Phylogenetic Models of Language Diversification

Robin Ryder

Phylogenetic Models of Language Diversification

Robin J. Ryder
The Queen's College

Supervisor:
Geoff K. Nicholls

Department of Statistics
University of Oxford, UK

A dissertation submitted in partial fulfillment
of the requirements for the degree of Doctor of Philosophy

April 9, 2010

Abstract

Language diversification is a stochastic process which presents similarities with phylogenetic evolution. Recently, there has been interest in modelling this process to help solve problems which traditional linguistic methods cannot resolve. The problem of estimating and quantifying the uncertainty in the age of the most recent common ancestor of the Indo-European languages is an example.

We model lexical change by a point process on a phylogenetic tree. Our model is specifically tailored to lexical data and in particular treats aspects of linguistic change which are hitherto unaccounted for and which could have a strong impact on age estimates: "catastrophic" rate heterogeneity and missing data. We impose a prior distribution on the tree topology, node ages and other model parameters, give recursions to compute the likelihood and estimate all parameters jointly using Markov Chain Monte Carlo.

We validate our methods using an extensive cross-validation procedure, reconstructing known ages of internal nodes. We make a second validation using synthetic data and show that model misspecifications due to borrowing of lexicon between languages and the presence of meaning categories in lexical data do not lead to systematic bias.

We fit our model to two data sets of Indo-European languages and estimate the age of Proto-Indo-European. Our main analysis gives a 95% highest posterior probability density interval of 7110 – 9750 years Before the Present, in line with the so-called "Anatolian hypothesis" for the expansion of the Indo-European languages. We discuss why we are not concerned by the famous criticisms of statistical methods for historical linguistics leveled by Bergsland and Vogt [1962]. We also apply our methods to the reconstruction of the spread of Swabian dialects and to the detection of "punctuational bursts" of language change in the Indo-European family.

Acknowledgements

This thesis would not have been possible without the help of many people.

Not only did the Life Sciences Interface Doctoral Training Centre provide an excellent preparation to doctoral work, they also provided invaluable help. In particular, Maureen York's commitment to providing assistance when looking for funding was impressive.

The examiners for my viva, my transfer and confirmation of status, Jotun Hein, Ziheng Yang, Gesine Reinert and Gilean McVean, provided me with helpful insights. Quentin Atkinson and Marlies Rother also made valuable comments.

My doctoral years have been some of the most enjoyable of my life thanks to my wonderful family and friends. I am very grateful to my parents for their continued support, and the presence along my side of my fiancée Elsa Paroissien-soon-to-be-Ryder has been invaluable.

I am amazed at how much Dr Geoff Nicholls has managed to teach me in these past three years. I feel privileged to have been supervised by him and I shall very much miss his advice, guidance and conversational wanderings.

Contents

Chapter 1

Introduction and background

1.1 The Indo-European family of languages

Most languages of Europe and several languages of India are members of the same family, called the Indo-European family.

Parsons [1767] discovered similarities in the words for basic numerals between Bengali, Persian, and 15 European languages, and noted that on the other hand, Chinese, Hebrew, Malay and Turkish have very different words for numerals. He came to the conclusion that those 17 languages are related, and all stemmed from a common ancestor, the language of Japhet, son of Noah. Some of these similarities are shown in Table 1.1.

Similarities between these languages can also be found in many other words, and in syntactical and phonetic features. Furthermore, the differences between languages show clear patterns, making it impossible for these similarities to be due to a coincidence. It is now accepted that these languages are related; Fig. 1.1 shows the distribution of the languages in the Indo-European family,

	1	2	3	9
Albanian	një	dy	tre	nëntë
Bengali	ek	dvi	tri	nay
English	one	two	three	nine
Greek	hen	duo	treis	ennea
Irish	aon	do	tri	naoi
Italian	uno	due	tre	nove
Persian	yak	do	se	noh
Russian	odin	dva	tri	devyat
Swedish	en	tva	tre	nio
Tocharian A	sas	wu	tre	nu
Chinese	yi	er	san	jiu
Hebrew	'ehad	s(e)nayim	selosa	tis'a
Turkish	bir	iki	üc	dokuz

Table 1.1: *Numerals in some of the languages in Parsons' sample. The first ten show significant similarities; they are all members of the Indo-European family. For languages which do not use the Latin alphabet, we show an approximate phonetic transcription.*

as it is understood nowadays.

Inside the Indo-European families, several genera can be distinguished, each grouping a few languages which are even more closely related. A number of models of those similarities were soon proposed. For example, Schmidt [1872] proposed a wave-model: each language (or genus) develops a certain number of innovations, which spread to some, but rarely all, languages in the family, as shown in Fig. 1.2. He grouped Balto-Slavic (Russian, Polish, Lithuanian...) with Germanic (English, German, Swedish...) because both have an /m/ at certain case endings, where other languages have a /bh/. The hypothesis was therefore that one Germanic or Balto-Slavic language had developed an /m/ case ending, and that that innovation had spread to its neighbours, without reaching the other Indo-European languages. On the

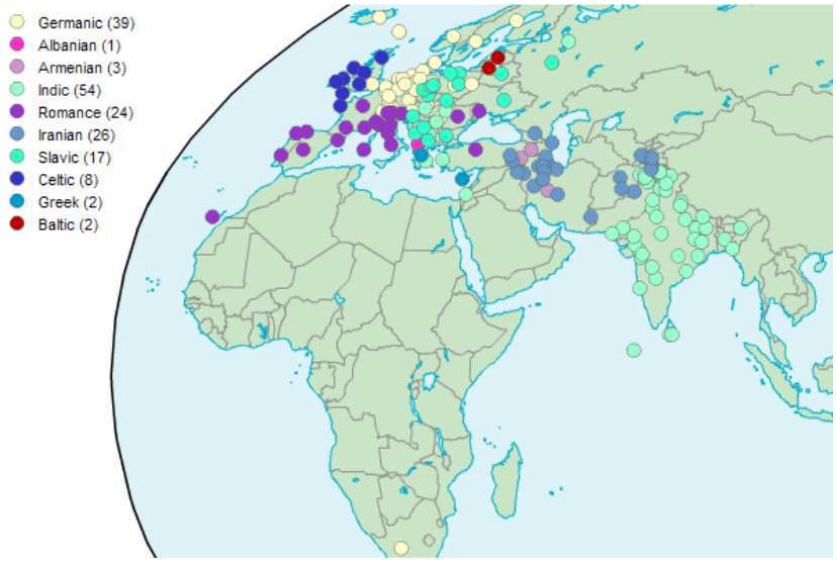

Figure 1.1: *Map of 176 Indo-European languages, shown by genus. Made using the electronic version of Dryer et al. [2003].*

other hand, he also grouped Balto-Slavic with Armenian and Indo-Iranian, because those languages have an /s/ where other languages have a /c/ [Mallory, 1989].

The wave model is appropriate in a small number of cases, but in general, a more appropriate model is a genetic one. There is a strong tree-like signal in linguistic data [McMahon and McMahon, 2005], and the observation of recent language change through written texts (e.g. Latin to Italian, French and Spanish) shows that modification with descent accounts for most of the changes. Schleicher [1850] was the first to propose such a model, and he introduced the tree representation. Borrowing vocabulary from biology, he introduced terms such as *genus*, *species* or *variety* to describe language groupings, and made the first attempt at an evolutionary tree of Indo-European languages. The similarities between language diversification and biological

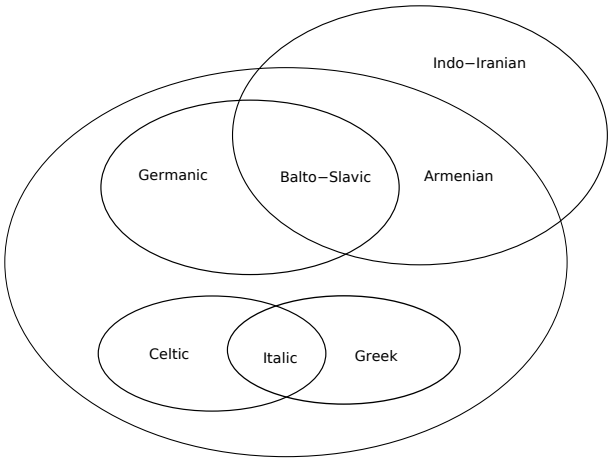

Figure 1.2: *Schmidt's model of relationships between Indo-European languages. Each language is part of one or several groupings, which may intersect.*

evolution were also noted by Darwin [1871]: "The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. ... We find in distinct languges striking homologies due to community of descent, and analogies due to a similar process of formation." Indeed, there are striking similarities between the processes of biological evolution and linguistic diversification: like genes, the vocabulary, phonology and morphosyntax of languages are passed on from parents to children in a process of descent with modification.

With a tree-like model, two questions need answering: What is the topology of the tree? How old is the root? Traditionally, the topology would be reconstructed by hand by an expert linguist on the languages under study, using the *comparative method*. By comparing words of identical or similar meanings in different languages, the linguist would try and identify *cognates*, *i.e.* meaning categories for which the languages have related words (see Section 2.1 and Table 2.1 for more details and an example of cognacy classes); at

the same time, they would identify the systematic phonetic correspondances through which the languages passed. They would then reconstruct the topology of the tree, and the sound changes that occured. Comparative linguists often view their subject more like an art than a science.

There was no direct way of estimating the age of the root, so archaeological evidence was used for dating issues. In the case of Indo-European, this has led to a controversy [Diamond and Bellwood, 2003]. The main hypothesis amongst linguists was postulated by Gimbutas and Hencken [1956] and holds that the most recent common ancestor of all known Indo-European languages branched no earlier than about $6000 - 6500$ years Before the Present (BP), with the expansion of the Kurgan horsemen, a people living in steppes north of the Black Sea. The proponents of the Kurgan hypothesis hold that domestication of the horse and of the wheel gave the Kurgan a significant military advantage allowing an enormous expansion [Mallory, 1989]. An alternative hypothesis suggests that the spread began around 8500 BP when the Anatolians mastered farming in the early Neolithic [Renfrew, 1987]. Dating the root of the Indo-European languages would therefore shed light on the events which allowed the Indo-European family to spread so far. This issue was the main question behind much of the work presented in this thesis.

The first systematic method for estimating the root age, glottochronology, was developed by Morris Swadesh in the 1950s. Swadesh used a list of about 200 meanings of "core vocabulary" (later refined to a list of 100 meanings), which were assumed to evolve at a constant (slow) rate. Given lists of core vocabulary for two languages, he would decide whether their words for a given meaning were cognate. The percentage of cognate words would then

directly translate into the date t at which the two languages had split, through the formula $t = \frac{\log C}{2 \log r}$, where C is the proportion of shared cognates, and r is the "retention constant", assumed to be a constant across all languages [Swadesh, 1952]; r could be estimated using any pair of languages for which the age of the common ancestor was known. Bergsland and Vogt [1962] gave a deadly blow to glottochronology when they used known dates to show that the retention constant is in fact not a constant, but takes very different values in the three groups of languages they selected (Old Norse, Icelandic and Norwegian; Old and Modern Georgian and Mingrelian; and Old and Modern Armenian). Glottochronology has now been discredited, despite criticisms of the issues raised by Bergsland and Vogt [1962] [Sankoff, 1970]. Indeed, it seems wishful thinking to hope to summarise language change in a single number. However, there are issues with Bergsland and Vogt [1962]'s work, and we show in Chapter 6 that modern methods are not subject to the same criticisms. Several decades later, the linguistic community remains nonetheless very skeptical about attempts at dating.

1.2 Phylogenetic models for linguistic data

Recent advances in phylogenetics have made it possible to fit models of much greater complexity and much closer to reality.

The first attempts at fitting phylogenetic models to linguistic data were described in articles by Gray and Jordan [2000] and Gray and Atkinson [2003]. Gray and Jordan [2000] apply maximum parsimony to a dataset of 5,185 lexical items from 77 Austronesian languages and find evidence supporting one of

the main hypotheses of Austronesian expansion, the so-called "express-train" hypothesis that Austronesia was colonized rapidly by farming people out of Taiwan. They do not attempt to date any of the internal nodes. Holden [2002] also apply maximum parsimony to 75 Bantu and Bantoid languages (spoken South of the Sahara); they find that the most parsimonious tree follows the expansion of farming in sub-Saharan Africa, indicating that it was the mastering of farming that allowed the Bantu people to colonize such a large area.

Gray and Atkinson [2003] apply the phylogenetic method to Indo-European language data collected by Dyen et al. [1997], using the finite sites model implemented in the MrBayes package by Huelsenbeck and Ronquist [2001]. The results obtained, using penalized maximum likelihood, are very close to the trees given by classical comparative linguistics methods. They impose part of the topology of the tree; other, unconstrained, genera yielded by their analysis correspond to those usually accepted by historical linguists. Several of the subfamilies they find also correspond to what was already believed to be true, such as a subfamily grouping Germanic, Romance and Celtic languages. Other parts of the tree are unresolved, such as the position of Albanian. While their results do not show any new groupings, the fact that the results correspond to what is accepted by linguists gave hope that phylogenetics could be successfully used in other, less studied, language families. More controversially, they attempt to date the internal nodes and root of the tree and estimate the age of Proto-Indo-European to be between 7800 BP and 9800 BP, in line with the Anatolian hypothesis. However, the finite sites model they use is not well-suited to lexical data; in particular, it allows for homoplasy, *i.e.* for a single

cognacy class to be born several times at different points of the tree, which is not appropriate. The same methods were applied with similar results to another data set of Indo-European languages by Atkinson et al. [2005], but there is little in terms of validation of the findings.

This line of research has spawned a lot of interest, including outside of scientific circles [Wade, 2004]. A number of other groups applied various methods from biology to linguistic datasets.

Rexova et al. [2003] perform a maximum parsimony analysis of the Dyen et al. [1997] data of Indo-European lexical items. They reconstruct all the known major features of the Indo-European family, but there is a lot of uncertainty about the topology close to the root. They note that the basic vocabulary of Indo-European is strikingly tree-like. Bryant et al. [2005] come to the same conclusion in their analysis of the data from Dyen et al. [1997] using NeighbourNet.

McMahon and McMahon [2005] also apply the Neighbour-joining method to 95 Indo-European languages and 200 lexical items. The unrooted tree they obtain shows the ten genera generally admitted for Indo-European languages, but no clear relationship appears between the genera.

Kitchen et al. [2009] use the BEAST software [Drummond and Rambaut, 2007] to estimate the age of the most recent common ancestor to the Semitic languages, but do not give the uncertainty of their estimates.

Dunn et al. [2005] apply maximum parsimony analysis to the "structural" (morphosyntactic and phonological) features of a set of Oceanic languages which are known to be closely related; the same authors later moved on to a Bayesian phylogenetic analysis using various models from biology [Dunn

et al., 2008]. Again, their results are similar to those linguists obtain using the comparative method. They argue that the tree model is better suited to structural features than to lexical traits, because borrowing of structure will only happen if lexical borrowing also occurs, whereas the converse is not true [Moravcsik, 1978] (although Nichols [1999] takes the opposite view that grammatical features evolve rapidly, and are more subject to regional influences than lexical data). Since the reconstruction of the Oceanic family is good, they extrapolate their method to a number of Papuan languages whose history is not well known. They conclude that the Papuan languages must either share a common ancestor, or that there must have been contact between the Papuan languages before 3200 BP.

Views diverge as to whether morphosyntactical or lexical items retain more signal from a language's ancestor. For example, Thomason [2000] notes that the Ma'a language of Tanzania was apparently originally in the Cushitic family, and retains much Cushitic basic vocabulary, but that its syntax has been modified beyond recognition through contact with neighbouring Bantu languages. The same author claims that this occurs in situations where speakers of the source language learn the receiving languages imperfectly, whereas borrowing of lexical but not morphosyntactical items occurs when speakers of the receiving language adopt new items from the source langage [Thomason, 2001].

