Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-022-01211-7 OPEN

Detecting contact in language trees: a Bayesian phylogenetic model with horizontal transfer

Nico Neureiter[1,2,3✉], Peter Ranacher[1,2,3,4], Nour Efrat-Kowalsky[1,5], Gereon A. Kaiping[2,3], Robert Weibel[1,2,3,4], Paul Widmer[1,3,4,5] & Remco R. Bouckaert[6,7]

Phylogenetic trees are a central tool for studying language evolution and have wide implications for understanding cultural evolution as a whole. For example, they have been the basis of studies on the evolution of musical instruments, religious beliefs and political complexity. Bayesian phylogenetic methods are transparent regarding the data and assumptions underlying the inference. One of these assumptions—that languages change independently—is incompatible with the reality of language evolution, particularly with language contact. When speakers interact, languages frequently borrow linguistic traits from each other. Phylogenetic methods ignore this issue, which can lead to errors in the reconstruction. More importantly, they neglect the rich history of language contact. A principled way of integrating language contact in phylogenetic methods is sorely missing. We present `contacTrees`, a Bayesian phylogenetic model with horizontal transfer for language evolution. The model efficiently infers the phylogenetic tree of a language family and contact events between its clades. The implementation is available as a package for the phylogenetics software BEAST 2. We apply `contacTrees` in a simulation study and a case study on a subset of well-documented Indo-European languages. The simulation study demonstrates that `contacTrees` correctly reconstructs the history of a simulated language family, including simulated contact events. Moreover, it shows that ignoring contact can lead to systematic errors in the estimated tree height, rate of change and tree topology, which can be avoided with `contacTrees`. The case study confirms that `contacTrees` reconstructs known contact events in the history of Indo-European and finds known loanwords, demonstrating its practical potential. The model has a higher statistical fit to the data than a conventional phylogenetic reconstruction, and the reconstructed tree height is significantly closer to well-attested estimates. Our method closes a long-standing gap between the theoretical and empirical models of cultural evolution. The implications are especially relevant for less documented language families, where our knowledge of past contacts and linguistic borrowings is limited. Since linguistic phylogenies have become the backbone of many studies of cultural evolution, the addition of this integral piece of the puzzle is crucial in the endeavour to understand the history of human culture.

A full list of author affiliations appears at the end of the paper.

Introduction

Phylogenetic trees are used to represent the evolutionary history of a language family from its descent from a common ancestor at the stem to the diversification into branches and the observed languages at the leaves. Traditionally, the linguistic comparative method would infer phylogenies qualitatively by comparing related languages on shared traits (Atkinson and Gray, 2005). Recent computational methods make it possible to model the process of evolution explicitly, disclosing both the assumptions and the data used for inference (Bowern, 2018). Bayesian phylogenetic inference is the most popular of these methods. It promises to reconstruct the relationships between languages in a family—represented in the phylogenetic tree, its topology and the branch lengths—while capturing the uncertainty of each of these parameters.

Bayesian phylogenetic inference has been applied to different problems of cultural evolution. As a method in comparative linguistics, Bayesian phylogenetics generates classifications of entire families or single branches, for which the sub-structure was previously unclear, e.g. the higher-order structure of Indo-European (Bouckaert et al., 2012), or Timor-Alor-Pantar (Kaiping and Klamer, 2022). In addition, phylogenetic inference explicitly reveals the dynamics of language evolution concerning technical innovation (Gray et al., 2009), cultural processes (Tehrani, 2020), cognitive processes (Widmer et al., 2017), migration pathways (Grollemund et al., 2015) and changes in subsistence (Sagart et al., 2019). Many findings stemming from phylogenetic inference have profound implications for linguistic theory and the field of cultural evolution as a whole. Phylogenetics has taught us, for instance, that changes in habitat facilitated the expansion of the Bantu family (Grollemund et al., 2015), that grammar evolves faster than the lexicon in Austronesian languages (Greenhill et al., 2017) and that the Sino-Tibetan language family originated in Northern China around 7200 B.P. (Sagart et al., 2019). Conversely, the more wide-ranging the implications of a method, the more critical it becomes that its assumptions are well-founded.

Bayesian phylogenetic inference has its roots in evolutionary biology and its model assumptions fit biological evolution but less so linguistics. The most obvious example is how (DNA) sequences are assumed to change according to continuous-time Markov chains (CTMC). In a CTMC, characters evolve continuously, independently changing from one state to another according to random mutation rates. Initially, these models from bioinformatics were applied by analogy to infer language trees from cognates (Gray and Atkinson, 2003), that is, words with a common etymological origin. The transfer also happened with other models grounded in biology, such as the coalescent tree prior (Chang et al., 2015; Chousou-Polydouri et al., 2016; Rama, 2018), which is motivated by descent from a common ancestor in a particular population model.

More recently, however, models have been tested or newly developed explicitly targeting language evolution. The Dollo-like evolution model requires that characters can only be gained once (Bouckaert and Robbeets, 2017), intuitively matching the evolution of cognates. The model of concerted evolution allows for parallel changes in several sites, capturing systematic sound changes (Hruschka et al., 2015). Ordinal models (Bouckaert, 2019) can be used to model the evolution of ordinal characters, e.g. the complexity of tonal systems ranging from "no tones" to "highly complex tone system". This is less the case for the overall shape of linguistic phylogenies, where systematic linguistics-based model recommendations are still lacking. There are very few recent tests of existing tree priors (Rama, 2018; Ritchie and Ho, 2019), and various questions have been posed, but not quantitatively and systematically investigated: Are there other (computationally tractable) tree priors that could be derived from demic expansion models, instead of assuming speciation at a constant rate, which would fit better with expanding language families (Neureiter et al., 2021)? Are binary trees a good model for language phylogenies (Maurits et al., 2019)? What clock model would be appropriate for language evolution, given known effects of schismogenesis (Atkinson et al., 2008; Bateson 1935)?

Most importantly, nearly all phylogenetic models applied to languages assume that evolution is entirely tree-like. If shared properties are found in two languages, they are assumed to have either derived from a common ancestor or arisen independently in the two branches. This assumption is problematic. In fact, a recent debate has been whether the tree model of language evolution has merit at all (François, 2015). While the strict tree model is an oversimplification of language history, the phylogenetic framework of inheritance from a common ancestor is still useful (Jacques and List, 2019). This paper aims to address a recurring criticism of phylogenetic models, the fact hat they do not account for contact. When languages interact, this often leads to horizontal transfer or borrowing where languages exchange linguistic material. The most readily recognisable borrowings are lexical items ("loanwords"), but borrowing may also involve other structural features from sounds to discourse strategies (Grossman et al., 2020; Muysken, 2011; Thomason and Kaufman, 1989). Horizontal transfer is incompatible with the assumption of a tree of independently evolving branches.

In the past, most language phylogenies bypassed this issue by marking forms that were known to be borrowed as not related to their origin form (Bouckaert et al., 2012; Kolipakam et al., 2018). Combining linguistic expertise with computational methods of borrowing detection ensures that the immediate conclusions drawn from phylogenies are likely to hold. In addition, simulation studies suggest that even a decent amount of borrowings has no significant influence on inferred phylogenies (Greenhill et al., 2009). However, these simulations assume a continuous rate of single borrowings, while it remains open whether concentrated borrowing events have a more significant impact on the reconstruction, in particular regarding estimated tree heights. More importantly, omitting horizontal transfer in phylogenetic models is deeply unsatisfying. Contact plays a substantial role in language histories and their interpretation and is thus a valuable outcome of a computational inference in and of itself.

Various attempts have been made at making horizontal transfer more transparent and explicit, but none has successfully added borrowing inference to large-scale language phylogenies. Outside the domain of phylogenetic inference, statistics such as the Q residuals (Gray et al., 2010), δ scores (Holland et al., 2002), and TIGER values (Syrjänen et al., 2021) quantify the amount of non-tree-like signal in language data. In addition, methods like NeighborNets (Bryant and Moulton, 2002) and splitsTree (Huson and Bryant, 2006) visualise to what extent and between what languages the data conflicts with a strict tree assumption.

The character-based visualisation method *Minimal Lateral Networks* (Dagan and Martin, 2007) shows which nodes in a tree share material beyond what a backbone tree would imply. In linguistics, the method is suitable for visualising language contact (Nelson-Sathi et al., 2011), but with the disadvantage of needing a pre-existing underlying tree.

In terms of phylogenetic methodology, Willems et al. (2016) presented a distance-based, non-Bayesian inference method of linguistic phylogenetic networks. Even earlier, Nakhleh et al. (2005) introduced "perfect phylogenetic networks", a character-based method to infer language phylogenies based on a maximum parsimony score that is modified to include borrowings. In addition, Dellert (2019) developed a causal inference method to

add edges of lexical flow to a pre-existing set of phylogenetic trees. None of these methods, however, is compatible with state-of-the-art probabilistic frameworks for phylogenetic inference. Recently though, some methods have been suggested in that domain. Kelly and Nicholls (2017) developed a Bayesian phylogenetic method that allows borrowing in a stochastic Dollo model. Due to the complexity of the numerical computations, however, their model is only practically applicable to small trees. Furthermore, the horizontal transfer edges are not considered individually, but only in sum of all their possibilities ("integrated out"), such that the inferred borrowing events cannot be inspected or interpreted.

The problem of horizontal transfer is not unique to cultural evolution. In biology, horizontal gene transfer (HGT) occurs in bacteria, which recombine by transferring part of the DNA from one microbe to another. Similar to linguistic borrowing, this violates the assumptions of tree-based phylogenetics. In phylogenetic models with HGT, a tree still describes the history of the sites, but each site can follow a different tree. This idea is similar to that of gene-trees within a species-tree in the multispecies coalescent (MSC) model (Heled and Drummond, 2009), which solves the problem of genetic polymorphism within species. However, the MSC does not allow a common ancestor of two genes to be more recent than the most recent common ancestor of the corresponding species. Species networks relax this assumption by allowing HGT at reticulation events (Wen et al., 2016; Zhang et al., 2018). Unfortunately, species networks are extremely computationally expensive, making inference infeasible for most language families.

