prussian	Old Prussian
Slavic	Old Church Slavonic	Slovenian, Lusatian L, Lusatian U, Czech, Slovak, Czech E, Ukrainian, Byelorussian, Polish, Russian, Macedonian, Bulgarian, Serbocroatian, Old Church Slavonic
Indic	Vedic	Romani, Singhalese, Kashmiri, Marathi, Gujarati, Lahnda, Hindi, Bengali, Nepali List, Assamese, Bihari, Marwari, Oriya, Sindhi, Urdu, Vedic Sanskrit
Irianian	Avestan, Old Persian	Afghan, Waziri, Persian List, Tadzik, Baluchi, Wakhi, Avestan, Digor Ossetic, Iron Ossetic, Kurdish, Old Persian