Lansing et al. [2007] compared lexical and genetic data from the Indonesian island of Sumba. They claim to show that the 29 Sumbanese languages they sampled form a subclade of the Austronesian family, though their argument is not clear. More interestingly, they find a positive correlation between retention

of Proto-Austronesian cognates and Austronesian Y chromosome lineages, as well as a correlation with the distance to the place where it is believed the ancestors of the modern populations first debarked on Sumba.

Although some have understood the large error bars resulting from statistical analysis of comparative data to be a sign of weakness in these methodologies, they are actually one of its main strengths: the comparative method as presently formulated does not give estimates of uncertainty [Pagel, 2000].

Another common criticism is that the tree model is not as well suited to languages as it is to biological species, even though Mallet [2005] found hybridization in 10% of animal and 25% of plant species reviewed. In response to Comrie [2006], who stated that the tree model "necessarily involves a simplification of the actual historical facts", Campbell [2006] contends that "historical linguists would say it is not a simplification; rather, [these methods] address directly only inherited material, while other methods and techniques help to complete the picture". This is all we can hope for: that the quantitative methods will describe the aspect of language history which they attempt to describe, rather than the complete picture. In general, the tree model is a good fit for core vocabulary.

Some linguists have asked for models more specifically tailored to linguistic data [Croft, 2008]. Noting the need for models specific to linguistic data, rather than borrowed from biology, Warnow et al. [2004] developed a model for lexical, phonological and structural data which includes some of the specificities of language diversification. For phonological and structural data, back-mutation to a specific "default" homoplastic state is allowed. They assume that language

evolution is mainly treelike, but add some edges between synchronic languages to model borrowing. They assume complete rate heterogeneity: the rates are different along different branches, and for different characters. They do not fit the model, but prove that the topology of the tree is identifiable (modulo the placement of the root, and the rate parameters). The same group propose to construct "perfect phylogenetic networks" [Nakhleh et al., 2005] by transforming a maximum parsimony tree into a network. Ben Hamed et al. [2005] and Bouchard-Côté et al. [2007] develop models for phonological data and use them to identify language families.

Nicholls and Gray [2008] fit a model specifically tailored to linguistic data, which is a stochastic extension of the model described by Dollo [1893] and used in a biological context by Le Quesne [1972] and Farris [1977] . Their results also support the Anatolian hypothesis, and they study a number of model misspecifications which could have introduced systematic bias. In particular, they note that rare but strong rate heterogeneity could have a big influence on date estimates, since it would be hard to detect and at the same time have a large effect. Our own analyses also show that the way missing data are handled by Nicholls and Gray [2008] is crude and it forces them to discard several languages from their analyses. Chapter 2 of this thesis expands the model they proposed. We apply it to several linguistic data sets in Chapter 5.

Language trees constructed with phylogenetic methods have been used to study other aspects of language change. Lieberman et al. [2007] study strong verbs in Old, Middle and Modern English and evaluate the rate at which strong verbs are regularized; they find that the rate of regularization changes with the square root of the usage frequency: a verb which is used 4 times more often

than an other will regularize at half the rate. The research by Lieberman et al. [2007] was made easier by the fact that they knew the ancestry of the three languages they were studying. Pagel et al. [2007] wished to estimate variation in rates for vocabulary. Since these rates are much lower than the rates of regularization of strong verbs, they needed to study a longer time period and they therefore chose to infer rates on the entire tree of the Indo-European family. They use posterior samples using the same finite sites model as Gray and Atkinson [2003] and compare the rates for each meaning category to the frequency of use in four modern Indo-European languages (English, Spanish, Russian and Greek); they find a significant negative correlation between frequency of use and diversification rate. They are able to explain 50% of the variation in the rates of evolution with two pieces of information: frequency of word use, and part of speech (verb, noun, number...) Pagel and Meade [2006] apply similar methods to the Bantu family.

Similarly, Atkinson et al. [2008] used Pagel et al. [2004]'s software BayesPhylogenies to build trees for the Indo-European, Bantu and Austronesian languages. They found that more changes occured on paths with more branching, and concluded that between 10 and 33% of language change happens in punctuational bursts, when two languages diverge, with the remainder of language change happening continually through time. We revisit this issue in Chapter 6.

There have also been some efforts to detect cognates automatically rather than by using the judgements of expert linguists [Mackay and Kondrak, 2005], but this area is still very much under-developed.

As opposed to the sound science in many of the articles cited above, there

also seems to be a tendency amongst some non-specialist scientists to consider phylogenetic analyses of linguistic data as an easy and fun problem, leading to a number of publications with little or no linguistic or statistical grounding.

Serva and Petroni [2008] notice that the comparative method can only be applied to languages which have been extensively studied by linguists and they claim that the classification into cognate classes leads to "subjectivity" which should be avoided. Rather than relying on the judgements of linguists, they use the spellings of words on the Swadesh list for Indo-European languages, compute the Levenshtein distance between each pair of words and apply the UPGMA method to the data they collect; they hope to apply the same methods to less studied language families. The intention is good, but their methodology is unfortunately flawed. Since they are dealing with languages which use at least five different alphabets, they transcribed these languages into the Latin alphabet. There are many ways of doing this, so the objectivity they were going for is under question. Unsurprisingly, the languages in their reconstructed tree are grouped by alphabet: Albanian is grouped with the other languages using the Latin alphabet, a position which no linguist could believe to be correct, whereas Greek and Armenian, which use alphabets of their own, are the outgroups. The Germanic, Celtic, Italic and Indo-Iranian subgroups are correctly reconstructed, but the details of these groups is very far from the known linguistic truth (e.g. the position of English and Polish). Finally, their formula for dating is simplistic, relies on only two known dates (though many others were available) and they do not even attempt to measure the uncertainty of their results. The method could probably be vastly improved by relying on a phonetic transcription rather than the spelling of the words in their data and

by defining a non-binary distance, so that similar phonemes are deemed closer than very distinct phonemes, like Ben Hamed et al. [2005].

Forster and Toth [2003] study Celtic and Romance languages but they are casual with their data collection and in particular with the cognacy judgements. Rather than use cognacy judgements made by expert linguists, they group together words which look somewhat similar even if they are known by linguists to come from different origins, and separate other words which are known to be cognates; see Evans et al. [2004] for an extensive criticism. Forster and Toth [2003] date the expansion of Celtic to 5200 ± 1500 BP. By adding a single language, Greek, to their list of Celtic and Romance languages, they estimate the age of Proto-Indo-European at 10100 ± 1900 BP. An underlying assumption of their estimate is that Greek is an outgroup of the Indo-European family, which is far from certain.

It is worth noting that none of these works, nor ours, attempts to discover new language families. Rather, given a set of languages already known to be related, we estimate its *internal* structure and dates.

1.3 Phylogenetics for cultural history and anthropology

A number of cultural aspects diversify in a way similar to languages [Mace and Holden, 2005, O'Brien and Lyman, 2002, Mesoudi et al., 2004]; as such, there have been many recent attempts to apply phylogenetic methods to cultural data sets, as suggested by Pagel [1992]. In general, software developed for biological data has been used without questioning the assumptions made by the models. It is often fair to say that cultural traits are closer to linguistic than

biological traits, so such research would greatly benefit from more developed and better publicized software for linguistic diversification. Examples include data sets on Paleoindian points [O'Brien et al., 2001], European neolithic pottery [Collard and Shennan, 2000], Native American baskets [Jordan and Shennan, 2003] and East African kinship and marriage traditions [Mulder et al., 2001]. Skelton [2008] examines differences in ways of writing Linear B on pottery to reconstruct the history of the language.

Cultural evolution can be harder to analyse: while it is clear how to split linguistic data into a list of traits, defining such a list of traits in a systematic and unbiased way for cultural data is not as easy. Another issue is that of taxon construction: is it easier to decide whether two animals are members of the same species than to decide whether two people speak two variants of the same language or two distinct but closely related languages; it is even harder to define what constitutes a single leaf in a tree of cultural evolution [O'Brien et al., 2002].

Gray et al. [2007] suggest a three-dimensional space to help choose cultural datasets which a tree model would suit best. They propose to look for cultural traits which minimize the rate of change in vertical transmission and the rate of horizontal transmission, and which maximize "the extent to which different aspects of culture are coupled together", but the methods to measure these characteristics are somewhat unclear.

There has also been interest in comparing evolutionary trees obtained from genetic data with those that come out of linguistic or cultural data sets [Jones, 2003]. This line of research was initiated by Cavalli-Sforza et al. [1988], who incorporated genetic data from all populations in the world and linguistic

data from all language families, and claimed to find "considerable parallelism between genetic and linguistic evolution". However, the linguistic groupings they used, which include the superfamilies of Greenberg [1987], are highly controversial.

Holden and Mace [2003] assume that language trees are a good model of cultural evolution as well: in order to test whether spread of cattle is correlated with matriliny[1] in Bantu populations, they plot traits coding for matriliny or patriliny, and for possession of cattle, on a phylogenetic tree of language diversification. Assuming that the cultural traits had evolved along the same tree, they show that matrilineal societies became patrilineal after they acquired domestic cattle. Similarly, Fortunato et al. [2006] use a phylogenetic tree of 51 Indo-European languages (a subset of the Dyen et al. [1997] data set which we analyse in later chapters) to estimate that dowry was likely to already exist in the Proto-Indo-European society.

Spencer et al. [2004] perform a simulation study where the true tree is known: they simulate copying of manuscripts by asking 20 modern "scribes" to copy a poem, each from a previous copy. They use Neighbour joining and maximum parsimony to reconstruct the evolutionary tree. Both methods yield results which are close to the true tree. These positive results were seen as a validation of the work by Barbrook et al. [1998], who apply an undisclosed phylogenetic method (presumably maximum parsimony) to various manuscripts of "The Wife of Bath's Prologue" from *The Canterbury Tales*, and show that some manuscripts which have been ignored by scholars are actually

[1] In a matrilineal society, group membership is inherited from an individual's mother rather than from their father.

the closest to Chaucer's original. Spencer et al. [2003] also build a maximum parsimony tree using the order in which the *Canterbury Tales* are written in a subset of the manuscripts.

While there is a lot of interest for applying phylogenetic methods to linguistic and cultural data, there are surprisingly few models specifically tailored to these, meaning that most of the research is conducted using models borrowed from biology. These certainly give a rough idea, but estimates using models which include specificities of language change would be more reliable and less prone to resistance by linguists.

In this thesis, we focus on a tree model for lexical data, which we call a stochastic Dollo model, after Dollo [1893], who first proposed a model where the loss of a trait is irreversible: once a species has lost a trait, it cannot reevolve that same trait. This principle, later called "Dollo's law", is thought to be almost universal for complex morphological traits in biology, although exceptions have been found [Pagel, 2004]. In the model we describe in Chapter 2, a trait can only be born once on a tree: in other words, if a trait is displayed at two leaves of the tree, these two instances of the trait must be homologous; this corresponds to how we expect lexical change to occur. This model is similar to models developed for biological trait data by Huson and Steel [2004] and Alekseyenko et al. [2008], and with minor adjustments, our model could probably be applied to certain biological data sets.

Recent improvements in computational power have allowed Bayesian inference to become preeminent in phylogenetics, following Yang and Rannala [1997]. In particular, many techniques for estimation via Markov Chain Monte Carlo have been developed [Larget and Simon, 1999]. In our context,

Bayesian inference presents two main advantages. First, it allows us to estimate uncertainties in a natural way. Second, it allows us to explicit our prior beliefs on key parameters. Since we are mostly interested in dating the root of the Indo-European family, we describe in Chapter 2 how to impose a uniform prior on this statistic.

Issues of dating on phylogenetic trees have been researched extensively for molecular phylogenetics. Thorne et al. [1998] study a model of evolution of the rate of molecular evolution and propose methods to estimate dates on a phylogenetic tree when the hypothesis of a constant molecular clock does not hold; Thorne and Kishino [2002] study the detection of correlations in the evolution of rates. Yang and Rannala [2006] examine issues with calibration data, especially the influence of "soft bounds", which allow (but discourage) known ancestral nodes to lie outside of a constraint. Rannala and Yang [2007] and Inoue et al. [2010] discuss the impact of the prior and of the size of the data on the uncertainty in posterior estimates. Much of this research is relevant to phylogenetic dating questions in Linguistics.

Some linguists have already embraced the idea of applying phylogenetic methods to linguistic data [Fitch, 2007], but others are fiercely opposed to the concept [Holm, 2007, Marris, 2008]. The doubts that have been expressed have not been alleviated by articles which ignore linguistic fact: for example, [Evans et al., 2004] criticize Forster and Toth [2003] quite sternly, but also indicate that this flawed methodology has led them to be wary of many other phylogenetic analyses of linguistic data, including analyses which do not have such obvious flaws. Because of the glottochronology fiasco, dating methods are met with even more skepticism [McMahon and McMahon, 2006]. As such,

particular attention must be given to the validation of the methods (what Gelman and Hill [2007] call "confidence-building"). Large parts of this thesis focus on such confidence-building.

In Chapter 2, we present a model tailored to lexical change. We give details on how we implemented the fit of the model in Chapter 3. We provide a number of validation checks in Chapter 4 and analyse two data sets of Indo-European languages in Chapter 5. In Chapter 6, we discuss other possible applications of our methods.

Chapter 2

Description of the data and of the model

2.1 Description of the data

Our initial focus is on two data sets of vocabulary for Indo-European languages, one collected by Ringe et al. [2002] and the other by Dyen et al. [1997]. Ringe et al. [2002] collected data from 328 meaning categories for 24 mostly ancient languages and coded the data in 3174 homology classes. Dyen et al. [1997] collected data for 84 modern languages in 207 meaning categories; Gray and Atkinson [2003] added 3 ancient languages (Hittite, Tocharian A and Tocharian B) to the list, bringing it up to 87 languages and 2449 homology classes. There is little overlap between the lists of languages in the two data sets.

Both datasets cover the "core" vocabulary: meanings such as *all*, *animal*, *ashes*... (A complete list is given in Appendix A.) These meaning categories are defined in advance and are expected to exist in all languages. Expert

linguists established which words share a common ancestor. Two words in the same meaning category which can be shown to be descended from a common ancestor through systematic phonological changes are called *cognates*. This equivalence relationship classifies words into *cognacy classes*. For example, for the meaning "head", the Italian *testa* and the French *tête* belong to the same cognacy class, while the English *head* and the Swedish *huvud* belong to another cognacy class. An element of a cognacy class is thus a word in a particular language. The vocabulary of a single language is represented as a set of distinct cognates. In some cases, cognacy classes can be very hard to detect, due to the amount of phonetic changes[1].

If there are N distinct cognacy classes in data for L languages, then the a'th class $M_a \subseteq \{1, 2, ..., L\}$ is a list of the indices of languages which possess a cognate in that class. The data are often coded as a binary matrix D. A row corresponds to a language and a column to a cognacy class, so that $D_{i,a} = 1$ if the a'th cognacy class has an instance in the i'th language, and $D_{i,a} = 0$ otherwise. See Table 2.1 for an example. This coding allows a language to have several words for one meaning (such as Old High German *stirbit* and *touwit* for "he dies", an instance of polymorphism), or no word at all (see Section 4.2.2 for a discussion of issues this raises). Missing matrix elements mostly arise because the reconstructed vocabularies of some ancient languages are incomplete. For some modern languages, small amounts of data are also missing; this may be because the linguists are unsure as to whether a word belongs to a cognate class. If we are unable to answer the question "does

[1]For instance, the English *wheel* and the Greek *κύκλος* (whence *cycle*) are cognate: both come from the Proto-Indo-European word reconstructed as $*k^wek^wlos$, but this is certainly not obvious to the untrained observer.