The ClonalOrigin model (Didelot et al., 2010) explicitly models horizontal transfer of consecutive DNA segments via conversion edges, which connect two branches of the underlying tree or "clonal frame". Due to potentially unsampled intermediate lineages, the transferred DNA segment reaches the receiver lineage with a time delay after branching from the donor lineage. Again, the different paths that the DNA segments can take at each conversion edge define different trees for different segments of DNA. The ClonalOrigin model can jointly infer the clonal frame and the conversion edges. An implementation of the model is available in the BEAST 2 package Bacter (Vaughan et al., 2017). However, the assumption that consecutive segments are transmitted at each conversion edge makes the model unsuitable for cultural evolution. Lexical features —or potentially other cultural traits—do not follow a natural order that would make this assumption plausible.

This paper presents `contacTrees`, a Bayesian phylogenetic model to infer the phylogenetic history of a language family and horizontal transfer between its branches. `contacTrees` builds on the ClonalOrigin model, with crucial changes to make the model suitable for linguistic data and with novel MCMC operators to improve sampling efficiency.

Methods and data

The `contacTrees` model. Phylogenetic models represent the ancestry of a language family by a rooted binary time tree $\mathcal{T}$. The tree's leaves represent languages in our data set, and the internal nodes represent ancestral languages. Each node has a corresponding height. For leaves, the height gives the sampling date; for internal nodes, it describes the age of all its descendants' most recent common ancestor. In the `contacTrees` model, the language tree $\mathcal{T}$ is complemented by a set of horizontal contact edges $\mathcal{C}$, representing events in history where one language borrowed words (or any linguistic traits) from another language. These contact edges relax the assumption that the history of all words needs to follow one shared language tree. A contact edge allows words to come from a different ancestor language than the one proposed by the language tree. Each word by itself still has only one ancestor and still follows a tree (Fig. 1), but this word tree does not always match the language tree.

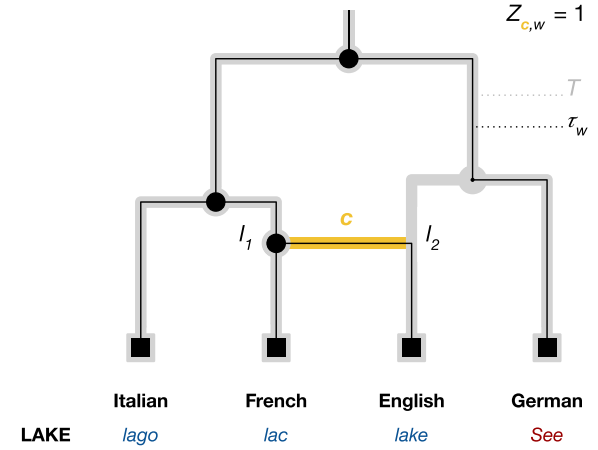

Fig. 1 Illustration of a word tree τ_w (black) within the a language tree T (grey). As indicated by $Z_{c,w} = 1$, the word is borrowed at the contact edge c, causing the word tree to deviate from the language tree. For example, the Middle English word *lake* was borrowed from Norman French *lac*, leading to matching form—meaning traits (indicated by blue/red colouring) in English and French.

A contact edge $c = (l_1, l_2, t) \in \mathcal{C}$ models a contact event at time t from a donor l_1 to a receiver l_2. For each contact edge c, a list of binary parameters Z_c defines which words are borrowed along that edge. Let $\mathcal{W}$ be the set of all words, with corresponding borrowing indicators $Z_{c,w} \in \{0, 1\}$ for every word $w \in \mathcal{W}$. $Z_{c,w} = 1$ indicates that w is borrowed from l_1 into l_2 at edge c, while $Z_{c,w} = 0$ indicates it is not borrowed and—barring other contact edges—it is inherited from l_2's parent instead. Together, the language tree $\mathcal{T}$, the edges $\mathcal{C}$ and the borrowing indicators $Z_{\cdot,w}$ define a word tree τ_w for word w, which deviates from the language tree $\mathcal{T}$ whenever $Z_{c,w}$ is 1 (see Supplementary Material, Section S1.2, for a detailed explanation of how word trees are computed).

Each word $w \in \mathcal{W}$ has associated data X_w. We denote the whole data set of all words together by $X = \{X_w | w \in \mathcal{W}\}$. The `contacTrees` likelihood can be computed as a product of standard tree likelihoods over all words in $\mathcal{W}$:

$$\mathrm{P}\left(X \mid \mathcal{T}, \mathcal{C}, Z, \theta_X\right) = \prod_{w \in \mathcal{W}} \mathrm{P}\left(X_w \mid \tau_w, \theta_X\right), \tag{1}$$

where θ_X are all parameters of the tree likelihood, including the substitution model parameters and branch rates (see Supplementary Material, Section S5.2).

The `contacTrees` model defines a network prior, a combination of a tree prior $\mathrm{P}\left(\mathcal{T} \mid \theta_{\mathcal{T}}\right)$ and a contact prior $\mathrm{P}(\mathcal{C} \mid \mathcal{T}, \Gamma)$:

$$\mathrm{P}\left(\mathcal{T}, \mathcal{C} \mid \theta_{\mathcal{T}}, \Gamma\right) = \mathrm{P}(\mathcal{C} \mid \mathcal{T}, \Gamma)\mathrm{P}\left(\mathcal{T} \mid \theta_{\mathcal{T}}\right) \tag{2}$$

The tree prior is an appropriate, arbitrary prior on $\mathcal{T}$ with parameters $\theta_{\mathcal{T}}$. There is a rich literature on tree priors for phylogenetic analysis (Drummond et al., 2005; Stadler et al., 2013), and discussions on which ones are appropriate for language trees (Rama, 2018). The contact prior is defined by a Poisson process along the branches of the tree T:

$$\mathrm{P}(C \mid T, \Gamma) = \Gamma^{|\mathcal{C}|} e^{-\Gamma} / L^{|\mathcal{C}|} \tag{3}$$

where L is the tree length and Γ is the expected number of edges. A derivation of this prior and an alternative contact prior are presented in the Supplementary Material, Section S1.1.

a)

$$T \sim \text{TreePrior}(\theta_T)$$
$$\mathcal{C} \sim \text{ContactPrior}(T, \Gamma)$$
$$Z_{\cdot,w} \sim \text{Binomial}(\beta)$$
$$\tau_w = \text{WordTree}(T, \mathcal{C}, Z_{\cdot,w})$$
$$X_w \sim \text{TreeLikelihood}(\tau_w, \theta_X)$$

b)

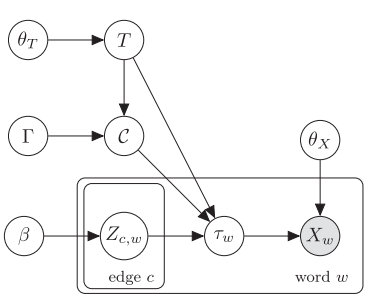

Fig. 2 The `contacTrees` **model.** The definition of the generative process (**a**) and its graphical representation (**b**).

At each contact edge $c \in \mathcal{C}$, an arbitrary subset of words is borrowed. The binary indicator variables $Z_{c,w} \in \{0, 1\}$ show whether word w was borrowed at c or not. We model the prior distribution of these word borrowings with a fixed borrowing probability $\mathrm{P}(Z_{c,w} = 1) = \beta$, resulting in a product of Bernoulli distributions:

$$\mathrm{P}(Z \mid \mathcal{C}, \beta) = \prod_{c \in \mathcal{C}} \prod_{w \in \mathcal{W}} \beta^{Z_{c,w}} (1-\beta)^{1-Z_{c,w}} = \beta^{\|Z\|_0} (1-\beta)^{\|1-Z\|_0} \tag{4}$$

While we use the term word to describe a unit of borrowing, it could in reality represent any trait or sequence of traits, as discussed in the sections "Data" and S3.

The full posterior of the `contacTrees` model combines the likelihood and priors defined above:

$$\mathrm{P}(\mathcal{T}, \mathcal{C}, Z, \beta, \Gamma, \theta_X, \theta_{\mathcal{T}} \mid X) \propto \mathrm{P}(X \mid \mathcal{T}, \mathcal{C}, Z, \theta_X) \cdot \mathrm{P}(Z \mid \mathcal{C}, \beta) \cdot \mathrm{P}(\mathcal{C} \mid \mathcal{T}, \Gamma) \tag{5}$$

$$\cdot \mathrm{P}(\mathcal{T} \mid \theta_{\mathcal{T}}) \cdot \mathrm{P}(\beta, \Gamma, \theta_X, \theta_{\mathcal{T}}). \tag{6}$$

Figure 2 denotes this posterior distribution as a generative process (a) and graphically as a Bayesian network using plate notation (b).

Inference. We use Markov-chain Monte Carlo (MCMC) to generate samples from the posterior distribution. While the posterior fully describes the desired behaviour of the model, adequate MCMC operators are crucial for efficient inference, especially for network approaches where the parameter space is exceptionally high-dimensional. We make use of standard MCMC operators provided in BEAST 2, adapt implementations of existing operators (mainly from the BEAST 2 package Bacter (Vaughan et al., 2017)) and implement additional operators specific to `contacTrees`. The added MCMC operators fall into three categories: (1) tree operators, (2) contact edge operators, and (3) borrowing operators. In category (1), we created adaptations of standard tree operators (e.g. subtree exchange, subtree slide (Drummond et al., 2002; Wilson and Balding, 1998), since any operator that changes the language tree needs to consider the effects that this change might have on existing contact edges. If a branch of the language tree moves, contact edges that are attached to this branch might become invalid and need to be adjusted accordingly. Contact edge operators leave the language tree unchanged, but they can add or remove contact edges, change their height or change their donor or receiver. A borrowing operator leaves the language tree and contact edges unchanged but proposes a new subset of words borrowed at a contact edge. Here, we introduce a new type of Gibbs operator that vastly improves the sampling performance over a naive random walk Metropolis implementation. For details on the MCMC operators, see Supplementary Material, Section S2.