Old English	*stierfþ*
Old High German	*stirbit, touwit*
Avestan	*miriiete*
Old Church Slavonic	*umĭretŭ*
Latin	*moritur*
Oscan	*?*

(*a*)

Old English	1	0	0
Old High German	1	1	0
Avestan	0	0	1
Old Church Slavonic	0	0	1
Latin	0	0	1
Oscan	?	?	?

(*b*)

Table 2.1: An example of data coding: (a), the word "he dies" in six ancient Indo-European languages; (b), the coding of this data as a binary matrix with ?'s for missing data. The first cognacy class is $M_1 \in \Omega_1$ with $\Omega_1 =$ {{Old English, Old High German}, {Old English, Old High German, Oscan}}

language i possess a cognate in cognacy class a?" then we set $D_{i,a} =?$.

We need notation for both matrix and set representations with missing data. Denote by B_a column a of $L \times N$ matrix B. For $a = 1, 2, ..., N$ let $\mathcal{D}_a$ be the set of all column vectors d^* allowed by the data D_a in column a of D,

$$\mathcal{D}_a = \left\{ d^* \in \{0,1\}^L : D_{i,a} \in \{0,1\} \Rightarrow d^*_a = D_{i,a},\ i = 1, 2, ..., L \right\}.$$

For $d^* \in \mathcal{D}_a$ let $m(d^*) = \{i : d^*_i = 1\}$. Denote by Ω_a the set of cognacy classes consistent with the data D_a, so that

$$\Omega_a = \{\omega \subseteq \{1, 2, \ldots, N\} : \omega = m(d^*), d^* \in \mathcal{D}_a\}.$$

The data D are then equivalently $\Omega = (\Omega_1, \Omega_2, ..., \Omega_N)$. The Ω_a-notation generalizes the M_a-notation to handle missing data.

Ringe et al. [2002] list 24 mostly ancient languages. For 11 of these languages (Latin, Modern Latvian, Old Norse...), all the data are known. For

the others, the proportion of missing entries varies between 1% (for Old Irish) and 91% (for Lycian, an ancient language of Anatolia). Of the 87 languages listed by Dyen et al. [1997], 27 have no missing data. Most other languages have very few missing data points; the proportion of missing entries varies between 0.2% (for Slovenian) and 25% (for Tocharian A).

Note that data are usually missing in small blocks corresponding to the cognacy classes for a given meaning category, as in Table 2.1. We do not model this aspect of the missing data. This is related to the model-error Nicholls and Gray [2008] call 'the empty-field approximation', under which cognacy classes in the same meaning category are assumed to evolve independently. See Section 4.2.2 for an analysis of possible systematic bias this could entail.

Figure 2.1 gives a visualisation of the data by Ringe et al. [2002] restricted to 100 meaning categories, which lists data for 24 languages across 872 cognacy classes. There are 8 blocks (large rows) in Figure 2.1, each corresponding to 109 cognacy classes. In each block, a row corresponds to language and a column to a cognacy class. A small black square corresponds to a 1 in the data; a white square corresponds to a 0 in the data, and a gray square corresponds to a ?. This figure shows a few of the features of the data: most of the data are 0's; most cognacy classes are only displayed at a very small number of languages (1 or 2), but a few appear at almost all languages; and data are missing in blocks.

It is also worth noting that we are the first to propose a model tailored to language diversification which correctly handles missing data. In previous research, missing data were generally assumed to be absent (?'s were replaced with 0's); Nicholls and Gray [2008] also had to discard 7 languages from their

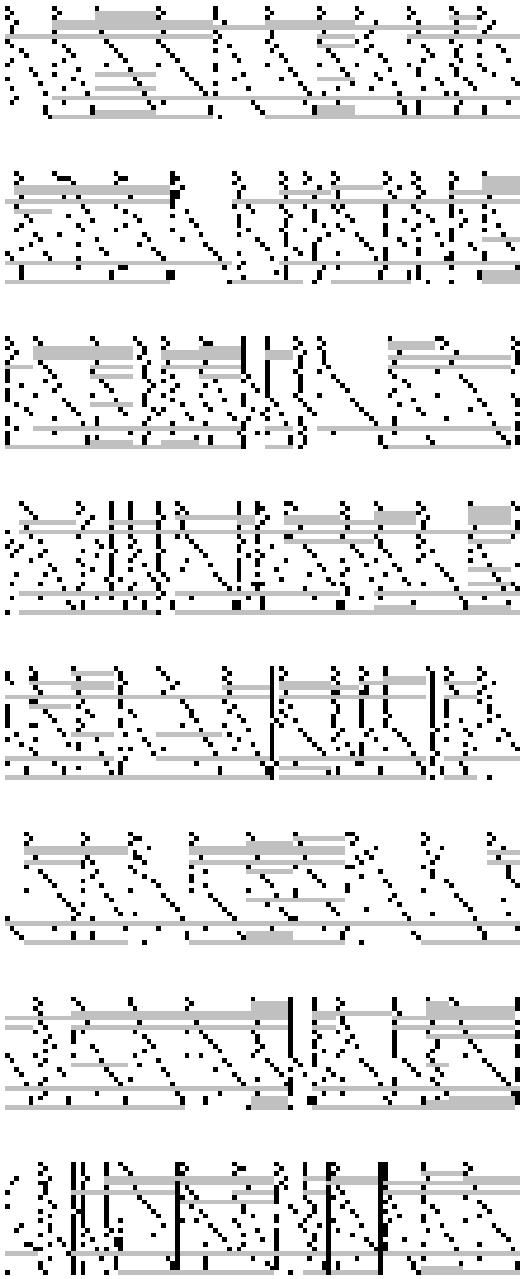

Figure 2.1: Visualization of the Ringe et al. [2002] data. For a given language (row), a cognacy class (column) can be present (black), absent (white), or the data point can be missing (gray).

study because they contained too many missing data, whereas we are able to include all languages in our analyses. Several linguists have expressed regret at the lack of correct handling of missing data, including Skelton [2008] and Michael Dunn (pers. comm.).

For these Indo-European phylogenies, some subtrees are known. For example, the Italic languages are known to form a subtree, and they are known to have diverged after the fall of Dacia in 112 AD [Gray et al., 2007]. Similarly, we have some knowledge of the dates at which ancestral languages were spoken; these take the form of a lower and an upper bound on the subtree root age. We also have constraints on the age of all non-contemporary leaves, since we know when these extinct languages were in use. The bounds are used to calibrate model parameters and to infer dates for other nodes. Table 2.2 lists the 15 constraints we use in our analyses of the Dyen et al. [1997] data. Jumping ahead to our results, Figure 2.2 is a sample from the posterior distribution we find for phylogenies in our analysis of the Ringe et al. [2002] data. Calibration constraints are represented by the black bars across nodes in this tree.

2.2 Model description

We specify a subjective prior for phylogenies, representing a state of knowledge of interest to us. We model vocabulary diversification down an evolutionary tree, where each leaf represents a language in our data.

The material in this subsection follows Nicholls and Gray [2008], with several extensions: we add catastrophic rate heterogeneity, a correct handling of missing data, and more registration processes.

Clade	Min age	Max age
Celtic	1700	∞
Brythonic	1450	1600
Italic	1700	1850
Iberian-French	1200	1550
Germanic	1750	1950
Balto-Slavic	1900	3400
Slavic	1300	∞
Indic	2200	∞
Indo-Iranian	3000	∞
Iranian	2500	∞
Greek	3500	∞
Tocharic	1650	2140
Hittite	3200	3700
Tocharian A	1250	1500
Tocharian B	1250	1500

Table 2.2: *Clade constraints for the Dyen et al. [1997] data set, first used by Gray and Atkinson [2003]. The first twelve rows are clades for which we have knowledge about the age of the root; the last three are ancient languages for which we know when they existed.*

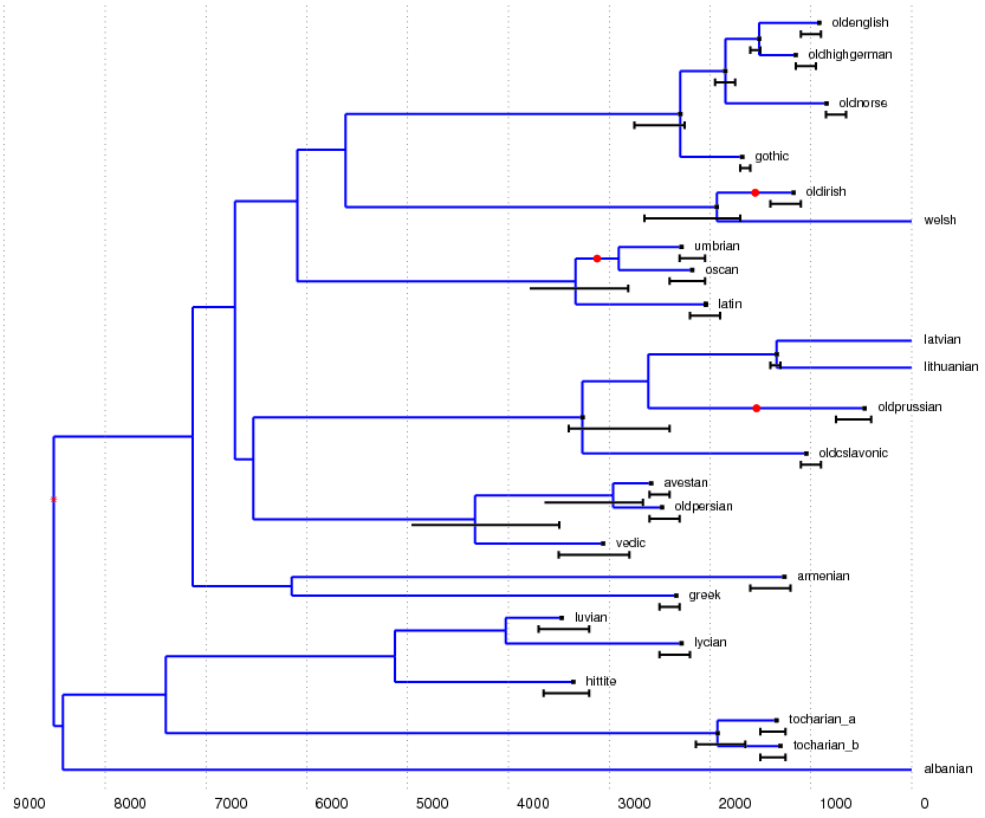

Figure 2.2: *A sample close to the maximum of the posterior for the analysis of the Ringe et al. [2002] data. All the constraints on the node ages are shown. The age constraints for the Italic, Indo-Iranian and Iranian groups do not have an upper bound; this is denoted by the absence of a tick on left side of the bound.*

2.2.1 Prior distribution on trees

Let g be a rooted tree with $2L$ nodes: L leaves, $L-2$ internal nodes, a root node $r = 2L-1$ and an Adam node $A = 2L$, which is linked to r by a edge of infinite length. Each node $i = 1, 2, \ldots, 2L$ is assigned an age t_i and $t = (t_1, t_2, ..., t_A)$; the units of age are years before the present; for the Adam node, $t_A = +\infty$. The edge between parent node j and child node i is a directed branch $\langle i, j \rangle$ of the phylogeny, with the ordering $t_i < t_j$. Let E be the set of all edges, including the edge $\langle r, A \rangle$, let V be the set of all nodes and let $V_L = \{1, 2, ..., L\}$ be the set of all leaf nodes.

We are initially interested in the set Γ of all rooted directed binary trees $g = (E, V, t)$ with distinguishable leaves $i = 1, 2, \ldots, L$. With this notation, (E, V) is the topology of a rooted directed binary tree. Let $\sigma(t)$ be the order of the internal ages t_i for $i = L+1, \ldots, 2L-2$. Then the triplet $(E, V, \sigma(t))$ is the labeled history of the tree. In general, several labeled histories correspond to one topology.

We restrict the set of allowable trees Γ by imposing C calibration constraints on the topology and on certain node ages. These are described at the end of Section 2.1. Each constraint c restricts Γ to the set of trees $\Gamma^{(c)}$ in which a certain set of leaves form a sub-tree, or clade. Some constraints also impose a lower and/or upper bound of the root of the clade. In some cases, the imposed clade contains only one leaf, so that the constraint is only imposing a range on the age of the leaf; this is used for ancient languages. We add to these constraints an upper bound on the root time at some age $T > 0$. In most of our analyses of Indo-European data, we use $T = 16,000$, which is

much greater than any possible root age considered plausible by linguists. Let $\Gamma^{(0)} = \{(E, V, t) \in \Gamma : t_r \leq T\}$. The space of calibrated phylogenies is then

$$\Gamma^C = \bigcap_{c=0}^{C} \Gamma^{(c)}.$$

Since we are interested in the root age, we choose a prior on trees for which the marginal prior on the root age t_r is uniform over an interval $t_L \leq t_r \leq T$. The uniform prior on Γ puts more weight on greater values of t_r. If all leaves i are isochronous with $t_i = 0$ and there are no calibration constraints, then Nicholls and Gray [2008] show that the prior $f(g|T) \propto t_r^{2-L} \mathbb{I}_{t_r \leq T}$ has the desired uniform marginal prior on t_r. However, the inclusion of calibration constraints complicates matters.

For node $i \in V$, let $t_i^+ = \sup_{g \in \Gamma^C} t_i$ and $t_i^- = \inf_{g \in \Gamma^C} t_i$ be the greatest and least admissible ages for node i, and let $S = \{i \in V : t_i^+ = T\}$, so that S is the set of nodes having ages not bounded above by a calibration (there are 12 such nodes in Figure (2.2), for example the most recent common ancestor to Latin, Umbrian and Oscan). Nicholls and Gray [2008] show with simulation studies that the prior probability distribution with density

$$f_G(g|T) \propto \prod_{i \in S} (t_r - t_i^-)^{-1}$$

gives a marginal density for t_r which is approximately uniform in $t_L < t_r < T$ if in addition $T \gg \max_{i \in V \setminus S} t_i^+$. This is the prior we use. Nicholls and Gray [2008] do not comment on the distribution determined by f_G over tree

topologies. We consider two priors on topologies: a uniform distribution on labeled histories (corresponding to the distribution for the Yule [1925] model), which favours balanced topologies [Velasco, 2008] and a uniform distribution on topologies, which favours small and large clades against medium-sized clades [Goloboff and Pol, 2005]. In both cases, the marginal prior is defined over Γ rather than Γ^C. The addition of constraints modifies this prior in two ways: topological constraints rule out certain topologies, setting the prior probability to 0; age constraints modify the volume available to each topology, so that our priors are in fact not exactly uniform over labeled histories or topologies over Γ^C. For the validation analyses presented in Chapter 4, we use only the uniform prior on labeled histories; we used both priors for our analyses of real data presented in Chapter 5.

The addition of catastrophes described in Section 2.2.2 has no impact on the marginal prior density for t_r, nor on the marginal prior distribution on topologies. Figure 2.3 shows a sample from the prior of the root age with the constraints from the Ringe et al. [2002] data. The prior is roughly uniform between 5000 and 16000 BP, which covers our region of interest; the prior does not correspond to any reasonable *a priori* belief before 4500 BP, but this does not matter since this region is ruled out by the likelihood.

The prior on the ages of internal nodes and leaves depends on the constraints. Take for example the Tocharian B leaf: it is constrained to lie between 1250 and 1500 BP, and its parent (the common ancestor with Tocharian A) is constrained to lie between 1650 and 2140 BP. The rest of the tree exerts little influence on the prior on the age of Tocharian B, and so that prior is approximately uniform over the allowed range (as shown by the

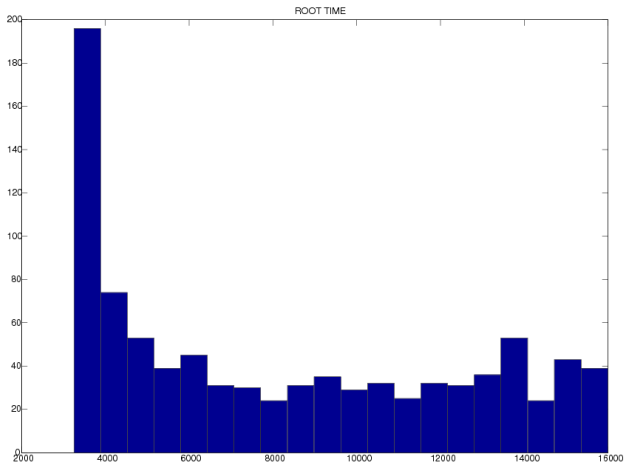

Figure 2.3: *Sample from the prior distribution on the root age. The prior is approximately flat over the region of interest (between 5000 and 16000 BP).*

sample from the prior in Figure 2.4). For other nodes which are constrained to lie between an upper and a lower bound, the prior is not necessarily exactly uniform, but simulations from the prior show that it is always very close to uniform, presumably because the authorized range is always quite small. This is not true, however, of nodes which have no upper bound on the age: Figure 2.5 shows a sample from the prior for the most recent common ancestor to the Italic languages, which is only constrained to be older than 3000 BP. The prior is strongly biased towards younger ages, which is unsurprising since younger ages leave more volume available for the nodes above the Italic clade. In general, we are not interested in dating internal nodes, so this is not an issue. The marginal prior on the age of this node is in fact comparable to priors obtained using "soft bounds" in molecular phylogenetics Yang and Rannala [2006]. We have no further information on the prior belief linguists have on the distribution of plausible ages for ancient nodes, bur it would be interesting

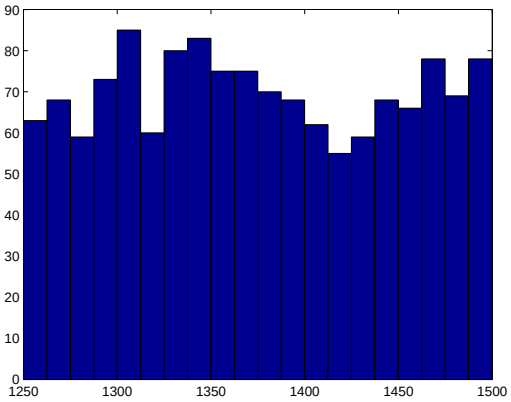

Figure 2.4: *Sample from the prior distribution on a leaf age (Tocharian B).*

to include it if such information became available.

2.2.2 Diversification of cognacy classes

In this subsection we extend the stochastic Dollo model of Nicholls and Gray [2008] to incorporate rate heterogeneity in time and space, via a catastrophe process.

Although this model was developed with languages in mind, it can also be applied to other kinds of trait data, such as some of the data sets mentioned in Section 1.3, or the morphological data sets of Glenner et al. [2004]. Alekseyenko et al. [2008] extended other aspects of the model to applications in genetics.

Cognacy classes are born and die along the tree, as shown in Figure 2.6. A cognacy class is born when a new word appears in a language, which is cognate with no other word in the process. A birth event occurs when a completely new word appears, but also when a word is borrowed from a language outside the study or when an existing word changes meaning. A cognate dies in a

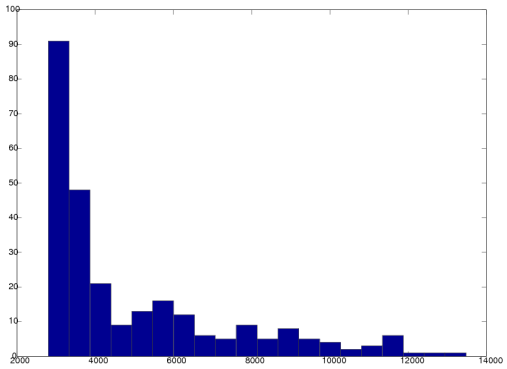

Figure 2.5: *Sample from the prior distribution on the age of the most recent common ancestor to the Italic languages.*

given language when it is no longer used for the meaning it was assigned to: it may completely cease to be used, or it may change meaning.

Cognates evolving in a single language (*i.e.* down a single branch of a language phylogeny) are born independently at rate λ, die independently at *per capita* rate μ, and are subject to point-like catastrophes, which they encounter at rate ρ along a branch. At a catastrophe, each cognate dies independently with probability κ, and a Poisson number of cognates with mean ν are born. A catastrophe corresponds to a brutal event, which might for example be a migration, an epidemic, or a large variation in population as those described by Shennan and Edinborough [2007] and Turney and Brown [2007]. At a branching event of the phylogeny, the set of cognates representing the branching vocabulary is copied into each of the daughter languages. See Figure 2.6.

The process we have described is not reversible, and this greatly complicates the analysis. We show in Section 2.5 that a necessary and sufficient condition

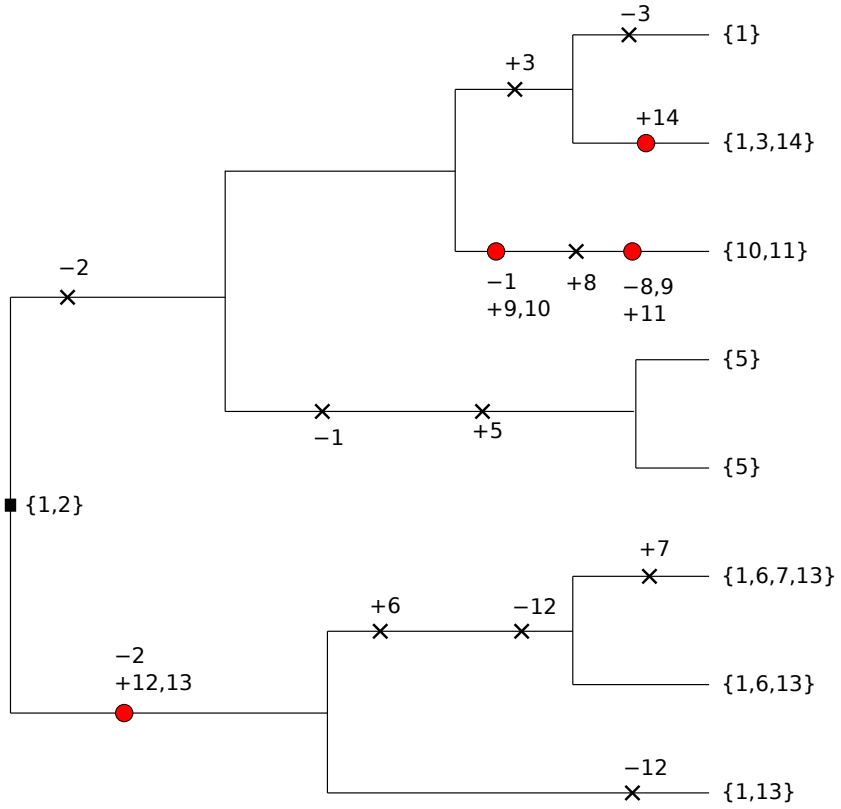

Figure 2.6: *Description of the model: births and deaths of cognacy classes are marked. The dots correspond to catastrophes, at which multiple births and deaths may occur simultaneously. The cognate sets this generates at leaves are shown on the right. Calendar time flows from left to right and the age variables t_i in the text increase from right to left. The root is represented by the square on the left.*

for the process to be time-reversible is $\nu = \kappa\lambda/\mu$. This is a reasonable modelling assumption: the anagenic process (without catastrophes) leads to a number of cognacy classes at any given point which is Poisson-Distributed with mean λ/μ. If we take $\nu = \kappa\lambda/\mu$, then at equilibrium[2], the distribution of the number of cognacy classes is unchanged by adding the catastrophe process. Under this condition, adding a catastrophe to an edge is equivalent

[2]Equilibrium has been reached since an infinite amount of time lapses on the Adam-root edge.

to lengthening that edge by $T_C(\kappa, \mu) = -\log(1-\kappa)/\mu$ years. This follows because the number of cognates generated by the anagenic part of the process in an interval of length T_C is Poisson distributed with mean $\frac{\lambda}{\mu}(1-e^{-\mu T_C})$ equal to $\kappa\lambda/\mu$, and the probability that a cognate entering an interval of length T_C dies during that interval is $1-e^{-\mu T_C}$, which equals κ.

Because a catastrophe simply extends its edge by a block of virtual time, the likelihood depends only on the number of catastrophes on an edge, and not their location in time. Let k_i be the number of catastrophes on edge $\langle i, j\rangle$, and $k = (k_1, \ldots, k_{2L-2})$ be the catastrophe state vector. We record no catastrophes on the $\langle r, A\rangle$ edge (its length is already infinite). The tree $g = (V, E, t, k)$ is specified by its topology, node ages and catastrophe state. Calibrated tree space extended for catastrophes is

$$\Gamma_K^C = \{(V, E, t, k) : (V, E, t) \in \Gamma^C, k \in \mathbb{N}_0^{2L-2}\}.$$

We drop the catastrophe process from the calculation in Section 2.3. It is straightforward to restore it, and we do this in the expression for the posterior distribution in Section 2.4.

2.2.3 The registration process

When linguists collect lexical data, some data are missing or are otherwise discarded through the registration process described in this section.

At leaf i, each data point is missing with probability ξ_i; we assume this probability depends only on the language i and not on the cognacy class. Data are usually missing in ancient languages which are only partially reconstructed,

[D*]	[I*]	[D~]	[I]	[D]
10000000000000	11111111111111	10000000000000	1 111 1 11 1	000 0 00
10100000000001	01011111111111	?0?00000000001	0 111 1 11 ?	000 0 01
00000000011000	10111001110110	0?000??001?00?	1 100 1 10 0	0?? 1 0?
00001000000000	11111111111111	00001000000000	1 111 1 11 0	100 0 00
00001000000000	11110111111111	0000?000000000	1 011 1 11 0	?00 0 00
10000110000010	10111111111111	1?000110000010	1 111 1 11 1	011 0 10
10000100000010	11111111111111	10000100000010	1 111 1 11 1	010 0 10
10000000000010	11111111111100	100000000000??	1 111 1 00 1	000 0 ??

Figure 2.7: Registration of the vocabulary realized in Figure 2.6 supposing a masking matrix I^*, as above. D^* is the unobserved full data with a column for each cognacy class and a row for each language of Figure 2.6 (thus cognate 1 is present in rows 1,2,6,7 and 8); zeros in the masking matrix I^* indicate missing matrix elements. Some cognacy classes are then thinned (in this example, the registration rule keeps cognacy classes with instances displayed in one or more language) to give the registered data D.

but we also observe small amounts of missing data in some modern languages.

Let $\mathbf{D}^*$ denote a notional *full* random binary data matrix, representing the outcome of the diversification process of Section 2.2.2. The number of columns in $\mathbf{D}^*$ is random, and equal to N^*. For the realization depicted in Figure 2.6, $\mathbf{D}^* = D^*$ with D^* displayed in Figure 2.7.

A column of D^* corresponds to a cognacy class which was either present at the root or was born below it, including some cognacy classes which died out completely. The number of columns N^* follows a $Poisson(|g| + k_T \cdot T_C)$ distribution, where $|g|$ is the total length of the tree, k_T is the total number of catastrophes on the tree, and T_C is the amount of time equivalent to one catastrophe.

The observed data D are a ternary matrix of 1's, 0's and ?'s with N columns, $N \leq N^*$. We call the mapping of the unknown complete data D^*

to the observed data D the registration process. This process occurs in two steps: first, some data go missing; second, some cognacy classes are removed from the data.

Let $\mathbf{I}^*$ be a random $L \times N^*$ indicator matrix of independent Bernoulli random variables for observed elements, such that for $i = 1 \ldots, L$ and $a = 1, \ldots, N^*$, $P[\mathbf{I}^*_{i,a} = 1] = \xi_i$, so that ξ_i is the probability that we can answer the question "Does language i display an instance of cognacy class a?". If we get an answer, it is assumed correct, and $\mathbf{I}^*_{i,a} = 1$; if we do not get an answer, then $\mathbf{I}^*_{i,a} = 0$: the zeros of $\mathbf{I}^*$ indicate which data points are missing. Let $\xi = (\xi_1, \ldots, \xi_L)$ and denote by I^* a realization of $\mathbf{I}^*$.

Now let $\tilde{D} = \tilde{D}(D^*, I^*)$ be the masked version of the full random data matrix: if $I^*_{i,a} = 1$ then $\tilde{D}_{i,a} = D^*_{i,a}$ and if $I^*_{i,a} = 0$ then $\tilde{D}_{i,a} =$?. The mapping of D^* to $\tilde{D}$ corresponds to the first step of the registration process.

In the second step, some columns are removed from $\tilde{D}$. For example, the second and third columns of $\tilde{D}$ consist entirely of 0's and ?'s: the corresponding cognacy classes were never observed at any leaf. Such cognacy classes are not included in the registered data. Other types of columns may also be removed from the data because they are deemed unreliable. Denote by R the registration rule $\mathbf{D} = R(\tilde{\mathbf{D}})$ mapping the full data to registered data. Let Y and Q be functions of the columns of D^* and I^* counting the visible 1's and ?'s respectively,

$$\begin{aligned} Y(\mathbf{D}^*_a, \mathbf{I}^*_a) &= \sum_{i=1}^{L} \mathbf{I}^*_{i,a} \mathbf{D}^*_{i,a}, \\ Q(\mathbf{I}^*_a) &= \sum_{i=1}^{L} (1 - \mathbf{I}^*_{i,a}). \end{aligned}$$

Given $a = 1, 2, ..., N^*$, let $Y_a = Y(\mathbf{D}_a^*, \mathbf{I}_a^*)$ and $Q_a = Q(\mathbf{I}_a^*)$.

We give an efficient algorithm for computing the likelihood for rules formed by compounding the following elementary thinning operations:

(1) $R_1(\tilde{D}) = (\tilde{D}_a : Y_a > 0)$ (discard classes with no instances at the leaves);

(2) $R_2(\tilde{D}) = (\tilde{D}_a : Y_a > 1)$ (discard classes - singletons - observed at a single leaf);

(3) $R_3(\tilde{D}) = (\tilde{D}_a : Y_a < L)$ (discard classes which are observed at all leaves);

(4) $R_4(\tilde{D}) = (\tilde{D}_a : Y_a < L - 1)$ (discard classes which are observed at all leaves or at all leaves but one);

(5) $R_5(\tilde{D}) = (\tilde{D}_a : Y_a + Q_a < L)$ (discard classes which are potentially present at all leaves);

(6) $R_6(\tilde{D}) = (\tilde{D}_a : Y_a + Q_a < L - 1)$ (discard classes which are potentially present at all leaves or at all leaves but one).

We assume the chosen rule includes Condition (1). The rule $D = R(\tilde{D})$ with $R(\tilde{D}) = R_6 \circ R_2(\tilde{D})$ collects "parsimony informative" cognacy classes. Ronquist et al. [2005] give the likelihood for the finite-sites trait evolution model of Lewis [2001] for registration rules like (1-6). The selection of columns is something we have in general no control over: the column selection rule simply describes what happened at registration, as defined in advance by the linguist collecting the data. In the example in Figure 2.7, and in our analysis of the Ringe et al. [2002] data, we fit data registered with $R(\tilde{D}) = R_1(\tilde{D})$, and also use rule

R_2 to validate our results. The Dyen et al. [1997] data use registration rule $R(\tilde{D}) = R_2(\tilde{D})$.

Many other registration rules could be thought of, and the likelihood calculations below are correct for a wide variety of rules: our only requirement is that the event that a class is registered is independent of all events in all other cognacy classes. We give recursions for rules (1-6) since these are likely to be the most frequently used.