Simulation study. We test the implementation of the model and operators in a simulation study, following a procedure described by Cook et al. (2006). In the simulation study, we demonstrate that the MCMC correctly samples from the posterior distribution and we show the effect of borrowing on tree-based phylogenetic models. The study works in three steps:

1. Trees, contact edges and parameters are simulated according to a pre-defined prior distribution.
2. Data is simulated according to the previously simulated parameters.
3. Trees, contact edges and parameters are sampled from the posterior distribution for the simulated data.

We repeat Step 3 with and without `contacTrees` and compare the performance. This way, we can assess the effect of borrowing on a phylogenetic reconstruction in an idealised setting, with ground truth and a clear statistical expectation for the posterior distribution.

In each of the 100 simulation runs, we simulate a tree with 25 languages according to a Yule process (Yule, 1925). The expected number of contact edges is 6.0. At every edge, each word can be borrowed with probability $\beta = 0.25$. Based on the tree, edges and other parameters, we simulate 100 words (representing the units of borrowing), with 20 sites per word. The sites of each word evolve according to a binary CTMC model along their word tree. The evolutionary rate can vary between branches according to a log-normal distribution with a standard deviation of 0.3. Priors on the height of three internal nodes calibrate the clock rate. For each of the 100 simulated data sets, we attempt to reconstruct the simulated parameters with the `contacTrees` model.

In a second setting, we reconstruct the same simulated data with the expected contact edges Γ set to 0, yielding a conventional phylogenetic model that assumes a purely vertical transfer of traits. Then we evaluate whether the 95% credible intervals inferred in the reconstruction cover the simulated parameters. When the model assumptions hold, and the simulation and model are well-calibrated, the coverage should fall between 91% and 99% for 95% of the parameters.

Case study. Finally, we present a case study on 39 Indo-European languages from the Celtic, Germanic and Romance branches (we will use the abbreviation *CGR* for these three clades). These languages are well documented and widely studied, which allows for a detailed discussion of the plausibility of our results.

The Indo-European languages have been subject to many phylogenetic studies before (Bouckaert et al., 2012; Chang et al., 2015; Gray and Atkinson, 2003), most of which were based on the IELex database (Dunn, 2012). For this study, we took the version of IELex published with Chang et al. (2015). We collected additional data on Medieval Latin, and we corrected clear coding errors (for details, see Supplementary Material, Section S5.1). The

data set also includes labels for known loanwords. We exclude the loanwords in one of the three runs described below.

We use a birth–death skyline model as a tree prior (Stadler et al., 2013). Together with the contact prior (Eq. (3)) and borrowing prior (Eq. (4)), this defines the prior distribution over networks and word trees within the network. We use an uncorrelated relaxed clock model (Drummond et al., 2006), which we calibrate through the ancient languages in our data set and through clade age priors on internal nodes (adapted from Bouckaert et al. (2012); see Supplementary Material, Section S5.3). As a substitution model, we use the binary covarion model (Tuffley and Steel, 1998) which demonstrated better performance than alternatives in previous studies (Bouckaert et al., 2018, 2012). For the contact prior, we define the expected number of contact edges to be $\Gamma = 0.25$ and the borrowing probability $\beta = 0.1$. The expected number of contact edges does not reflect our true expectations but represents a regularising prior to avoid overfitting by adding too many edges. For comparison with conventional phylogenetic methods, we run the analysis in three settings:

- CT: The `contacTrees` model with $\Gamma = 0.25$, as described above.
- noCT: A conventional phylogenetic set-up (without contact) by setting $\Gamma = 0$.
- noCT-filtered: A conventional phylogenetic set-up, with known loanwords removed from the data.

The first two scenarios are identical, apart from the different contact prior and, for efficiency, omission of MCMC operators that only affect contact edges, which makes them immediately comparable. However, the noCT scenario is not representative of other phylogenetic studies, which often remove loanwords from the data (Gray et al., 2009; Kolipakam et al., 2018). Hence we add the noCT-filtered scenario where we replace known loans by constant 0-strings (i.e. absence of all forms in that meaning class). We detail the MCMC set-up and runtime in the Supplementary Material, Section S5.4 and discuss general scalability in Section S4.3.

The `contacTrees` model is able to jointly infer the phylogeny and the contact edges. We show the corresponding reconstructions in the Supplementary Material, Section S5.5. For the discussion of our results, we focus on the ability of `contacTrees` to infer contact events. Since it is difficult to visualise and interpret both the uncertainty about the tree topology and the contact edges, we report the results of an analysis where we fixed the tree topology. Precisely, we infer contact events and node heights conditioned on a fixed topology based on the summary tree by Chang et al. (2015), a Bayesian phylogenetic analysis of the same languages.

Data. The `contacTrees` model defines a general framework that allows the phylogenetic tree of a linguistic trait to deviate from the language tree. However, the model does not make specific assumptions about traits and how they change over time. Most phylogenetic studies are based on lists of basic vocabulary for universal, culture-independent concepts such as "MOTHER", "EAT", or "SUN". These word lists, sometimes known as Swadesh lists (Swadesh, 1955), code traits by meaning class (as *form-meaning traits*): when two different languages share a cognate—a form with a common etymology—for a meaning, the trait is coded as present ("1"), otherwise as absent ("0"). For example, in Table 1, German and Old English share the cognate "*flaisk-" for the meaning class MEAT. Multiple forms may be synonyms and refer to the same meaning, e.g. *madi- and *keta- in Swedish (Table 1), but usually, word lists only include the most common form.

In this case study, we apply `contacTrees` to a subset of the IELex data (Dunn, 2012) that encodes the presence/absence of 1419 cognates in 206 meaning classes across 39 languages. The coding of the data has several implications on reconstruction with `contacTrees`, which we briefly address below.

Table 1 Three cognate classes of the meaning class MEAT in four sample languages.

	MEAT-1 *madi-	MEAT-2 *keta-	MEAT-3 ... *flaisk-
English	1	0	0
German	0	0	1
Old English	0	0	1
Swedish	1	1	0
...			

Since form-meaning traits are coded according to meaning classes, the model also needs to explain semantic shifts or changes in meaning. Cognates move in and out of meaning classes more easily than being gained or lost altogether. Hence, semantic shifts can result in homoplasy, parallel innovation in two or more languages. The Russian *xod* and Ancient Greek *hodós*, for example, both derive from the Proto-Indo-European word **sodó-*, which meant 'sitting'. In both languages, the meaning changed to 'walking, journey' independently.

Next, IELex uses meaning classes as the unit of borrowing. This implies that a loanword replaces all previously used forms for a given meaning, which is plausible if the data contains only the most common form. However, it results in the undesirable effect that synonyms are always borrowed or replaced together.

Synonymy in ancestral states can result in different patterns depending on the synonym selected by the researcher. If different synonyms are selected for closely related languages, this implies very volatile and parallel changes along the language tree. Continental Germanic, for example, has three synonymous words for the concept SMALL. In the IELex data, two of them are selected for Old High German exclusively (*small, luzzil*). At the same time, all other Continental Germanic languages—including German, the closest relative of Old High German—are assigned only the third synonym (*klein*). This phenomenon is comparable to incomplete lineage sorting in genomics. `contacTrees` could mistake parallel innovations or retentions as evidence for borrowing. However, parallel semantic change can also be a contact effect, when an innovation through semantic shift is reinforced through contact. Languages may innovate or preserve form-meaning pairs so as to mirror the form-meaning pairs of their neighbours. Such processes are fundamentally based on copying and imitating (calquing) rather than on transfer of concrete material (Johanson, 1992). This can be seen in the parallel shift from 'straight, good' to 'opposite of left' in French (*droite*) and English (*right*). In the case of semantic borrowings only the meaning is borrowed and placed on an existing word. German *Maus*, for example, acquired the meaning 'small mobile manual device that controls functions on a computer display' from its English cognate. Such semantic contact effects are reflected in the inferred contact edges.

There are other data coding methods relevant for phylogenetic studies involving language contact, which we discuss in Section S3.

Results

Simulation study. The purpose of the simulation study is two-fold: to demonstrate that the inference works correctly under the `contacTrees` model and to reveal potential errors when using conventional phylogenetic models on data that contains undetected traces of borrowing.

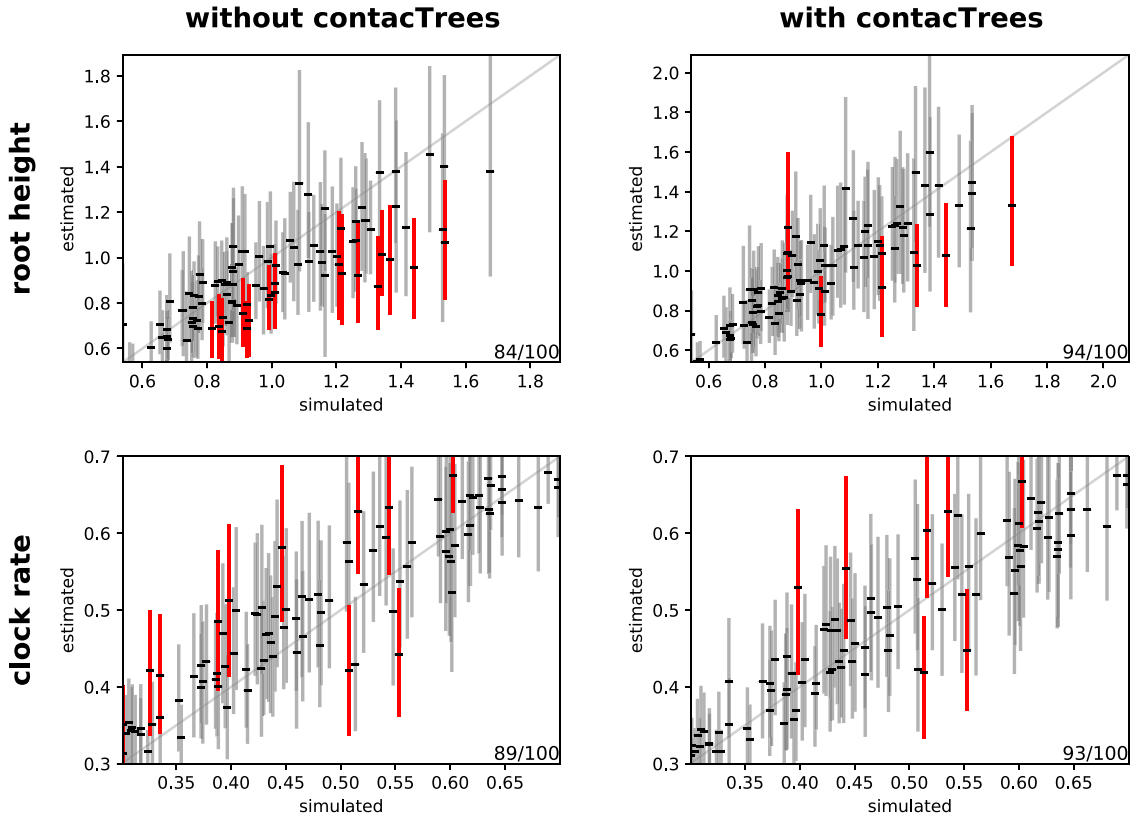

Fig. 3 Simulated values (*x*-axes) and reconstructed credible intervals (*y*-axes) of the root height and clock rate. Reconstructions are based on a conventional phylogenetic model without contact (noCT) (left) or on the `contacTrees` model (CT) (right). The simulations and reconstructions are repeated 100 times, out of which 91–99 should yield a credible interval (grey/red bars) covering the simulated value (diagonal) in a well-calibrated model. Grey bars indicate that the simulated value is within the interval and red that it is outside.