The thinning $D = R(\tilde{D})$ corresponds to the second step of the registration process and gives rise to the observed data. Let $\mathbf{I}$ be a $L \times N$ matrix containing those columns of $\mathbf{I}^*$ which survived the thinning. We observe the realization I of $\mathbf{I}$.

Column indices $a = 1, 2, ..., N^*$ are exchangeable. It is convenient to renumber the columns of D^*, I^* and $\tilde{D}$ after registration, so that $\tilde{D}_a = D_a$ and $I^*_a = I_a$ for $a = 1, 2, ..., N$. The information needed to evaluate Y_a and Q_a is available in the column D_a and set Ω_a representations. We write $Y(D_a) = Y(\Omega_a) = Y_a$ and $Q(D_a) = Q(\Omega_a) = Q_a$.

2.2.4 Point process of births for registered cognacy classes

Fix a catastrophe-free phylogeny $g \in \Gamma^C_K$, with $k = (0, 0, ..., 0)$, and let an edge $\langle i, j \rangle$ and a time $\tau \in [t_i, t_j)$ be given. Denote by $[g]$ the set of all points (τ, i) on the phylogeny, including points (τ, r) with $\tau \geq t_r$ in the edge $\langle r, A \rangle$. The locations $z_D = \{z_1, z_2, ..., z_N\}$ of the birth events of the N registered cognacy classes are a realization of an inhomogeneous Poisson point process Z_D on $[g]$. Let $Z \in [g]$ be the birth location of a generic cognacy class $M \subseteq \{1, 2, ..., L\}$,

corresponding to a column of $\tilde{\mathbf{D}}$ with Y observed 1's and Q ?'s, and let $\mathcal{E}_Z$ be the event that this class generates a column of the registered data.

The point process Z_D of birth locations of registered cognacy classes has intensity

$$\tilde{\lambda}(z) = \lambda \Pr(\mathcal{E}_Z | g, \mu, \lambda, \xi, Z = z)$$

at $z \in [g]$ and probability density

$$f_{Z_D}(z_D) = \frac{1}{N!} e^{-\Lambda([g])} \prod_{a=1}^{N} \tilde{\lambda}(z_a)$$

with respect to the element of volume $dz_D = dz_1 dz_2 ... dz_N$ in $[g]^N$, where

$$\begin{aligned} \Lambda([g]) &= \int_{[g]} \tilde{\lambda}(z) dz \\ &= \sum_{\langle i,j \rangle \in E} \int_{t_i}^{t_j} \tilde{\lambda}((\tau, i)) d\tau. \end{aligned}$$

The number N of registered cognacy classes is $N \sim \text{Poisson}(\Lambda([g]))$.

2.3 Likelihood calculations

We give the likelihood for g, μ, λ, κ, ρ and ξ given the data, along with an efficient algorithm to compute the sum over all missing data.

We need to compute $P[D|g, \mu, \lambda, \xi, R(D)]$, the likelihood of the observed data given the tree g, the birth and death rates and λ and μ, the observation model parameters ξ, and the event that the traits we consider have been registered. We restore the birth locations (and so omit λ from the conditioning),

and factorize using the joint independence of $D_a,\ a = 1, 2, ..., N$, under the given conditions.

$$
\begin{aligned}
&P[\mathbf{D}=D|g,\mu,\lambda,\xi,\mathbf{D}=R(\tilde{\mathbf{D}})] \\
&\quad = \int f_{Z_D}(z_D) P[\mathbf{D}=D|g,\mu,\xi,Z_D=z_D,\mathbf{D}=R(\tilde{\mathbf{D}})]\, dz_D \\
&\quad = \frac{e^{-\Lambda([g])}}{N!} \prod_{a=1}^{N} \int_{[g]} \tilde{\lambda}(z_a) P[\mathbf{D}_a=D_a|g,\mu,\xi,Z_a=z_a,\mathcal{E}_{Z_a}]\, dz_a \\
&\quad = \frac{e^{-\Lambda([g])}}{N!} \prod_{a=1}^{N} \int_{[g]} \lambda P[\mathcal{E}_{Z_a}|g,\mu,\xi,Z_a=z_a]\, P[\mathbf{D}_a=D_a|g,\mu,\xi,Z_a=z_a,\mathcal{E}_{Z_a}]\, dz_a \\
&\quad = \frac{e^{-\Lambda([g])}}{N!} \prod_{a=1}^{N} \lambda \int_{[g]} P[\mathbf{D}_a=D_a,\mathcal{E}_{Z_a}|g,\mu,\xi,Z_a=z_a]\, dz_a. \\
&\quad = \frac{e^{-\Lambda([g])}}{N!} \prod_{a=1}^{N} \lambda \int_{[g]} P[\mathbf{D}_a=D_a|g,\mu,\xi,Z_a=z_a]\, dz_a.
\end{aligned}
$$

The last line follows because all traits in the data have been registered, hence the event $\{\mathbf{D}_a = D_a\}$ is a subset of the event $\mathcal{E}_{Z_a}$. The likelihood depends on the awkward condition $\mathbf{D} = R(\tilde{\mathbf{D}})$ only through the mean number $\lambda([g])$ of registered cognacy classes. The calculation has so far extended Nicholls and Gray [2008] to give the likelihood for a greater variety of column thinning rules. We now add the missing element component of the registration process.

We sum over possible values of the missing matrix elements in the registered data, *i.e.* other all elements in $\mathcal{D}$, the set of possible values for $\tilde{D}$ given D. Since $P[\mathbf{D}_a = D_a|g,\mu,\lambda,\xi,Z_a = z_a]$ is not conditioned on the requirement that the column D_a gets registered, the entries of the corresponding column I_a are

determined by the unconditioned Bernoulli process, and we have

$$\begin{aligned} P[\mathbf{D}_a = D_a | g, \mu, \xi, Z_a = z_a] &= \sum_{d^* \in \mathcal{D}_a} P[\mathbf{I}^*_a, \mathbf{D}^*_a = d^* | g, \mu, \xi, Z_a = z_a] \\ &= \prod_{i=1}^{L} \xi_i^{I_{a,i}} (1-\xi_i)^{1-I_{a,i}} \sum_{d^* \in \mathcal{D}_a} P[\mathbf{D}^*_a = d^* | g, \mu, \xi, Z_a = z_a]. \end{aligned}$$

The likelihood is

$$\begin{aligned} &P[\mathbf{D} = D | g, \mu, \lambda, \xi, \mathbf{D} = R(\tilde{\mathbf{D}})] = \\ &\quad \frac{e^{-\Lambda([g])}}{N!} \prod_{a=1}^{N} (\prod_{i=1}^{L} \xi_i^{I_{a,i}} (1-\xi_i)^{1-I_{a,i}}) \lambda \int_{[g]} \sum_{\omega \in \Omega_a} P[M = \omega | g, \mu, \xi, Z = z_a] dz_a, \quad (2.1) \end{aligned}$$

where we have switched from summing $d^* \in \mathcal{D}_a$ to the equivalent set representation $\omega \in \Omega_a$.

For the two integrated quantities in Equation (2.1) we have tractable recursive formulae. We are using a pruning procedure akin to Felsenstein [1981]. We begin with $\Lambda([g])$.

We make the reasonable assumption that the registration rule includes at least Condition R_1 from Section 2.2.3. It follows that a cognacy class born at $Z = (\tau, i)$ in $[g]$ must survive down to the node below, at $Z = (t_i, i)$, in order to be registered, and so

$$P[\mathcal{E}_Z | Z = (\tau, i), g, \mu, \xi] = P[\mathcal{E}_Z | Z = (t_i, i), g, \mu, \xi] e^{-\mu(\tau - t_i)}.$$

We can substitute this into the expression for $\Lambda([g])$, and integrate, to get

$$\Lambda([g]) = \frac{\lambda}{\mu} \sum_{\langle i,j \rangle \in E} P[\mathcal{E}_Z | Z = (t_i, i), g, \mu, \xi] \left(1 - e^{-\mu(t_j - t_i)}\right). \tag{2.2}$$

Given a node i, let $V_L^{(i)}$ be the set of leaf nodes descended from i, including i itself if i is a leaf. Let $s_i = \text{card}(V_L^{(i)})$. Denote by $u_i^{(n)} = P[Y = n | Z = (t_i, i), g, \mu, \xi]$ and $v_i^{(n)} = P[Y + Q = n | Z = (t_i, i), g, \mu, \xi]$. We can compute $\Lambda([g])$ for rules made up of combinations of Condition R_1 with any combination of Conditions (R_2-R_6), from $u_i^{(0)}$, $u_i^{(1)}$, $u_i^{(s_i-1)}$, $u_i^{(s_i)}$, $v_i^{(s_i-1)}$ and $v_i^{(s_i)}$. For example,

$$P[\mathcal{E}_Z | Z = (t_i, i), g, \mu, \xi] = \begin{cases} 1 - u_i^{(0)} & R = R_1, \\ 1 - u_i^{(0)} - u_i^{(1)} & R = R_2, \\ 1 - u_i^{(0)} - u_i^{(1)} - u_i^{(L-1)} - u_i^{(L)} & R = R_4 \circ R_2. \end{cases} \tag{2.3}$$

Notice that $u_i^{(n)} = 0$ unless $s_i \geq n$, so for example $u_i^{(L)}$ is non-zero at $i = r$ the root node only.

For nodes i and j, let $\delta_{i,j} = e^{-\mu(t_j - t_i)}$ be the probability for a cognate class present at (t_j, j) to survive down to (t_i, i). Consider a pair of edges $\langle c_1, i \rangle$,

$\langle c_2, i\rangle$ in E. Then

$$
\begin{aligned}
u_i^{(0)} &= \left((1-\delta_{i,c_1}) + \delta_{i,c_1} u_{c_1}^{(0)}\right)\left((1-\delta_{i,c_2}) + \delta_{i,c_2} u_{c_2}^{(0)}\right) \\
u_i^{(1)} &= \delta_{i,c_1}(1-\delta_{i,c_2}) u_{c_1}^{(1)} + \delta_{i,c_2}(1-\delta_{i,c_1}) u_{c_2}^{(1)} + \delta_{i,c_1}\delta_{i,c_2}(u_{c_1}^{(1)} u_{c_2}^{(0)} + u_{c_1}^{(0)} u_{c_2}^{(1)}) \\
u_i^{(s_i)} &= \delta_{i,c_1} u_{c_1}^{(s_{c_1})} \delta_{i,c_2} u_{c_2}^{(s_{c_2})} \\
u_i^{(s_i-1)} &= \left(\delta_{i,c_1} u_{c_1}^{(s_{c_1}-1)} + \mathbb{I}_{\{s_{c_1}=1\}}(1-\delta_{i,c_1})\right) \delta_{i,c_2} u_{c_2}^{(s_{c_2})} \\
&\quad + \delta_{i,c_1} u_{c_1}^{(s_{c_1})} \left(\delta_{i,c_2} u_{c_2}^{s_{c_2}-1} + \mathbb{I}_{\{s_{c_2}=1\}}(1-\delta_{i,c_2})\right) \\
v_i^{(0)} &= \left(\delta_{i,c_1} v_{c_1}^{(0)} + (1-\delta_{i,c_1}) \prod_{j \in V_L^{c_1}} \xi_j\right)\left(\delta_{i,c_2} v_{c_2}^{(0)} + (1-\delta_{i,c_2}) \prod_{j \in V_L^{c_2}} \xi_j\right) \\
v_i^{(s_i)} &= \left(\delta_{i,c_1} v_{c_1}^{(s_{c_1})} + (1-\delta_{i,c_1}) \prod_{j \in V_L^{c_1}} (1-\xi_j)\right)\left(\delta_{i,c_2} v_{c_2}^{(s_{c_2})} + (1-\delta_{i,c_2}) \prod_{j \in V_L^{c_2}} (1-\xi_j)\right) \\
v_i^{(s_i-1)} &= \left(\delta_{i,c_1} v_{c_1}^{(s_{c_1}-1)} + (1-\delta_{i,c_1}) \sum_{j \in V_L^{c_1}} \xi_j \prod_{k \neq j} (1-\xi_k)\right)\left(\delta_{i,c_2} v_{c_2}^{(s_{c_2})} + (1-\delta_{i,c_2}) \prod_{j \in V_L^{c_2}} (1-\xi_j)\right) \\
&+ \left(\delta_{i,c_1} v_{c_1}^{(s_{c_1})} + (1-\delta_{i,c_1}) \prod_{j \in V_L^{c_1}} (1-\xi_j)\right)\left(\delta_{i,c_2} v_{c_2}^{(s_{c_2}-1)} + (1-\delta_{i,c_2}) \sum_{j \in V_L^{c_2}} \xi_j \prod_{k \neq j} (1-\xi_k)\right)
\end{aligned}
$$

The recursion is evaluated from the leaves $i \in V_L$, at which

$$
\begin{aligned}
u_i^{(0)} &= u_i^{(s_i-1)} = 1 - \xi_i \\
u_i^{(1)} &= u_i^{(s_i)} = \xi_i \\
v_i^{(0)} &= v_i^{(s_i-1)} = 0 \\
v_i^{(s_i)} &= 1
\end{aligned}
$$

We now give the equivalent recursions for $\lambda \int_{[g]} \sum_{\omega \in \Omega_a} P[M = \omega_a | Z = z_a, g, \mu]\, dz_a$. Consider the set $m_a = \bigcap_{\omega \in \Omega_a} \omega$ of leaves *known* to have a cognate

in the a'th registered cognacy class (m_a is the set of leaves i such that $D_{i,a} = 1$). Let E_a be the set of branches on the path from the most recent common ancestor of the leaves in m_a up to the Adam-node A above the root. Cognacy class a must have been born on an edge in E_a. For $a = 1, 2, ..., N$, class M_a is non-empty, so a must have survived from its birth point down to the node below. We can shift the birth location to the node below and convert the integral to a sum,

$$\lambda \int_{[g]} \sum_{\omega \in \Omega_a} P[M = \omega | Z = z_a, g, \mu] \, dz_a = \frac{\lambda}{\mu} \sum_{\langle i,j \rangle \in E_a} \sum_{\omega \in \Omega_a} P[M = \omega | Z = (t_i, i), g, \mu](1 - \delta_{i,j}).$$

For each $a = 1, 2, ..., N$ and $\omega \in \Omega_a$, let $\omega^{(i)} = \omega \cap V_L^{(i)}$ and

$$\Omega_a^{(i)} = \{\omega^{(i)} : \omega^{(i)} = \omega \cap V_L^{(i)}, \omega \in \Omega_a\}$$

denote the set of all subsets $\omega^{(i)}$ of the leaves $V_L^{(i)}$ which are cognacy classes consistent with the data available for those leaves. Consider two child branches $\langle c_1, i \rangle$ and $\langle c_2, i \rangle$ at node i. Since $\Omega_a = \Omega_a^{(c_1)} \times \Omega_a^{(c_2)}$, and we have assumed that events are independent along the two branches,

$$\begin{aligned} \sum_{\omega \in \Omega_a} P[M = \omega | Z = (t_i, i), g, \mu, \xi] \;\; = \;\; & \sum_{\omega^{(c_1)} \in \Omega_a^{(c_1)}} P[M = \omega^{(c_1)} | Z = (t_i, c_1), g, \mu] \\ & \times \sum_{\omega^{(c_2)} \in \Omega_a^{(c_2)}} P[M = \omega^{(c_2)} | Z = (t_i, c_2), g, \mu]. \end{aligned}$$

Having moved the birth event at (t_i, i) to (t_i, c_1) and (t_i, c_2) (off the node and onto its child edges) we now move the birth event at (t_i, c) to (t_c, c) for $c = c_1, c_2$

(down an edge) as follows:

$$\sum_{\omega\in\Omega_a^{(c)}} P[M=\omega|Z=(t_i,c),g,\mu] =$$

$$\begin{cases} \delta_{i,c}\times\displaystyle\sum_{\omega\in\Omega_a^{(c)}} P[M=\omega|Z=(t_c,c),g,\mu] & \text{if } Y(\Omega_a^{(c)})\geq 1 \\ (1-\delta_{i,c})+\delta_{i,c}\times\displaystyle\sum_{\omega\in\Omega_a^{(c)}} P[M=\omega|Z=(t_c,c),g,\mu] & \text{if } Y(\Omega_a^{(c)})=0 \text{ and } Q(\Omega_a^{(c)})\geq 1 \\ (1-\delta_{i,c})+\delta_{i,c}v_c^{(0)} & \text{if } Y(\Omega_a^{(c)})+Q(\Omega_a^{(c)})=0 \\ & (i.e.\ \Omega_a^{(c)}=\{\emptyset\}) \end{cases}$$

The recursion is evaluated from the leaves. If c is a leaf, then

$$\sum_{\omega\in\Omega_a^{(c)}} P[M=\omega|Z=(t_c,c),g,\mu] = \begin{cases} 1 & \text{if } \Omega_a^{(c)}=\{\{c\},\emptyset\} \text{ or } \{\{c\}\}\ (i.e.\ D_{c,a}\in\{?,1\}) \\ 0 & \text{if } \Omega_a^{(c)}=\{\emptyset\}\ (i.e.\ D_{c,a}=0). \end{cases}$$

In order to restore catastrophes to this calculation, and given $g\in\Gamma_K$, with k_i catastrophes on edge $\langle i,j\rangle\in E$, replace t_j-t_i with $t_j-t_i+k_iT_C(\kappa,\mu)$ and all conditions on μ with conditions on μ,κ throughout.

2.4 Posterior distribution

Our prior distribution on the catastrophe death probability κ and on each of the missing data parameters $\xi_i,\ i=1\ldots L$ is a uniform distribution on the interval $[0,1]$. For the rate parameters μ, λ and ρ, we impose an improper prior distribution $p(\mu,\lambda,\rho)\propto\frac{1}{\mu\lambda\rho}$. This prior is scale-invariant. If we scale all the times t by a factor η ($t'=\eta t$), then the scaling $(\mu',\lambda',\rho')=(\mu/\eta,\lambda/\eta,\rho/\eta)$

leaves the likelihood unchanged. It is therefore reasonable to put the same prior weight for each rate parameter on an interval $[a, b]$ as on the interval $[a/\eta, b/\eta]$. In some analyses, we used a $\Gamma(1.5, 0.0002)$ prior on ρ. The 95% highest probability density interval for this distribution corresponds to catastrophes occuring between every 1,300 years and every 28,000 years; changing the prior did not affect our results. Our prior on the tree g is described in Section 2.2.1.

We take a uniform prior over $[0, 1]$ for the death probability at a catastrophe κ and each missing data parameter ξ_i.

Substituting using equations (2.2)-(2.3) into equation (2.1) and multiplying by the prior $f_G(g|T)p(\lambda, \mu, \rho)$, we obtain the posterior distribution

$$
\begin{aligned}
& p(g, \mu, \lambda, \kappa, \rho, \xi | \mathbf{D} = D) \\
&= \frac{1}{N!}\left(\frac{\lambda}{\mu}\right)^N \exp\left(-\frac{\lambda}{\mu}\sum_{\langle i,j\rangle \in E} P[\mathcal{E}_Z | Z = (t_i, i), g, \mu, \kappa, \xi](1 - e^{-\mu(t_j - t_i + k_i T_C)})\right) \\
&\quad \times \prod_{a=1}^{N}\left(\sum_{\langle i,j\rangle \in E_a}\sum_{\omega \in \Omega_a} P[M = \omega | Z = (t_i, i), g, \mu](1 - e^{-\mu(t_j - t_i + k_i T_C)})\right) \\
&\quad \times \frac{1}{\mu\lambda} p(\rho) f_G(g|T) \frac{e^{-\rho|g|}(\rho|g|)^{k_T}}{k_T!} \prod_{i=1}^{L}(1 - \xi_i)^{Q_i}\xi_i^{N - Q_i} \qquad (2.4)
\end{aligned}
$$

for parameters $\mu, \lambda, \rho > 0$, $0 \leq \kappa, \xi_i \leq 1$ and trees $g \in \Gamma_K^{(C)}$.

With the prior $p(\rho) \propto 1/\rho$, the posterior is improper without bounds on ρ since $k_T = 0$ is allowed. We place very conservative bounds on ρ. Results are not sensitive to this choice. We discuss the propriety of the posterior in Section 3.1.

2.5 Time reversibility

In this section, we show that the process is time-reversible if and only if $\nu = \kappa\lambda/\mu$, as claimed in section 2.2. For this, we look at the transition rates $r_{i,j}$ from the state with i cognacy classes in a language to the state with j cognacy classes. These transition rates are the sum of the transition rates for the anagenic process, which allows transitions from i to $i+1$ and $i-1$, and the transition rates for the catastrophe process, which allows transitions from any i to any j.

- for $|i-j| \neq 1$, the transition from i to j has to go through a catastrophe (these occur at rate ρ). For a catastrophe occurring at time τ, say that the deaths happen at τ and the births happen shortly after, at $\tau - \epsilon$; let k be the number of cognacy classes existing after the deaths have occurred, but before the births. Only deaths occurred to go from i to k, so $k \leq i$, and only births occur to go from k to j, so $k \leq j$; k can take any value between 0 and $\min(i,j)$. Summing over all these values, we get:

$$r_{i,j} = \rho \sum_{k=0}^{k=\min(i,j)} Bin(k; i, 1-\kappa) \times Poi(j-k; \nu) \tag{2.5}$$

(starting with i cognacy classes, k cognacy classes have to survive the thinning process with survival probability $1-\kappa$; then the remaining $j-k$ cognacy classes are born through the $Poi(\nu)$ process). This becomes

$$r_{i,j} = \rho\kappa^i e^{-\nu}\nu^j i!\ {}_2F_0\left(-i, -j; \frac{1-\kappa}{\nu\kappa}\right) \tag{2.6}$$

where

$$ {}_2F_0(-i,-j;\theta) = \sum_{k=0}^{k=\min(i,j)} \theta^k \frac{1}{k!(i-k)!(j-k)!} $$

is a generalized hypergeometric function [Abramowitz and Stegun, 1964]. Note that ${}_2F_0(-i,-j;\theta) = {}_2F_0(-j,-i;\theta)$.

- for $j = i+1$, the transition rates are

$$ r_{i,j} = \lambda + \rho e^{-\nu} \frac{\nu^j \kappa^{j-1}}{j!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right) \quad (2.7) $$

$$ r_{j,i} = \mu j + \rho e^{-\nu} \frac{\nu^{j-1}\kappa^j}{(j-1)!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right). \quad (2.8) $$

First, assume that $\nu = \kappa\lambda/\mu$. We have proven in Section 2.2.2 that at equilibrium, the probability for any point on the tree to display exactly i cognate classes is $\pi_i = e^{-\frac{\nu}{\kappa}}\nu^i/(i!\kappa^i)$. If $|i-j| \neq 1$, it is straightforward to check that $\pi_i r_{i,j} = \pi_j r_{j,i}$. If $j = i+1$, then

$$ \begin{aligned} \frac{\pi_j r_{j,i}}{\pi_i r_{i,j}} &= \frac{\lambda}{\mu}\frac{1}{j} \times \frac{\mu j + \rho e^{-\nu}\frac{\nu^{j-1}\kappa^j}{(j-1)!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \rho e^{-\nu}\frac{\nu^j\kappa^{j-1}}{j!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)} \\ &= \frac{\lambda}{\mu} \times \frac{\mu + \rho e^{-\nu}\frac{\nu^{j-1}\kappa^j}{j!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \rho e^{-\nu}\frac{\nu^j\kappa^{j-1}}{j!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)} \\ &= \frac{\frac{\lambda}{\mu}\mu + \frac{\nu}{\kappa}\rho e^{-\nu}\frac{\nu^{j-1}\kappa^j}{j!} \, {}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \rho e^{-\nu}\frac{\nu^j\kappa^{j-1}}{j!} \, {}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)} \\ &= 1. \end{aligned} $$

For all i and j, $\pi_i r_{i,j} = \pi_j r_{j,i}$ and so the process is time-reversible.

Suppose conversely that the process is time-reversible. Take $i \geq 2$ and let

$j = i + 1$. Then

$$\begin{aligned} \pi_i &= \frac{\pi_0 r_{0,i}}{r_{i,0}} \text{ (by time-reversibility)} \\ &= \frac{\nu^i}{\kappa^i}\frac{\pi_0}{i!} \text{ (by equation (2.6))} \qquad (2.9) \\ \pi_j &= \frac{\nu^j}{\kappa^j}\frac{\pi_0}{j!} \qquad (2.10) \end{aligned}$$

Hence $\pi_i/\pi_j = (i+1)\kappa/\nu$. Equations (2.7) and (2.8) give

$$\frac{r_{j,i}}{r_{i,j}} = \frac{\mu(i+1) + \frac{\rho e^{-\nu}\nu^i\kappa^{i+1}}{i!}\,{}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \frac{\rho e^{-\nu}\nu^{i+1}\kappa^i}{(i+1)!}\,{}_2F_0\left(-j,-i;\frac{1-\kappa}{\nu\kappa}\right)} \qquad (2.11)$$

Since the process is time reversible, we have $\frac{\pi_i}{\pi_j} = \frac{r_{j,i}}{r_{i,j}}$, *i.e.*

$$(i+1)\frac{\kappa}{\nu} = \frac{\mu(i+1) + \frac{\rho e^{-\nu}\nu^i\kappa^{i+1}}{i!}\,{}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)}{\lambda + \frac{\rho e^{-\nu}\nu^{i+1}\kappa^i}{(i+1)!}\,{}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)} \qquad (2.12)$$

Dividing the numerator of both sides by $(i+1)\kappa$ and the denominator of both sides by ν gives

$$1 = \frac{\frac{\mu}{\kappa} + \frac{\rho e^{-\nu}\nu^i\kappa^i}{(i+1)!}\,{}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)}{\frac{\lambda}{\nu} + \frac{\rho e^{-\nu}\nu^i\kappa^i}{(i+1)!}\,{}_2F_0\left(-i,-j;\frac{1-\kappa}{\nu\kappa}\right)}. \qquad (2.13)$$

It follows that $\mu/\kappa = \lambda/\nu$.

This shows that the process is time reversible if and only if $\nu = \lambda\kappa/\mu$.

Chapter 3

Implementation

This chapter checks the posterior probability distribution is proper, sets out the Markov Chain Monte Carlo algorithm, explains how it is implemented, describes debugging checks and briefly lists other aspects of the software TraitLab in which Geoff Nicholls, David Welch and this author have implemented the stochastic Dollo model described in Chapter 2.

3.1 Propriety of the posterior

Recall that our prior on the birth parameter λ and the death parameter μ is $p(\lambda, \mu) \propto \frac{1}{\lambda\mu}$, which is improper. It is therefore possible that the posterior distribution is improper. Equation 2.4 gives the posterior distribution of the

tree and parameters given the data as

$$
\begin{aligned}
& p(g,\mu,\lambda,\kappa,\rho,\xi|\mathbf{D}=D) \\
&= \frac{1}{N!}\left(\frac{\lambda}{\mu}\right)^N \exp\left(-\frac{\lambda}{\mu}\sum_{\langle i,j\rangle\in E} P[\mathcal{E}_Z|Z=(t_i,i),g,\mu,\kappa,\xi](1-e^{-\mu(t_j-t_i+k_iT_C)})\right) \\
&\times \quad \prod_{a=1}^{N}\left(\sum_{\langle i,j\rangle\in E_a}\sum_{\omega\in\Omega_a} P[M=\omega|Z=(t_i,i),g,\mu](1-e^{-\mu(t_j-t_i+k_iT_C)})\right) \\
&\times \quad \frac{1}{\mu\lambda}p(\rho)f_G(g|T)\frac{e^{-\rho|g|}(\rho|g|)^{k_T}}{k_T!}\prod_{i=1}^{L}(1-\xi_i)^{Q_i}\xi_i^{N-Q_i}
\end{aligned}
$$

for parameters $\mu,\lambda,\rho>0$, $0\le\kappa,\xi_i\le 1$ and trees $g\in\Gamma_K^{(C)}$.

We give conditions under which the posterior can be shown to be proper. Our strategy is to first integrate out λ, and then find an upper bound on the integral of the posterior when $\mu\to\infty$ and $\mu\to 0$. More specifically, we shall find μ_0 and μ_1 such that there is a bound for the integral on $[\mu_0,\infty)$ and a bound on the integral on $[0,\mu_1]$. (The integral on $[\mu_1,\mu_0]$ is the integral of a bounded function over a compact set and is therefore finite.) For ease of readability, we let $\theta=(g,\kappa,\rho,\xi)$ and $d\theta=d\kappa d\rho d\xi\prod_i dt_i$ (with counting measure on topologies).

The marginal for λ is Gamma-distributed: $\lambda|\mu,\theta,D\sim\Gamma(N,\beta)$, where

$$
\beta \;=\; \frac{\mu}{\sum_{\langle i,j\rangle\in E} P[\mathcal{E}_Z|Z=(t_i,i),\theta,\mu](1-e^{-\mu(t_j-t_i+k_iT_C)})}
$$

and the probability density function of a $\Gamma(\alpha,\beta)$ is $f(x)=\frac{x^{\alpha-1}e^{-\beta x}\beta^\alpha}{\Gamma(\alpha)}$

When we integrate out λ we pick up the factors normalizing this distribution:

$$\begin{aligned} p(\mu,\theta|D) \quad \propto \quad & \left(\sum_{\langle i,j\rangle \in E} P[\mathcal{E}_Z|Z=(t_i,i),\theta,\mu](1-e^{-\mu(t_j-t_i+k_iT_C)}) \right)^{-N} \\ & \times \prod_{a=1}^{N} \left(\sum_{\langle i,j\rangle \in E_a} \sum_{\omega \in \Omega_a} P[M=\omega|Z=(t_i,i),g,\mu](1-e^{-\mu(t_j-t_i+k_iT_C)}) \right) \\ & \times \frac{1}{N!}\frac{1}{\mu} p(\rho) f_G(g|T) \frac{e^{-\rho|g|}(\rho|g|)^{k_T}}{k_T!} \prod_{i=1}^{L} (1-\xi_i)^{Q_i} \xi_i^{N-Q_i} \end{aligned} \tag{3.1}$$

(and this is in fact the form we used in the implementation described later in this chapter).

We first examine the asymptotic behaviour when $\mu \to +\infty$. In order to prove that the posterior is proper, we will assume the data includes a trait displayed in two leaves in different clades. Such a trait must survive for a finite time, and t his is impossible at $\mu \to \infty$.

We rewrite Equation 3.1

$$p(\mu,\theta|D) \propto \frac{C}{\mu} \prod_{a=1}^{N} \frac{\sum_{\langle i,j\rangle \in E_a} \sum_{\omega \in \Omega_a} P[M=\omega|Z=(t_i,i),\theta,\mu](1-e^{-\mu(t_j-t_i+k_iT_C)})}{\sum_{\langle i,j\rangle \in E} P[\mathcal{E}_Z|Z=(t_i,i),\theta,\mu](1-e^{-\mu(t_j-t_i+k_iT_C)})} \tag{3.2}$$

where $C = \frac{\prod_{i=1}^{L}(1-\xi_i)^{Q_i}\xi_i^{N-Q_i}}{N!} p(\rho) f_G(g|T) \frac{e^{-\rho|g|}(\rho|g|)^{k_T}}{k_T!}$.