Figure 3 shows the simulated parameters (*x*-axes) and the reconstructed mean values and 95% credible intervals (*y*-axes). In a well-calibrated model, the credible intervals should reflect the true uncertainty about the parameters and contain the true value in 91–99 out of the 100 simulation runs.

When using `contacTrees` for the reconstruction, 94 runs contain the simulated values for the root height, and 93 contain those for the clock rate, which is within the expected range (Fig. 4). Furthermore, the root height and clock rate estimates are unbiased. Figure S4 in the Supplementary Material shows the consistency of `contacTrees`-specific parameters: the number of contact edges, the number of borrowed words, the expected contact edges (Γ) and the borrowing probability (β). Without `contacTrees`, only 84 runs contain the simulated values for root height, and 89 contain those for the clock rate, which is outside the expected range. The errors are biased towards lower root heights and higher clock rates (Fig. 4).

Finally, we also compared the potential for errors in the tree topology. Figure 4 shows the distribution of the RNNI distance (Collienne and Gavryushkin, 2021) to the simulated tree. In both scenarios, some errors are possible, but without `contacTrees` the expected distance is 2.35, while with `contacTrees` it reduces to 1.64. Of course, the exact results of this comparison depend on the simulation settings. The errors are amplified when increasing the expected values for Γ and β. However, we aimed for a plausible setting, and we would not expect severe topological errors as a consequence of borrowing as also discussed in Bowern (2018) and Greenhill et al. (2009).

Case study. In our analysis of the Celtic, Germanic and Romance languages, we compare a conventional phylogenetic reconstruction to a reconstruction with `contacTrees`. We are mainly interested in three questions: (1) Does the use of `contacTrees` affect the reconstructed language tree and model parameters? (2) Does the data support the use of `contacTrees`? (3) Does `contacTrees` recover known or plausible contact events and loanwords?

When comparing the resulting language trees, we see that their topologies mostly agree (see Supplementary Material, Section S5.5), but the CT tree is significantly younger. The posterior distribution of the tree height (Fig. 5b) and the clock rate (Fig. 5c) are clearly lower according to the CT model, which is not surprising. In a conventional phylogenetic model, borrowings have to be explained by parallel innovations of words, which can either be reflected in more changes per time unit, and thus a higher clock rate, or in longer time to explain the changes, thus a higher tree. Which of the two is reflected in the posterior distribution depends on the calibrations. Specifically, it depends on whether borrowings similarly affect the parts of the tree below and above calibration points.The parameters of the model suggest that without CT, the covarion needs to give room for more variable substitution rates, with $\alpha \approx 0.044$ for the CT model and a much lower $\alpha \approx 0.006$ for the noCT model. Furthermore, a higher switch rate washes out the difference between hot and cold states in the CT model.

We computed a Bayes factor (BF) to confirm that modelling contact improves the model fit and better explains the data X. The BF is the ratio between the marginal likelihood of two models or hypotheses. Specifically, we want to compare the marginal likelihood of a model without contact ($|\mathcal{C}| = 0$) and one with contact ($|\mathcal{C}| > 0$). Since both hypotheses are part of the same parameter space explored by the MCMC in CT, we

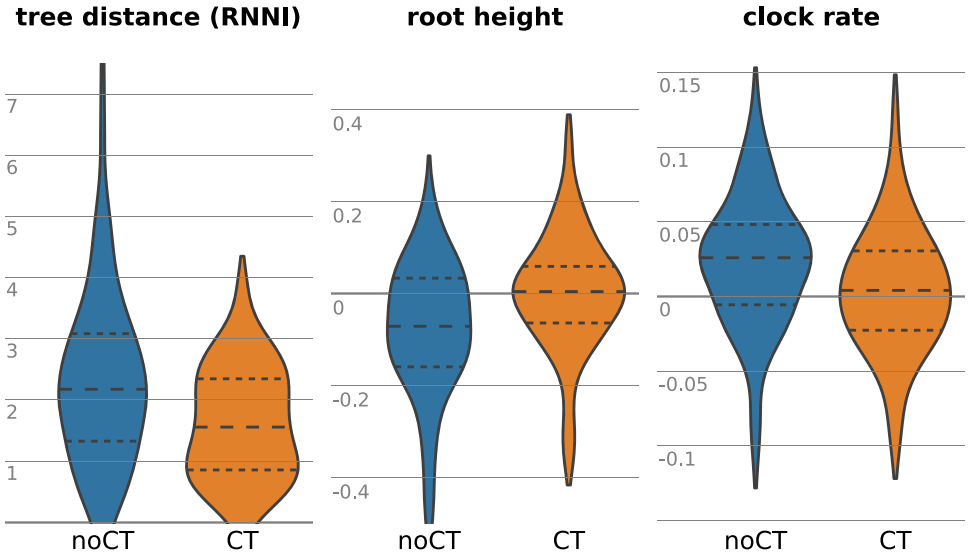

Fig. 4 The error distribution for the tree topology, root height and clock rate, respectively, in the simulation study. For a conventional phylogenetic model (noCT, blue), the error is larger than for the `contacTrees` model (CT, orange).

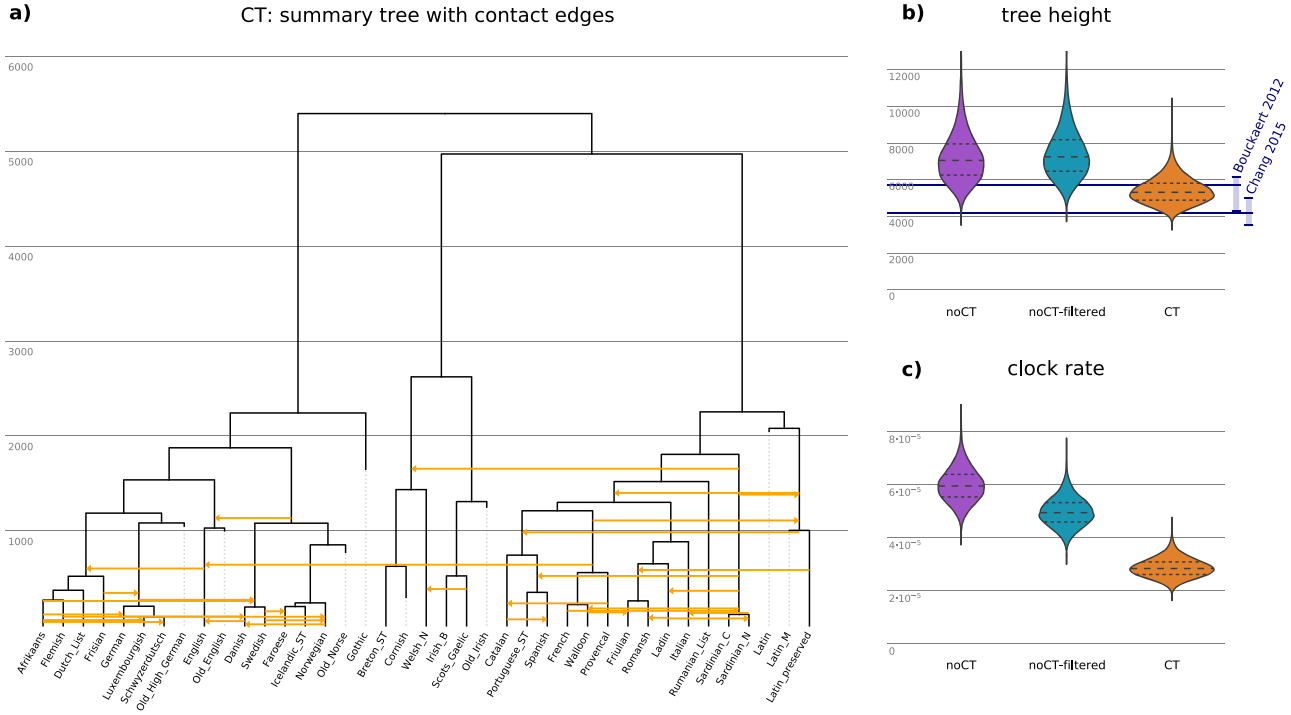

Fig. 5 Reconstructed tree, contact edges and parameters from the case study on the Celtic-Germanic-Romance (CGR) languages. a The reconstructed language tree and contact edges for the CGR languages. **b** The posterior distributions of the root height for the noCT, noCT-filtered, and CT models (mean height of 7206.2, 7449.8, and 5384.0 years, correspondingly). For comparison, we added the reconstructed root age of CGR as estimated by Bouckaert et al. (2012) and Chang et al. (2015) (mean and 95% credible intervals). **c** The posterior distribution of the clock rates (see Supplementary Material, Section S5.2) with a mean of 5.97×10^{-5} (noCT), 4.96×10^{-5} (noCT-filtered), and 2.84×10^{-5} (CT) substitutions per year.

can estimate the BF by counting the samples in each hypothesis:

$$K_{|\mathcal{C}|>0} = \frac{\mathrm{P}(X \mid |\mathcal{C}|>0)}{\mathrm{P}(X \mid |\mathcal{C}|=0)} = \frac{\mathrm{P}(|\mathcal{C}|>0 \mid X)\mathrm{P}(|\mathcal{C}|=0)}{\mathrm{P}(|\mathcal{C}|=0 \mid X)\mathrm{P}(|\mathcal{C}|>0)} \quad (7)$$

We can estimate all terms in Eq. (7) from samples of the prior and the posterior distribution in the CT model. We obtain an infinite Bayes factor when using the raw counts as a maximum-likelihood estimate of the probabilities since all posterior samples contained at least one contact edge. Using a more conservative Bayesian estimate with a uniform prior for the counts (also known as Laplace correction) yields a BF of $K_{|\mathcal{C}|>0} = \frac{300+1}{0+1} / \frac{66.36}{233.64} = 1056.24$, which indicates decisive support for CT.