The numerator is bounded, so we get a bound for the Right Hand Side if we can find a lower bound for the denominator. Since for all $\omega \in \Omega_a$, the event $\{M_a = \omega|Z=(t_i,i)\}$ is contained in the event $\{\mathcal{E}_Z|Z=(t_i,i)\}$, and since $E_a \subseteq E$, all the terms in this product are less than 1, but this upper bound is not quite enough to show that the integral over μ is proper.

Let us now consider the case where the registration condition is $Y(a) > 0$

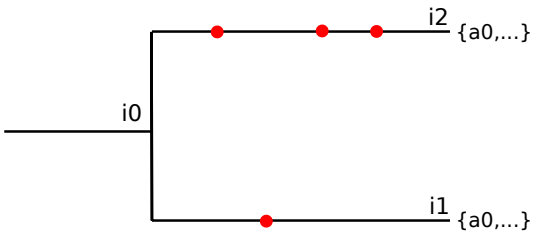

Figure 3.1: *An example tree for Eq. 3.3, where cognate class a_0 is displayed only at leaves i_1 and i_2, with respectively $k_1 = 1$ and $k_2 = 3$ catastrophes above them.*

(discard cognacy classes which are never observed; we call this Condition R_1 in Section 2.2.3). Suppose there is a cognacy class a_0 such that a_0 is displayed at exactly two leaves, i_1 and i_2.[1] For now, assume that i_1 and i_2 are siblings, as in Figure 3.1, let i_0 be their common parent node and let k_1 and k_2 be the number of catastrophes on the branches $\langle i_1, i_0\rangle$ and $\langle i_2, i_0\rangle$ respectively (see Figure 3.1). We can then obtain a lower bound on the denominator of Equation (3.2):

$$\begin{aligned}
&\sum_{\langle i,j\rangle \in E} P[\mathcal{E}_Z | Z = (t_i, i), \theta, \mu](1 - e^{-\mu(t_j - t_i + k_i T_C)}) \\
&\qquad = \sum_{\langle i,j\rangle \in E} P[Y(a) > 0 | Z = (t_i, i), \theta, \mu](1 - e^{-\mu(t_j - t_i + k_i T_C)}) \\
&\qquad \geq P[Y(a) > 0 | Z = (t_{i_1}, i_1), \theta, \mu](1 - e^{-\mu(t_{i_0} - t_{i_1} + k_1 T_C)}) \\
&\qquad \geq \xi_1 \times (1 - e^{-\mu(t_{i_0} - t_{i_1} + k_1 T_C)}) \\
&\qquad \geq \xi_1 (1 - e^{-\mu(t_{i_0} - t_{i_1})}) \qquad (3.3)
\end{aligned}$$

We now need an upper bound on the numerator of Equation (3.2). The double sum $\sum_{\langle i,j\rangle \in E_{a_0}} \sum_{\omega \in \Omega_{a_0}} P[M = \omega | Z = (t_i, i), \theta, \mu](1 - e^{-\mu(t_j - t_i + k_i T_C)})$

[1] In other words, $\Omega_{a_0} = \{\{i_1, i_2\}\}$. In this case, the double sum over E_a and Ω_a will therefore only be a sum over E_a, since there are no missing data for this cognacy class.

contains at most $L-2$ terms. Since $1-e^{-\mu(t_j-t_i+k_iT_C)} \leq 1$ and since the maximal value (over i) of the terms $P[M=\omega|Z=(t_i,i),\theta,\mu]$ clearly corresponds to a birth at the most recent common ancestor, i_0,

$$\begin{aligned}\sum_{\langle i,j\rangle\in E_{a_0}}\sum_{\omega\in\Omega_{a_0}} & P[M=\omega|Z=(t_i,i),\theta,\mu](1-e^{-\mu(t_j-t_i+k_iT_C)}) \\ &\leq (L-2)P[M=\{i_1,i_2\}|Z=(t_{i_0},i_0),\theta,\mu] \\ &\leq (L-2)\xi_{i_1}\xi_{i_2}e^{-\mu(t_{i_0}-t_{i_1}+k_1T_C)}e^{-\mu(t_{i_0}-t_{i_2}+k_2T_C)}.\end{aligned}$$

By using this upper bound for the term a_0, and the fact that all the terms in the product (3.2) for $a \neq a_0$ are each less than 1, we get that

$$p(\mu,\theta|D) \leq C(L-2)\xi_{i_1}\xi_{i_2}\frac{e^{-\mu(2t_{i_0}-t_{i_1}-t_{i_2}+(k_1+k_2)T_C)}}{\mu(1-e^{-\mu(t_{i_0}-t_{i_1})})} \tag{3.4}$$

A similar result can be obtained if i_1 and i_2 are not siblings.

We can now use that, for $\mu \geq 1$,

$$\frac{e^{-\mu(2t_{i_0}-t_{i_1}-t_{i_2}+(k_1+k_2)T_C)}}{\mu} \leq e^{-\mu(2t_{i_0}-t_{i_1}-t_{i_2}+(k_1+k_2)T_C)} \leq e^{-\mu(2t_{i_0}-t_{i_1}-t_{i_2})}.$$

Therefore,

$$\begin{aligned}p(\mu,\theta|D) &\leq C(L-2)\xi_{i_1}\xi_{i_2}\frac{e^{-\mu(2t_{i_0}-t_{i_1}-t_{i_2})}e^{\mu(t_{i_0}-t_{i_1})}}{e^{\mu(t_{i_0}-t_{i_1})}-1} &(3.5)\\ &\leq C(L-2)\xi_{i_1}\xi_{i_2}\frac{e^{-\mu(t_{i_0}-t_{i_2})}}{e^{t_{i_0}-t_{i_1}}-1} &(3.6)\end{aligned}$$

We can integrate μ out of $[1, \infty)$, and obtain

$$p(\theta|D)d\theta \leq C(L-2)\xi_{i_1}\xi_{i_2}\frac{e^{-(t_{i_0}-t_{i_2})}}{(t_{i_0}-t_{i_2})e^{(t_{i_0}-t_{i_1})}-1}d\theta \tag{3.7}$$

As long as we can put a lower bound on $t_{i_0} - t_{i_1}$ and $t_{i_0} - t_{i_2}$, this will be proper. Assuming that i_1 and i_2 are modern leaves (so that their age is constrained to be 0), such a lower bound can be obtained if there is a lower bound on t_{i_0}, *i.e.* if we have a constraint on the root of a clade containing i_1 but not i_2. The integral is therefore proper for $\mu \in [1, \infty[$ if at least one cognacy class is displayed in exactly two languages in two different clades. For example, in Figure 2.2, a cognacy class displayed in Latvian and Welsh would satisfy this condition.

This result can be extended to the case where a_0 is displayed at more than 2 leaves and where the data are missing at some leaves. In general, if the registration condition takes the form $Y(a) > d$ for some value of d, it can be extended to the case where there exists a cognacy class a_0 displayed at at least $d + 2$ leaves in two different clades. We suspect that this condition could be made less stringent, but it is already satisfied by our data. If the registration condition includes one of conditions R_3-R_6 of section 2.3 (*e.g.* $Y(a) < L$, discard any cognacy class displayed at every leaf), the calculation still holds, as long as i_0 has less than L descendants (or $L-1$, depending on the condition used). The case $\mu \to \infty$ is covered.

It is straightforward to show that, under registration condition R_1, the posterior distribution for data in which all cognacy classes are displayed at just one leaf is improper. We do not know what the propriety of the integral

is when cognacy classes are displayed at more than $d+1$ leaves, but with all the leaves in the same clade; we suspect it is improper.

We now consider the case $\mu \to 0$. With catastrophes included, the posterior is improper. Indeed, for a tree with one catastrophe on each branch and with fixed κ, the likelihood does not go to 0 when $\mu \to 0$, since such a tree is equivalent to a tree in which all branches have equal length $T_C = -\log(1-\kappa)/\mu$. We show that for a tree without catastrophes, and under reasonable conditions on the data, the posterior is proper when $\mu \to 0$. This time, the key is to assume a cognacy class which must have at least one death on the tree. Any cognacy class which is not monophylitic fulfills this criterion. We now define $\theta = (g, \xi)$ and look only at catastrophe-free trees.

Assume that the registration condition is of the form $Y(a) > d$ for some d. We shall use two properties of the exponential function:

$$\forall x \in \mathbb{R}, \quad 1 - e^{-x} \leq x \tag{3.8}$$

$$\forall x \in [0,1], \quad 1 - e^{-x} \geq \frac{x}{2} \tag{3.9}$$

These translate into

$$\forall \mu \in \mathbb{R}^+, \forall \langle i,j \rangle \in E, \qquad 1 - e^{-\mu(t_j - t_i)} \leq \mu(t_j - t_i) \tag{3.10}$$

$$\forall \mu \leq 1/T, \forall \langle i,j \rangle \neq \langle r, A \rangle, \qquad 1 - e^{-\mu(t_j - t_i)} \geq \frac{\mu(t_j - t_i)}{2} \tag{3.11}$$

(Recall that $\langle r, A \rangle$ is the Root-Adam branch, and T is the upper bound on the root age. The second equation holds because $\forall \langle i,j \rangle \neq \langle r, A \rangle,\ t_j - t_i \leq T$, hence $\mu(t_j - t_i) \leq 1$.)

Also, note that

$$\forall \langle i,j\rangle \in E,\ 1 \geq P[\mathcal{E}_Z | Z = (t_i, i), \theta, \mu] \geq e^{-\mu[(d+1)T]} \tag{3.12}$$

(the last term corresponds to a cognacy class present at the root, and surviving along $d+1$ paths of length T, which is one way of having the event $\mathcal{E}_Z$ occur). Equations (3.11) and (3.12) give a lower bound on the denominator of (3.2):

$$\begin{aligned} \sum_{\langle i,j\rangle \in E} P[\mathcal{E}_Z | Z = (t_i, i), \theta, \mu](1 - e^{-\mu(t_j - t_i)}) &\geq e^{-\mu[(d+1)T]} \cdot \frac{\mu}{2} \sum_{\langle i,j\rangle \neq \langle r,A\rangle} (t_j - t_i) \\ &\geq e^{-\mu[(d+1)T]} \frac{\mu}{2} t_r \end{aligned} \tag{3.13}$$

where t_r is the root age.

Given a tree g and a cognacy class a, we say that the cognacy class a is *monophylitic* if and only if there is a subtree h of g such that a is potentially displayed at all the leaves of h and nowhere else [2]. Now suppose that at least one cognacy class, a_m say, born at (t_i, i) is not monophylitic. Then at least one death must have occurred in the subtree below (t_i, i). The length of that subtree is at most $t_r \cdot L$, so for any $\omega \in \Omega_{a_m}$

$$P[M = \omega | Z = (t_i, i), \theta, \mu] \leq 1 - e^{-\mu t_r L} \leq \mu t_r L \tag{3.14}$$

This is true for all possible values of i for $\langle i,j\rangle \in E_{a_m}$, of which there are at most $L-1$, and for all possible values of $\omega \in \Omega_{a_m}$, of which there are at most

[2] For leaves in the subtree, the data can be 1 or ?; for leaves outside the subtree, the data can be 0 or ?.

2^{L-3}. Hence, using equation (3.10),

$$\begin{aligned}
&\sum_{\langle i,j\rangle\in E_{a_m}} \sum_{\omega\in\Omega_{a_m}} P[M=\omega|Z=(t_i,i),\theta,\mu](1-e^{-\mu(t_j-t_i)}) \\
&\qquad \leq \sum_{\langle i,j\rangle\in E_{a_m}} \sum_{\omega\in\Omega_{a_m}} P[M=\omega|Z=(t_i,i),\theta,\mu]\mu(t_j-t_i) \\
&\qquad \leq \mu(t_r L)\cdot \mu T\cdot (L-1)\cdot 2^{L-3} \qquad (3.15)
\end{aligned}$$

Coming back to the product in the Left Hand Side of Equation 3.2, the term for a_m in the product can be bounded:

$$\frac{\sum_{\langle i,j\rangle\in E_{a_m}} \sum_{\omega\in\Omega_{a_m}} P[M=\omega|Z=(t_i,i),\theta,\mu](1-e^{-\mu(t_j-t_i)})}{\sum_{\langle i,j\rangle\in E} P[\mathcal{E}_Z|Z=(t_i,i),\theta,\mu](1-e^{-\mu(t_j-t_i)})} \leq \frac{\mu(t_r L)\cdot \mu T\cdot (L-1)\cdot 2^{L-3}}{e^{-\mu[(d+1)T]\frac{\mu}{2}t_r}} \qquad (3.16)$$

As previously, every other term of this product is bounded by 1, hence Equation 3.2 becomes, for some constant C',

$$p(\theta,\mu|D) \leq C' \frac{\mu(L)\cdot T\cdot (L-1)\cdot 2^{L-3}}{e^{-\mu[(d+1)T]\frac{1}{2}}} \frac{1}{\mu} t_r^{2-L} f_G(g|T) \qquad (3.17)$$

which is proper for $\mu \in [0, 1/T]$.

Suppose now that there are two registration conditions: $Y(a) > d$ and $Y(a) < L - d'$ (for example, and using the notation from Section 2.2.3, $R = R_3 \circ R_2$). Let B be the set of edges which have at least $L - d'$ descendants (these are edges close to the root of the tree). Inequality (3.12) does not hold anymore for the edges in B.

Suppose that there is a clade g' such that:

- There are less than $L - d'$ leaves in g'

- We can put a lower bound T' on the root age of g' (this can be done through the age constraints we know for certain ancestral nodes)

Let E' be the set of edges in g'. Inequality (3.12) still holds for the elements of E'. Inequality (3.13) becomes

$$\begin{aligned}\sum_{\langle i,j\rangle\in E} P[\mathcal{E}_Z|Z=(t_i,i),g,\mu](1-e^{-\mu(t_j-t_i)}) &\geq \sum_{\langle i,j\rangle\in E'} P[\mathcal{E}_Z|Z=(t_i,i),g,\mu](1-e^{-\mu(t_j-t_i)}) \\ &\geq e^{-\mu[(d+1)T]}\cdot\frac{\mu}{2}\sum_{\langle i,j\rangle\in E'}(t_j-t_i) \\ &\geq e^{-\mu[(d+1)T]}\frac{\mu}{2}T' \qquad (3.18)\end{aligned}$$

A slightly modified version of inequality (3.17) shows that the integral is still proper.

A very similar argument holds for the case where the registration condition is of the form $Y(a)+Q(a)<L-d'$. The case $\mu\to 0$ is covered.

On the other hand, if all the cognacy classes are monophylitic, it is clear that the posterior distribution is improper when $\mu\to 0$.

To sum up: under the prior $p(\mu,\lambda)\propto\frac{1}{\mu\lambda}$, with catastrophes excluded, and with the registration condition $d<Y(a)<L-d'$, the following conditions are sufficient for the posterior to be proper:

1. there exists a cognacy class a_0 such that a_0 is displayed at at least $d+2$ different leaves in 2 different clades;

2. there exists a non-monophylitic cognacy class a_m;

3. there exists a clade g' such that there are less than $L - d'$ leaves in g, and we can put a lower bound T' on the root age of g'

These conditions are met by any realistic data. For example, in the Ringe et al. [2002] data, for which $R = R_1$ (*i.e.* $0 < Y(a)$, corresponding to $d = 0$),

1. The Latin *moritur* and the Old Church Slavonic *umĭretŭ* (both meaning "he dies") are cognate and in different clades.

2. This cognacy class is not monophylitic, since Old Prussian is constrained to be in the Balto-Slav clade with Old Church Slavonic but without Latin, and there is no instance of this class in Old Prussian (the only Old Prussian term for "he dies" is *aulaūt*, which is in a different cognacy class).