The Bayes factor confirms that `contacTrees` statistically fits the IE-data better than an otherwise identical model without contact. Next, we estimated whether `contacTrees` correctly identifies contact events and loanwords that reflect the true

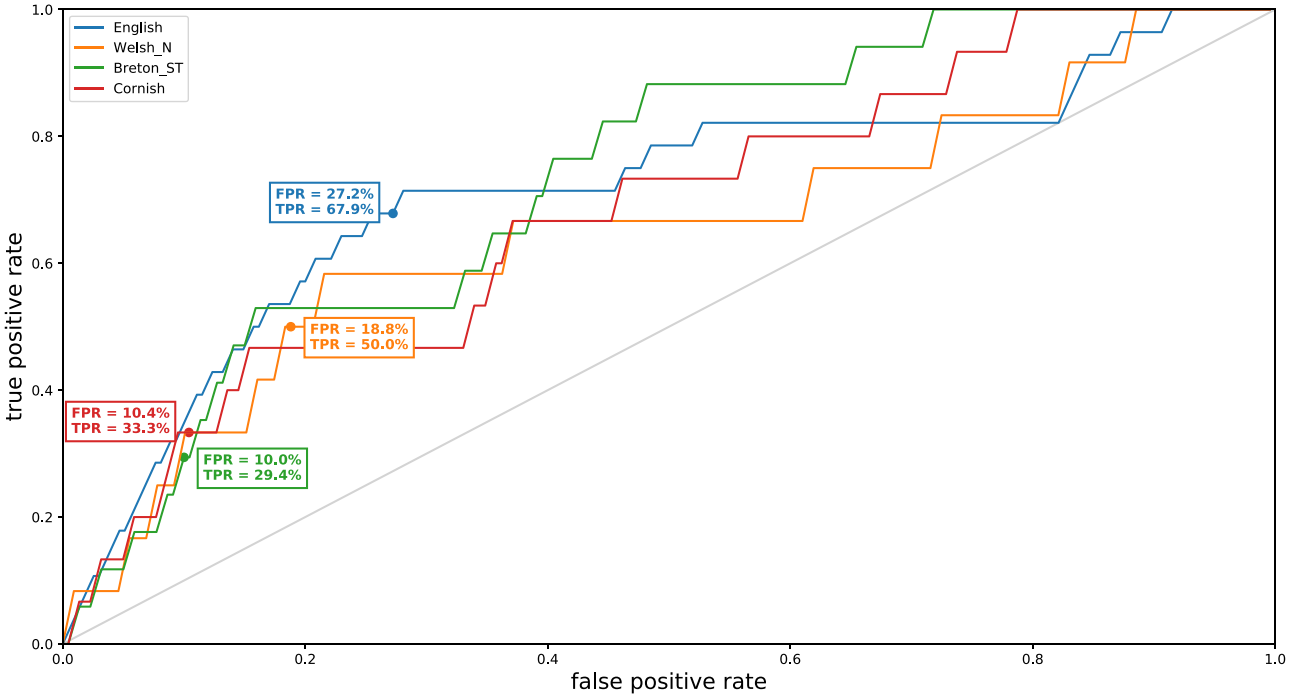

Fig. 6 The receiver operating characteristic (ROC) curve comparing known and predicted loanwords. The curve shows all possible trade-offs between the true positive rate (TPR) and the false positive rate (FPR) under varying posterior probability thresholds. The labelled points are the FPR and TPR for a posterior probability threshold of 0.33.

history of these languages. While there is no exact ground truth for loanwords, we can use the words marked as loans in the data set to evaluate the model's ability to predict borrowings. Unfortunately, this reference is likely incomplete, as we will discuss later. For a given threshold $\bar{k}_{\text{loan}} \in [0, 1]$, we mark a loanword as predicted in a language if in at least $\bar{k}_{\text{loan}}$ of the posterior samples, any of its ancestor branches borrowed it. Varying this threshold, we can achieve different trade-offs between the true-positive rate (TPR) and the false-positive rate (FPR). Figure 6 shows the receiver operating characteristic (ROC) curve for the obtained TPR and FPR values. For example, marking every word as borrowed if it appears in a third of the posterior samples would correctly recover 67.9% of the known loanwords in the English data. At the same time, only 27.2% of the non-loanwords (including words that are actually loans, but not marked as such) would be marked as loans (blue label in Fig. 6). In the Supplementary Material, Section S5.6, we included a list of the most significant false positives (posterior probability of >80% to be a loanword, but not marked as such in the data set) and false negatives (posterior probability of <20% to be a loanword, but marked as one in the data set). The inferred contact events correspond to multiple borrowings between two languages. There is no comparable reference to evaluate them, but we will discuss each contact event and its historical plausibility in the next section and the Supplementary Material.

The `contacTrees` reconstruction of the Celtic, Germanic and Romance languages shows 32 contact edges. Figure 5a gives an overview of all the edges in the tree. Each edge is visualised with its corresponding loanwords in the Supplementary Material 2. Among the 32 total edges, only two connect different top-level clades (Gallo-Romance > English, Romance > Brittonic). Within clades there are 14 contact edges in the Germanic, 15 in the Romance and 1 in the Celtic branch. The contact edge in the Celtic branch shows the influence of Scots Gaelic on Welsh. The Germanic contact edges fall into three categories. Two edges represent the influence of Norse speakers on the English language and one shows an English influence on the Low Franconian languages. Eleven edges connect proximate languages within continental Germanic. In the Romance clade, nine edges indicate a broad influence of Latin and Sardinian on other Romance languages. Most of the remaining contact edges fall between closely related and geographically proximate languages (e.g. Ladin > Romansh, Provencal > Catalan). Exceptions are the edges Walloon > Friulian, French > Sardinian and Ladin > Sardinian_N. We discuss the historical plausibility of these edges and explain what the reconstructions are based on in the section "Reconstructed contact events".

Discussion

In Section "Methods and data" we introduced `contacTrees`, a new model to incorporate language contact and borrowing in linguistic phylogenies. We provide an implementation of the model in BEAST 2 (Bouckaert et al., 2019) (https://github.com/NicoNeureiter/contacTrees). Based on the simulation study, we argue that the model can faithfully reconstruct simulated borrowing events and resolve biases arising from borrowing in the phylogenetic reconstruction. In the case study, we demonstrated that (1) contact edges improve the statistical fit of the data and (2) the reconstructed contact edges and borrowings include known contact events and loanwords within the Indo-European languages. We discuss these findings, offer additional reflections on the suitability of the data set, and give an outlook on future research in the remainder of this section.

Phylogenetic reconstructions with `contacTrees`. In the phylogenetic reconstruction of the Celtic, Germanic, and Romance (CGR) languages, we see that the inclusion of contact in the model significantly impacts the reconstruction. The reconstructed tree is younger, has a lower clock rate (see Fig. 5) and the

inferred frequency of the 'present' state is much lower. This means that the effective transition rate from 0 to 1 is very low and, hence, the character evolution is closer to a Dollo process. This confirms the intuition that borrowing is one type of violation of the Dollo assumption (i.e. a source of multiple innovations). The lower clock rate is expected as well, since borrowed words have to be explained by additional changes in a strict tree model. The simulation study shows that undetected borrowings can affect the tree height when using a conventional phylogenetic model. However, in the simulation, the bias was towards *younger* trees, while in the IE case study, conventional models return *older* trees. In the Supplementary Material, Section S4.2, we explain that the direction of bias depends on the relative position of contact edges to calibration points on the tree. In that sense, the reconstruction of the CGR phylogeny without `contacTrees` overestimated the root height due to undetected loanwords. Using `contacTrees`, we can reduce this bias. A similar effect was reported in Kelly (2016), where a stochastic Dollo model with lateral transfer was applied on the CGR languages, and the authors observed a similar reduction in the tree height. Previous phylogenetic estimates of the age of the CGR clades range from about 4200 (Chang et al., 2015) to about 5700 (Bouckaert et al., 2012). In contrast to our analyses, these estimates benefit from the larger context of the whole Indo-European language family to inform the height of the CGR root. While the two estimates correspond to very different hypotheses on the expansion of the Indo-European languages, both of them are more compatible with the estimate of 5384.0 years before present obtained in the CT analysis, as opposed to the 7206.2 years before present obtained in the noCT analysis (Fig. 5). The age of the Indo-European languages is highly debated. It would be interesting to apply `contacTrees` in a case study on all Indo-European languages to see how our results for the CGR languages translate to the whole language family.

Unsurprisingly, the clock rate is significantly lower in the CT reconstruction compared to noCT and noCT-filtered. In a strict tree model, loanwords need to be explained by multiple parallel innovations and potentially a loss of the previously used word, driving up the clock rate. In the noCT-filtered analysis, removing known borrowings reduces this effect but likely misses some loanwords or parallel semantic shifts, as discussed below. A lower rate of change and more consistent absence–presence patterns across the word trees lead to a higher likelihood. This higher likelihood comes at the cost of additional model parameters, the contact edges and borrowing indicator variables, bearing the danger of overfitting. However, the Bayes factor (1056.24) is still significantly higher when contact is permitted, indicating overwhelming support for the CT hypothesis. The BF considers the marginal likelihood, which integrates over all possible parameter values, confirming that overfitting is not an issue.

Reconstructed contact events. The CT reconstruction proposes 32 contact edges, all of which are listed and visualised in the Supplementary Material, Section S5.7. Some edges clearly represent important historical examples of language contact: Latin influences on the Brittonic languages after the invasion of the Roman Empire, Norse influences on English due to Viking settlers in Great Britain or Norman influences on English after the invasion of William the Conqueror. These major contact events across clades are rare. Most reconstructed contact edges represent contact between closely related languages: Influences of Swedish on Norwegian, Catalan on Spanish, German on Danish and many more. The cross-clade contact events are well supported by known loanwords, while contact between closely related languages is more nuanced.

The two examples shown in Fig. 7 are the only contact events between two of the three top-level clades. In both cases Romance languages entered the British isles. We use these well-studied contact events as a reference for discussing the details `contacTrees` is able to infer under comparable conditions. The fact that we find major contact events in the British isles is no coincidence. The British isles are a known area of close language contact involving Celtic, Germanic and Romance languages (Dedio et al., 2019). The remaining 30 edges represent contact between more closely related languages. There are 14 contact edges in the Germanic, 15 in the Romance, and 1 in the Celtic branch. We will first focus on the two cross-clade edges and then discuss several examples of contact between closely related languages.

In the 11th century CE, William the Conqueror, then Duke of Normandy, invaded England and was crowned king. Norman French was introduced as the language of the elites, causing many French words to be introduced into English Black (2017). Some of these loanwords also found their way into the core vocabulary: in the IElex data set 11 English words were marked as loans from Norman French. The contact edge in Fig. 7a represents the borrowings that resulted from the Norman Conquest. Medieval Norman French not being included in IElex, the edge identifies its closest relative, an ancestor of French, as the donor language. However, regarding the timing of the contact event, there is uncertainty. The bulk of the probability mass is accurately placed around 600–1000 years ago, but there is also minor support (about 9%) for a more recent date. We attribute this to words whose borrowed form was lost or not included in Provencal or Walloon (e.g. *push*, *turn* and *round*), making a more recent edge appear plausible to the model.

The seven reconstructed loanwords with the highest support at this edge (Fig. 7a) are mostly in line with previous knowledge. The words *lake*, *animal*, *count*, *round*, *vomit* and *fruit* are all known borrowings from the Gallo-Romance cultural sphere. The only false positive is *right* (RIGHTSIDE), a case of parallel innovation. In spite of *right* belonging to the same cognate class as the identified source word, `contacTrees` reconstructs a borrowing event because two ancient varieties (Old English, the closest relative, and Old High German) feature items from other cognate classes. As a consequence, `contacTrees` assumes that the ancestor of *right* was replaced in the prehistory of English and infers its reintroduction by lateral transfer.

The Roman Empire expanded over a vast area, including most of Europe. From 43 CE on, the Romans conquered parts of Great Britain, where they ruled until around 410 CE. Over this period, a Romano-British culture developed, influencing the culture and language of the local Celtic people, introducing many Latin loanwords to the Brittonic languages. In the IELex data, 10 Breton, 10 Cornish and 8 Welsh words are marked as borrowings from Latin. Our reconstruction shows a contact edge from Romance languages to Proto-Brittonic around 350 CE (Fig. 7b). There is uncertainty regarding the exact placement of the donor language, spread between all ancestors of the current Romance languages at that time. The ancestor of Sardinian is the most likely donor (posterior support of 0.26), while Latin does not receive significant support. We attribute this to some shared Brittonic and Romance cognates missing from Latin. For example, Welsh_ST and Sardinian_N share the cognates *mam*/*mamma* for MOTHER and *ffrwyth*/*frutta* for FRUIT. In both cases, the Latin lexicon contains retentions of these cognates —*mamma* and *fructus*—which where omitted in the IELex data in favour of *mater* and *pomum*, two more salient words in the corresponding meaning class. While *ffrwyth* is correctly classified as a loanword, *mam* is a parallel innovation, a well-known phenomenon for nursery words and onomatopoeia. A closer look

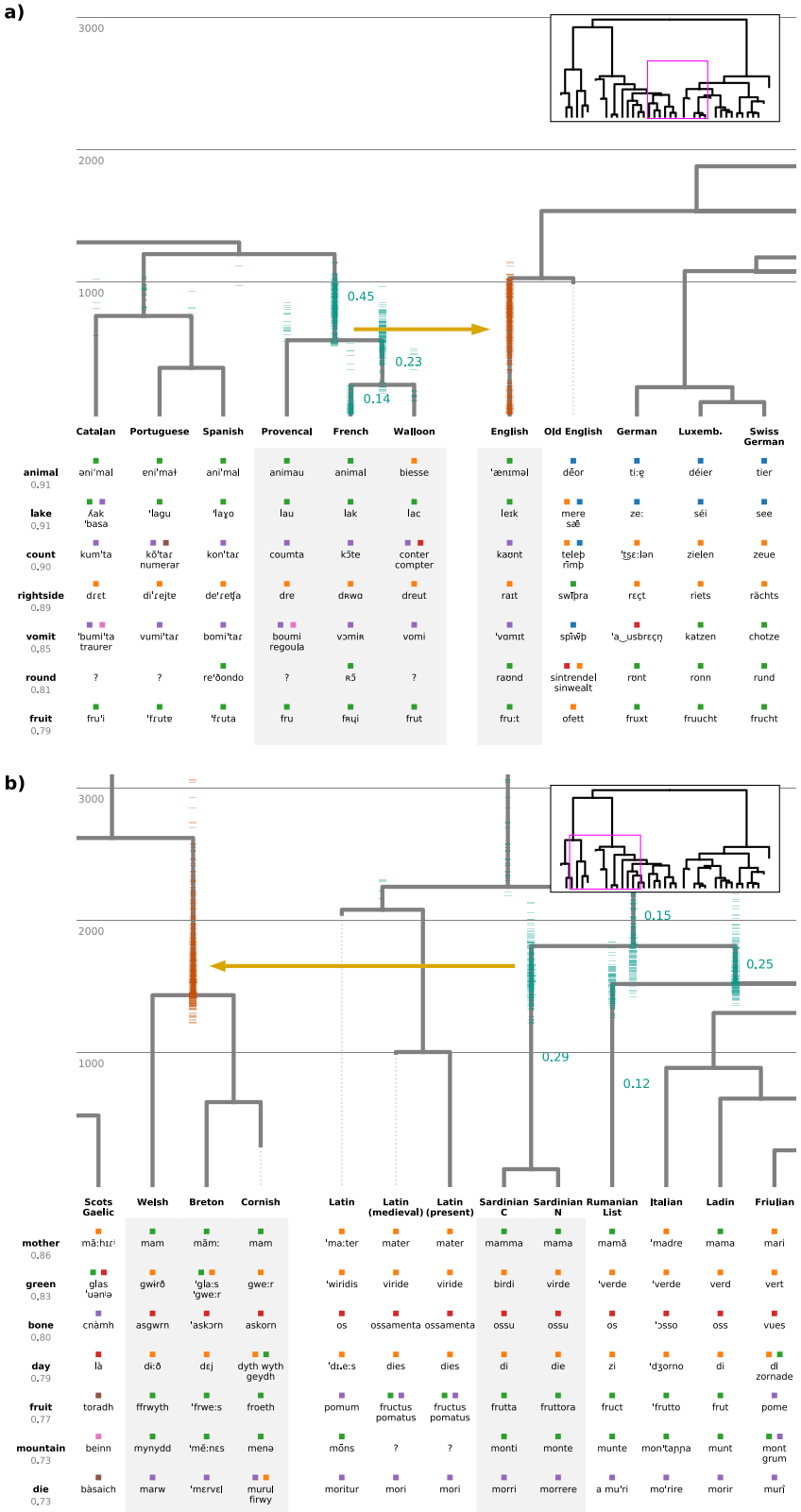

Fig. 7 The two reconstructed contact events connecting different top-level clades. a A contact edge from Old French to Middle English about 640 years before present. **b** A contact edge from ancestral Romance to Proto-Brittonic about 1670 years before present. Green and red marks represent the possible placements of the donor and receiver language, respectively. The word list shows the most likely loanwords at these contact events. The row labels show the meaning class and the posterior support. Each cell lists the lexemes, colorcoding the cognates.

at the remaining reconstructed loanwords for the Romance > Proto-Brittonic contact edge shows a similar picture. Some are marked loanwords, like the Proto-Brittonic *gwirð* (GREEN) which was borrowed from Vulgar Latin *virdis*. Others are likely inherited, but still tell a story that contact effects and parallel innovations can explain. The Proto-Indo-European word *$h_3éh_1os$* (BONE) was retained in Romance and Brittonic (e.g. French *os* and Breton *as-corn*), but was replaced in the much more isolated Goidelic branch of the Celtic languages (e.g. by *cnáim* in Old Irish), including Irish and Scots Gaelic. Persistent contact with Romance may have reinforced the use of the cognate in the one branch while the more isolated Goidelic languages were more readily replacing it.

Given the well-documented persistent interactions between Germanic and Celtic populations since the Middle Ages (Black, 2017; Jackson, 1953), we expect `contacTrees` to identify a cross-clade edge, which is not the case. We can but speculate about possible reasons. Germanic-Celtic loans are absent from the IELex and apparently there is not sufficient support from cases of shared retention and parallel innovation either. This leads us to suspect that not every kind of contact leaves the same traces in form-meaning data sets.

We now turn to the edges within each clade, which we visualise in the Supplementary Material, Section S5.7. Two of the Germanic edges represent the known historical influence of Norse speakers on the English language, including multiple known loanwords. The older of the two edges very accurately fits the Viking settlements in northern and eastern England around 1000 years before present, which were the source of most known Norse loanwords in English. The younger edge (Danish > English) seems to be an artefact due to Norse loanwords that are not coded in Old English. Eleven edges broadly outline a contact area within the continental Germanic languages. In these cases, almost all support comes from shared inheritance, parallel semantic shift with and without contact and cascading loans, e.g. the Romance successor of Latin *rotundus* 'round' spread to most Germanic languages. One edge indicates an English influence on Low Franconian and Frisian. Here, too, the relevant support comes from shared inheritance, (*hold*), parallel shift (*right*) and multiple loans (*river*). The contact event in the Celtic branch shows the influence of Goidelic on Welsh. While Intra-Celtic contact is well known to have happened at least since the early Middle Ages (Bauer, 2015) and contact between Goidelic and Welsh fits historical facts, the intra-Celtic edge has support primarily from shared inheritance (EAR) in a common socio-economic and cultural context, conservative coding and multiple loans (PERSON).

In the Romance clade, we find edges that match known or likely language contact. They connect Latin to other Romance languages due to its continued use in the Catholic church, or very closely related languages (e.g. Ladin > Romansh) and geographically proximate ones (e.g. Provencal > Catalan). However, five of the Romance contact edges have Sardinian as a donor language. Most of the loanwords reconstructed at this edge represent shared inheritance. We suspect that Sardinian—known to be a very conservative language—serves as a proxy for archaic word forms. This shows how the model accommodates for the specific data coding procedures (e.g. avoiding synonyms, selecting archaic forms) as well as processes not explicitly modelled, such as parallel semantic shifts.

To summarise, the `contacTrees` model detects parallel developments in two branches of the tree. Parallel innovations are especially informative, but the model could also infer contact based on surprisingly many shared retentions (as suggested in the case of Romance > Brittonic). Parallel innovations can either occur as contact effects or independent parallel innovations. Independent parallel innovations are not infrequent in form-meaning traits. Especially when a language contains multiple forms for a meaning (synonyms), the most common form for this meaning can change easily or coding decisions can vary. This can explain cases where the same form is dominant in two distinct languages by coincidence. But independent parallel innovations can occur more systematically due to semantic or derivational drift, where the meaning of a word changes into a different, but semantically related meaning. It is not uncommon that a word in two related languages undergoes the same semantic changes in this way, for example in common metaphors (SEE > KNOW, etc.). Contact can lead to parallel innovations in multiple ways. Apart from typical loanwords we might detect more subtle contact effects like semantic loans and latent contact effects. Furthermore, when a word is widely borrowed, it might be introduced independently in two languages, due to a borrowing from a third language.

While independent parallel innovations can be falsely classified as borrowings, we would not expect many such coincidental changes to occur in the same languages at the same time. Evidence for a contact edge is based on multiple parallel innovations, which are more plausibly explained by actual contact effects.

Reconstructed contact effects. We have compared the reconstructed borrowings to the annotated loanwords in IELex and shown the resulting true positive and false negative rates in ROC curves in Fig. 6. The curves are all clearly above the diagonal, which indicates that `contacTrees` can detect known loanwords better than chance.

However, loanwords are only one possible form of borrowing. The type of borrowings that can be detected by `contacTrees` depends on the traits used. Since form-meaning traits are coded by meaning classes, the reconstructed borrowings can reflect borrowed meanings. This can take the form of a semantic loan or more subtle effects. E.g. the usage of a shared cognate could be reinforced through contact, turning it into the dominant form for a meaning (which is what is eventually represented in the data). These semantic contact effects are valid and expected results, but they are difficult to detect and consequently not listed as loanwords in IELex (or loanword databases like WOLD) and, hence, are hard to verify.

Due to the limited information on each word (only absence/presence) provided by form-meaning traits, the model needs to infer contact from a broader pattern across multiple words. For example, a single English word that differs from its Germanic relatives, but shares cognates in Gallo-Romance, could be ascribed to coincidental parallel innovation rather than contact. However, the convergent evidence from multiple Gallo-Romance cognates in the English vocabulary can still lead to a convincing proposal of a contact event. As a positive side effect, these statistical reconstructions make it possible to detect latent contact effects. When we observe a surprising similarity between two branches, it seems sensible to consider contact effects as an explanation, even without concrete evidence for a specific loanword. For example, the use of a certain form for a meaning could be gradually reinforced under contact with a language that uses a cognate for the same meaning. This is the case in the Goidelic-Welsh edge. In typology, this distributional view on contact effects is more common Bickel (2015), Dedio et al. (2019), Ranacher et al. (2021). For example, the Balkan languages share many structural features, likely as a result of contact. However, the evidence for contact does not rely on the exact reconstruction of a specific borrowed feature, but rather on the statistical similarity across a range of features. The contact events inferred by `contacTrees` are based on the same principle, but add a diachronic model to explain these statistical patterns.

Linguistic phylogenetic analyses are not always based on form meaning traits. The reconstructed contact effects may be different when using different linguistic traits. In the Supplementary Material, Section S3 we describe the possible contact effects for *etymon traits*, *phonemic transcription traits* or *typological traits*. Any of these data types would yield different types of borrowings and requires different efforts in data collection and suitable models of evolution. Form-meaning traits are usually more readily available and established for phylogenetic reconstructions.

Future work. The `contacTrees` model extends existing phylogenetic models to allow horizontal transfer between languages in a very general way. The implementation is compatible with many phylogenetic models available as BEAST 2 packages. This makes a range of future applications possible. The case study shows how `contacTrees` can be used to reconstruct dated phylogenies. The relevance of this is more striking when considering less studied language families, where the knowledge of past interactions and resulting loanwords is limited. Furthermore, the use of `contacTrees` for tree building becomes more relevant for longer word lists, since words outside the core vocabulary are thought to be more prone to borrowing (Pagel et al., 2007).

Another outcome of the case study are the reconstructed contact events and loanwords, which are of interest themselves. As discussed, the type of borrowings that will be reconstructed and the reliability of the reconstruction depends on the type of data used in the analysis. While form-meaning traits provide the potential to infer semantic loans and latent contact effects, etymon traits or phonemic transcriptions might be more reliable at detecting loanwords. We hope for future studies to evaluate the quality of the reconstructed contact events and borrowings for etymon traits, phonemic transcriptions and typological data.

The inferred contact edges complement the language tree and together they provide a more complete picture of the history of the languages, which can be a benefit to downstream applications. For example, ancestral state reconstruction has been used to infer states of linguistic traits in ancestral languages, like grammatical features (Carling and Cathcart, 2021) or the size of the phonological inventory (Moran et al., 2021). Assuming that these traits can be borrowed in the same way as words, a comparative phylogenetic study should allow for contact. The contact events inferred from lexical data could be informative for the comparative study of other features, too. The applications even extend beyond linguistics: contact between languages can be indicative of contact between cultures in general. A broad range of studies on cultural evolution (e.g. on political complexity (Currie et al., 2010) or marital residence patterns (Fortunato and Jordan, 2010)) already uses linguistic phylogenies and could benefit from linguistic contact edges to incorporate potential contact effects.

Finally, further development could go into different priors for the `contacTrees` model. In particular, assumptions about which languages are expected to be in contact, could be cast into a contact prior. Since language contact must involve contact between speakers, we would expect more contact between geographically proximate languages. Using phylogeography to estimate historical locations, it would be possible to define a geographically informed contact prior, i.e. a prior where the rate of contact decreases with geographic distance. Stolz et al. (2021) followed a similar idea and recently introduced a structured coalescent model for viral reassortment networks. In a similar way closely related languages are more likely to be in contact. This would motivate a contact prior that decreases with phylogenetic distance. In the results of our case study, we can already observe a tendency of related and proximate languages to be in contact. However, this tendency is arising purely from the data. Multiple contact edges with Afrikaans as a donor language demonstrate that the data cannot always exclude contact that is geographically implausible. Here, a geographically informed contact prior would improve the reconstruction. A different extension of the model could allow the borrowing probability (β) to vary between features. This will be crucial when mixing different types of data—e.g. lexical and typological features—in a single analysis. But even within the lexicon, there are multiple hypotheses about varying borrowability (Pagel et al., 2007). A model which allows β to vary between features would be a suitable tool to test such hypotheses. The modular implementation, the compatibility with existing BEAST 2 packages and the open source availability make `contacTrees` readily extendable in these directions.

Data availability

The data used for the case study is available at https://github.com/NicoNeureiter/contacTrees-IndoEuropean along with instructions for how to generate BEAST XML files and how to run the analysis (archived at https://doi.org/10.5281/zenodo.6563028). The implementation of `contacTrees` is available at https://github.com/NicoNeureiter/contacTrees including instructions for installation and usage (archived at https://doi.org/10.5281/zenodo.6563025).

Received: 14 January 2022; Accepted: 23 May 2022;
Published online: 17 June 2022

References

Atkinson QD, Gray RD (2005) Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. Syst Biol 54(4):513–526

Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M (2008) Languages evolve in punctuational bursts. Science 319(5863):588 https://doi.org/10.1126/science.1149683

Bateson G (1935) Culture contact and schismogenesis. Man 35, 199 (178–183) https://doi.org/10.2307/2789408

Bauer B (2015) Intra-Celtic loanwords, Ph.D. thesis, Wien, A, Universität, Wien

Bickel B (2015) Distributional typology: statistical inquiries into the dynamics of linguistic diversity. In: Heine B, Narrog H (eds) The Oxford handbook of linguistic analysis, 2nd edn. Oxford University Press, Oxford, pp. 901–923

Black J (2017) A history of the British Isles, 4th edn. Palgrave, London & New York

Bouckaert RR, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N et al. (2019) BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 15(4):e1006650

Bouckaert RR (2019) Babel: BEAST analysis backing effective linguistics https://github.com/rbouckaert/Babel

Bouckaert RR, Bowern C, Atkinson QD (2018) The origin and expansion of Pama-Nyungan languages across Australia. Nat Ecol Evol 2(4):741–749

Bouckaert RR, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD (2012) Mapping the origins and expansion of the Indo-European language family. Science 337(6097):957–960

Bouckaert RR, Robbeets M (2017) Pseudo dollo models for the evolution of binary characters along a tree, bioRxiv https://doi.org/10.1101/207571

Bowern C (2018) Computational phylogenetics. Annu Rev Linguist 4:281–296

Bryant D, Moulton V (2002) NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. In: Guigó R, Gusfield D (eds) International workshop on algorithms in bioinformatics. Springer, Belin, Heidelberg, pp. 375–391

Carling G, Cathcart C (2021) Reconstructing the evolution of Indo-European grammar. Language 97(3), https://doi.org/10.1353/lan.0.0253

Chang W, Hall D, Cathcart C, Garrett A (2015) Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. Language 91(1):194–244

Chousou-Polydouri N, Birchall J, Meira S, O'Hagan Z, Michael L (2016) A test of coding procedures for lexical data with Tupí-Guaraní and Chapacuran

languages. In: Bentz C, Jäger G, Yanovich I (eds) Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics. Philosophische Fakultät, Tübingen
Collienne L, Gavryushkin A (2021) Computing nearest neighbour interchange distances between ranked phylogenetic trees. J Math Biol 82(1):1–19
Cook SR, Gelman A, Rubin DB (2006) Validation of software for Bayesian models using posterior quantiles. J Comput Graph Stat 15(3):675–692
Currie TE, Greenhill SJ, Gray RD, Hasegawa T, Mace R (2010) Rise and fall of political complexity in island South-East Asia and the Pacific. Nature 467(7317):801–804
Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. PNAS 104(3):870–875
Dedio S, Ranacher P, Widmer P (2019) Evidence for Britain and Ireland as a linguistic area. Language 95(3):498–522
Dellert J (2019) Information-theoretic causal inference of lexical flow. In: Wieling M, D'Arcy A (eds) Language variation 4. Language Science Press, Berlin
Didelot X, Lawson D, Darling A, Falush D (2010) Inference of homologous recombination in bacteria using whole-genome sequences. Genetics 186(4):1435–1449
Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4(5):e88
Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161(3):1307–1320
Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22(5):1185–1192
Dunn M (2012) Indo-European lexical cognacy database (IELex). Max Planck Institute for Psycholinguistics, Nijmegen
Fortunato L, Jordan F (2010) Your place or mine? A phylogenetic comparative analysis of marital residence in Indo-European and Austronesian societies. Philos Trans R Soc B: Biol Sci 365(1559):3913–3922
François A (2015) Trees, waves and linkages. In: Bower C, Evans B (eds) The Routledge handbook of historical linguistics. Routledge, London, pp 161–189
Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426(6965):435–439
Gray RD, Bryant D, Greenhill SJ (2010) On the shape and fabric of human history. Philos Trans R Soc B: Biol Sci 365(1559):3923–3933
Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323(5913):479–483
Greenhill SJ, Currie TE, Gray RD (2009) Does horizontal transmission invalidate cultural phylogenies? Proc R Soc B 276(1665):2299–2306
Greenhill SJ, Wu CH, Hua X, Dunn M, Levinson SC, Gray RD(2017) Evolutionary dynamics of language systems Proc Natl Acad Sci USA 114(42):E8822–E8829
Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M (2015) Bantu expansion shows that habitat alters the route and pace of human dispersals. Proc Natl Acad Sci USA 112(43):13296–13301
Grossman E, Eisen E, Nikolaev D, Moran S (2020) SegBo: a database of borrowed sounds in the world's languages. In: Proceedings of the 12th language resources and evaluation conference. European Language Resources Association, Marseille, France, pp. 5316–5322
Heled J, Drummond AJ (2009) Bayesian inference of species trees from multilocus data. Mol Biol Evol 27(3):570–580
Holland BR, Huber KT, Dress A, Moulton V (2002) δ plots: a tool for analyzing phylogenetic distance data. Mol Biol Evol 19(12):2051–2059
Hruschka DJ, Branford S, Smith ED, Wilkins J, Meade A, Pagel M, Bhattacharya T (2015) Detecting regular sound changes in linguistics as events of concerted evolution. Curr Biol 25(1):1–9
Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23(2):254–267
Jackson KH (1953) Language and history in early Britain. University Press, Edinburgh
Jacques G, List JM (2019) Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them). J Hist Linguist 9(1):128–167
Johanson L (1992) Strukturelle Faktoren in türkischen Sprachkontakten. Steiner, Stuttgart
Kaiping GA, Klamer M (2022) The dialect chain of the Timor–Alor–Pantar language family. Lang Dyn Change https://doi.org/10.1163/22105832-bja10019
Kelly L (2016) A stochastic Dollo model for lateral transfer. Ph.D. thesis, University of Oxford
Kelly LJ, Nicholls GK (2017) Lateral transfer in stochastic Dollo models. Ann Appl Stat 11(2):1146–1168
Kolipakam V, Jordan FM, Dunn M, Greenhill SJ, Bouckaert RR, Gray RD, Verkerk A (2018) A Bayesian phylogenetic study of the Dravidian language family. R Soc Open Sci 5(3):171504
Maurits L, de Heer M, Dunn M, Vesakoski O (2019) Using contact linguistics for relative calibration of phylogenies. In: International Conference on Historical Linguistics 24, Canberra, Australia
Moran S, Grossman E, Verkerk A (2021) Investigating diachronic trends in phonological inventories using bdproto. Lang Resour Eval 55(1):79–103
Muysken P (2011) Three processes of borrowing: borrowability revisited. De Gruyter Mouton, pp. 229–246
Nakhleh L, Ringe D, Warnow T (2005) Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. Language 81(2):382–420
Nelson-Sathi S, List JM, Geisler H, Fangerau H, Gray RD, Martin W, Dagan T (2011) Networks uncover hidden lexical borrowing in Indo-European language evolution. Proc R Soc B: Biol Sci 278(1713):1794–1803
Neureiter N, Ranacher P, van Gijn R, Bickel B, Weibel R (2021) Can Bayesian phylogeography reconstruct migrations and expansions in linguistic evolution? R Soc Open Sci 8(1):201079
Pagel M, Atkinson QD, Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature 449(7163):717–720
Rama T (2018) Three tree priors and five datasets: A study of Indo-European phylogenetics. Lang Dyn Change 8(2):182–218
Ranacher P, Neureiter N, van Gijn R, Sonnenhauser B, Escher A, Weibel R, Muysken P, Bickel B (2021) Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact. J R Soc Interface https://doi.org/10.1098/rsif.2020.1031
Ritchie AM, Ho SYW (2019) Influence of the tree prior and sampling scale on Bayesian phylogenetic estimates of the origin times of language families. J Lang Evol 4(2), 108–123, (2021) https://doi.org/10.1093/jole/lzz005
Sagart L, Jacques G, Lai Y, Ryder RJ, Thouzeau V, Greenhill SJ, List JM (2019) Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proc Natl Acad Sci USA 116(21):10317–10322
Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ (2013) Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci USA 110(1):228–233
Stolz U, Stadler T, Müller NF, Vaughan TG (2021) Joint inference of migration and reassortment patterns for viruses with segmented genomes. Mol Biol Evol https://doi.org/10.1093/molbev/msab342
Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. Int J Am Linguist 21(2):121–137
Syrjänen K, Maurits L, Leino U, Honkola T, Rota J, Vesakoski O (2021) Crouching TIGER, hidden structure: exploring the nature of linguistic data using TIGER values. J Lang Evol 6(2):99–118
Tehrani JJ (2020) Descent with Imagination: oral traditions as evolutionary lineages. Springer International Publishing, Cham, pp. 273–289
Thomason SG, Kaufman T (1989) Language contact, creolization and genetic linguistics. University of California Press, Berkeley, Los Angeles & Oxford
Tuffley C, Steel M (1998) Modeling the covarion hypothesis of nucleotide substitution. Math Biosci 147(1):63–91
Vaughan TG, Welch D, Drummond AJ, Biggs PJ, George T, French NP (2017) Inferring ancestral recombination graphs from bacterial genomic data. Genetics 205(2):857–870
Wen D, Yu Y, Nakhleh L (2016) Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS Genet 12(5):e1006006
Widmer M, Auderset S, Nichols J, Widmer P, Bickel B (2017) Np recursion over time. Language 93(4):799–826
Willems M, Lord E, Laforest L, Labelle G, Lapointe FJ, Di Sciullo AM, Makarenkov V (2016) Using hybridization networks to retrace the evolution of Indo-European languages. BMC Evol Biol 16(1):1–18
Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. Genetics 150(1):499–510
Yule GU (1925) II—A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. Philos Trans R Soc Lond Ser B 213(402-410):21–87
Zhang C, Ogilvie HA, Drummond AJ, Stadler T (2018) Bayesian inference of species networks from multilocus sequence data. Mol Biol Evol 35(2):504–517

Acknowledgements

We thank Timothy Vaughan, Jordan Douglas, Alexei Drummond and Natalia Chousou-Polydouri for valuable discussions and feedback. Funding supports for this work were provided by the URPP Language and Space, University of Zurich, the NCCR Evolving Language with Swiss NSF Agreement No. 51NF40_180888, the Swiss NSF Sinergia Project No. CRSII5_183578 (Out of Asia) and the Royal Society of New Zealand Marsden Grant 18-UOA-096.

Author contributions

NN conceived of the idea, implemented the algorithm and carried out the case studies. RRB, PR and GAK contributed to the idea, implementation and case study design. PW and NE interpreted the results from a linguistic perspective. RW coordinated the study. NN led the writing of the manuscript with contributions from all co-authors.

Competing interests
The authors declare no competing interests.

Ethical approval
This research did not require any ethical approval.

Informed consent
This article does not contain any studies with human participants performed by any of the authors.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-022-01211-7.

Correspondence and requests for materials should be addressed to Nico Neureiter.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]University Research Priority Program (URPP) Language and Space, University of Zurich, Zurich, Switzerland. [2]Department of Geography, University of Zurich, Zurich, Switzerland. [3]NCCR Evolving Language, University of Zurich, Zurich, Switzerland. [4]Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich, Zurich, Switzerland. [5]Department of Comparative Language Science, University of Zurich, Zurich, Switzerland. [6]Centre for Computational Evolution, University of Auckland, Auckland, New Zealand. [7]School of Computer Science, University of Auckland, Auckland, New Zealand. ✉email: nico.neureiter@gmail.com