3. There are 8 clades which satisfy the last condition, for example the Germanic clade.

With catastrophes included, the posterior is improper when $\mu \to 0$, because then $T_C \to \infty$: the effective length can go to infinity, with all changes happening on catastrophes and no change occurring through the anagenic process. This is most easily resolved by imposing a lower bound on μ, but could also be resolved by imposing conditions on the catastrophes. However, this is not necessary in practice since we never observe $\mu \to 0$ in our analyses.

3.2 Implementation in MatLab

The stochastic Dollo model described in Chapter 2 was implemented in MatLab by Geoff Nicholls, David Welch, and this author, resulting in a piece of

public-domain software named TraitLab. The tree, catastrophes and model parameters are estimated jointly via Markov Chain Monte Carlo, as described in Section 3.3.

TraitLab takes files in Nexus format. It analyses the data under the stochastic Dollo model, with options to include some or all the clade constraints in the Nexus file, as well as to exclude specific traits or languages. The inclusion of catastrophes is optional, and the user may choose to handle missing data correctly or to treat missing data as absent (replacing all ?'s with 0's).

Data may be synthetized from the model directly from the Graphical User Interface (GUI). Most of the types of out-of-model data described in Chapter 4 may also be synthetized directly from the GUI, including data with borrowing and data missing in blocks.

The Analysis part of the GUI gives tools to check convergence and mixing of the MCMC, with autocorrelations plots for six statistics: the root age t_r, the death rate μ, the catastrophe parameters ρ and κ, the log-likelihood and the prior. Several tools for analysis of the data are included, including the ability to construct consensus trees for catastrophe models as described in Chapter 4. An option allows to save all results to HTML format for ease of sharing.

3.3 Markov Chain Monte Carlo

We use Markov Chain Monte Carlo (MCMC) to sample the posterior distribution and estimate summary statistics. We do not estimate λ, which we integrate out following Equation 3.1. The MCMC state is then $x = ((E, V, t, k), \mu, \kappa, \rho, \xi)$.

Given a prior distribution p and a likelihood function $L(\cdot|D)$, the strategy in Markov Chain Monte Carlo is to build a Markov chain whose stationary distribution is the posterior distribution we wish to sample from. Let x_n be the state of the Markov chain at step n; the state x_{n+1} is constructed as follows:

1. Propose to move from x_n to x', where x' is drawn from a proposal distribution $q(x'|x_n)$.

2. Compute
$$\alpha = \min\left(1, \frac{p(x')L(x'|D)q(x_n|x')}{p(x_n)L(x_n|D)q(x'|x_n)}\right).$$

3. Set $x_{n+1} = x'$ with probability α and $x_{n+1} = x_n$ with probability $1 - \alpha$.

Under weak conditions, this Markov chain converges to the posterior distribution.

For the proposal distribution, we use the MCMC moves described by Drummond et al. [2002] and Nicholls and Gray [2008], to which we add moves to take into account our additional parameters. Our complete list of MCMC moves is as follows (a star denotes moves added by the present author):

- Moves on the topology:
  - Exchange the positions of two close nodes (a node and its "niece", the child of its sibling)
  - Exchange the positions of any two nodes
  - Move a subtree to a close position
  - Move a subtree to any other position in the tree
- Moves on the ages:

  - Change the age of an internal node
  - Change the age of a non-modern leaf
  - Rescale the whole tree
  - Rescale a subtree
  - Rescale the top of the tree

- Moves on catastrophes:
  - Add a new catastrophe (*)
  - Delete a catastrophe (*)
  - Move a catastrophe to a neighbouring edge (*)

- Moves on the parameters:
  - Random walk (log scale) of the death rate μ
  - Random walk (log scale) of the catastrophe rate ρ (*)
  - Random walk on $[0, 1]$ for the catastrophe death probability κ (*)
  - Rescale the missing data parameter ξ_i for a single leaf i (*)
  - Rescale the entire vector of missing data parameters ξ (*)

The probability $0 < \xi_i < 1$ for an element of the registered data matrix to be observable is, for many leaves, close to one, so we update those parameters by scaling $1 - \xi_i$. We also include a move which scales all ξ_i $(i = 1 \dots L)$ simultaneously away from 1 by a factor of η, where $\eta \sim U([1/2, 2])$. In this case, the Jacobian for the transformation from $(\xi_1, \dots, \xi_L, \eta)$ to $(1 - \eta(1 -$

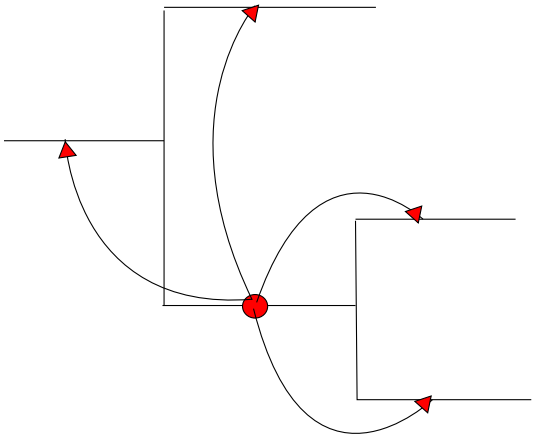

Figure 3.2: An example of a MCMC move: moving a catastrophe from an edge to one of its neighbours.

$\xi_1), \ldots, 1-\eta(1-\xi_L), 1/\eta)$ is η^{L-2}, so the acceptance probability is

$$\alpha(x'|x) = \min\left(1, \frac{p(x'|D)}{p(x|D)}\eta^{L-2}\right)$$

For the addition and deletion of catastrophes, we do not need to use reversible jump Markov Chain Monte Carlo, as the state vector specifies the numbers, and not the locations, of catastrophes on edges.

We omit the details of these moves but give, as an example, the update that moves a catastrophe from an edge to a parent, child or sibling edge. Let $k_T = \sum_{i=1}^{2L-2} k_i$ give the total number of catastrophes. Given a state $x = (g, \mu, \kappa, \rho, \xi)$ with $g = (V, E, t, k)$, we pick edge $\langle i, j\rangle \in E$ with probability k_i/k_T. Let $E_{\langle i,j\rangle}$ be the set of edges *neighbouring* edge $\langle i, j\rangle$ (child, sibling and parent edges, but excluding the edge $\langle r, A\rangle$ since we put no catastrophes on this edge of infinite length) and let $q_i = \text{card}(E_{\langle i,j\rangle})$. We have in general $q_i = 4$. However, for i the index of a leaf node, $q_i = 2$ (1 parent, 1 sibling, no children). If j is the root and i is non-leaf, then $q_i = 3$ (1 sibling, 2 children) and if j is the

root and i is a leaf we have $q_i = 1$ (a sibling edge). Choose a neighbouring edge $\langle \tilde{i}, \tilde{j} \rangle$ uniformly at random from $E_{\langle i,j \rangle}$ and move one catastrophe from $\langle i, j \rangle$ to $\langle \tilde{i}, \tilde{j} \rangle$. The candidate state is $x' = ((V, E, t, k'), \mu, \kappa, \rho)$, with $k'_i = k_i - 1$ and $k'_{\tilde{i}} = k_{\tilde{i}} + 1$ and $k'_l = k_l$ for $l \neq i, \tilde{i}$. This move is accepted with probability

$$\alpha(x'|x) = \min\left(1, \frac{q_i \, k'_{\tilde{i}} \, p(x'|D)}{q_{\tilde{i}} \, k_i \, p(x|D)}\right).$$

Markov Chain Monte Carlo presents two issues: the chain samples from the posterior distribution only once it has reached equilibrium, and the samples it outputs are not independent. We assessed convergence with the asymptotic behaviour of the autocorrelation for the parameters μ, κ, ρ and t_r and the log-likelihood, as suggested by Geyer [1992]. Given Markov chain (X_n) from which we have a run of length N_{samp} and given statistic S, the autocorrelation function r_S and the integrated autocorrelation time τ_S are defined as

$$r_S(t) = \frac{cov\left(S(X_n), S(X_{n+t})\right)}{var(S)} \tag{3.19}$$

$$\tau_S = \sum_{t=-\infty}^{+\infty} r_S(t). \tag{3.20}$$

We estimate τ_S by restricting the sum between $-M$ and M for some M, $1 \ll M \ll N_{samp}$. Then N_{samp}/τ is the effective sample size, a measure of the size of an independent sample which would provide the same variance of the sample mean as the sample output by the MCMC. This method indicates that we can use runs of about 10 million samples (thinned down to 1000 samples); we also let the MCMC run for 100 million samples and checked that the computed statistics did not vary. As an example, we give in Figure 3.3

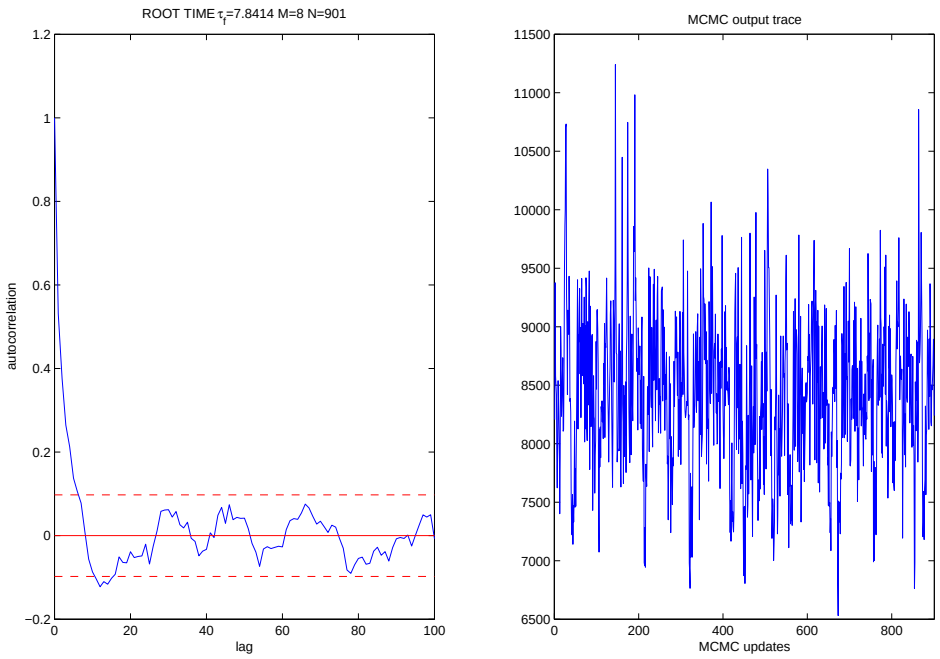

Figure 3.3: *Some of the output used to assess convergence for the results shown in Section 5.3. These two plots help to visualize a statistic S, here the root age t_r. A run of 10 million samples was thinned down to $N_{samp} = 1000$ samples, of which the first 100 were discarded as burn-in. For the 900 remaining samples, the left plot shows the autocorrelations $r_S(t)$ against t. The two dotted red lines delimit a 95% credible interval for the correlation of independent samples, so that once the autocorrelation lies between these two lines, the MCMC output samples are equivalent to independent samples. The right plot shows the trace of the root age in the MCMC output.*

a part of the output we used to assess convergence for the results shown in Section 5.3.

3.4 Debugging tests

We made a number of debugging checks on our code. We analysed an empty data file and set the log-likelihood to 0, then checked that the posterior

distributions for the death parameter μ, the catastrophe parameters ρ and κ and the root age t_r were identical to the prior distributions. On small trees with small data sets, the likelihood can be computed by hand: we checked the results from our code against manual calculations for a few data sets with 1 or 2 traits on a tree with 4 leaves.

On a tree with 5 leaves, we generated all possible registered data sets with non-zero probability (including all possible patterns of missing data) following registration rule R_1 from Section 2.2.3. We chose a high death rate, so that the likelihood of data sets in which 6 or more traits have survived is negligible. With all parameters fixed, we checked that the likelihoods of all data sets summed to 1. This is a very strict requirement, unlikely to be satisfied if there are bugs, for a point-process of the kind modeling our data. We repeated this with registration rule R_2. This provides a check of the mathematical derivation as well as of the code.

Analyses of synthetic data are presented in Chapter 4.

Chapter 4

Validation

In this chapter, we present results showing the validity of our methods, using synthetic data as well as cross-validation on the real data set collected by Ringe et al. [2002].

In each analysis, the sample from the posterior distribution on trees is summarized with a consensus tree, which shows all splits with at least 50% support in the posterior sample. When the support for a split is between 50% and 95%, the split is labeled; unlabeled splits receive at least 95% support in the posterior. The displayed length of a branch is the average length of that branch in the posterior, conditional on that branch existing; this means that the time depths shown on consensus trees can occasionaly be confusing, especially when there is high uncertainty in the topology. The number of catastrophes displayed on a branch is the average number of catastrophes on that branch in the posterior sample, conditional on that branch existing, and rounded to the nearest integer. The samples from the posterior distribution for the parameters of the model are summarized with either histograms from

the posterior or 95% Highest Probability Density (HPD) intervals.

4.1 In-model testing

We made a number of tests using synthetic data. Fitting the model to synthetic data simulated according to the likelihood $P(\mathbf{D} = D|g, \mu, \lambda, \rho, \kappa, \xi, \mathbf{D} = R(\tilde{\mathbf{D}}))$, (in-model data), shows us just how informative the data is for the topology, node ages, catastrophe placement, and parameter values, as well as making a debug-check on our implementation.

For clarity, we first look at data simulated on trees with catastrophes but where no data go missing, then at data simulated on trees with no catastrophes, but where some data go missing. The results still hold when both issues are combined.

4.1.1 Catastrophes

We simulated data under our model for different values of λ, μ, κ and ρ, in order to explore the different possible scenarios of rate heterogeneity, such as few large catastrophes or many small catastrophes. We synthesized data for 20 languages, with on average $\lambda/\mu = 100$ traits per language and 5 clade constraints; the real data sets we analyse in Chapter 5 have more traits than this. We set $\mu = 2 \cdot 10^{-5}$ deaths/year, since the estimates for that parameter in Chapter 5 are around that value. We studied small ($\kappa = 0.1$), medium ($\kappa = 0.2$) and large ($\kappa = 0.5$) catastrophes (corresponding to catastrophes equivalent to 520 years, 1160 years and 3280 years of change respectively). We studied values of the catastrophe rate ρ corresponding to rare catastrophes (between

1 and 5 catastrophes on a tree with 38 branches), occasional catastrophes (around 10 catastrophes on the tree) and very frequent catastrophes (up to 100 catastrophes on the tree, or about 3 per branch on average); we also studied parameter values taken from the posterior distribution for our analysis of the Ringe et al. [2002] data described in Chapter 5. We show typical results in Figures 4.1 and 4.2. The true topology was almost perfectly reconstructed in every case; the position of catastrophes was perfectly reconstructed; the posterior evaluations of the parameters were close to the true values. This remains true for a wide variety of parameter values.

When catastrophes are present in the true tree, the signal for them is strong. We wished to check the influence of catastrophes on our reconstructions, so we tried fitting a model without catastrophes to data which did, in fact, evolve with catastrophes. Figure 4.3 is typical: if we do not include catastrophes in the model, our reconstructed parameters (root age and death rate μ) are very far from the true values. The reconstructed topology is also much further from the truth than when we fit the model with catastrophes.

Note that these results also show that our methods are not subject to the criticisms of Blust [2000] who claimed that issues of rate heterogeneity meant that "lexicostatistics doesn't work", *i.e.* that the topology of the tree cannot be reconstructed through statistical modelling. The main issue raised by Blust [2000] is effectively long branch attraction. Though his notation is different, he is concerned with trees such as the one shown in Figure 4.4. When there is enough rate heterogeneity, it may appear that language C is closer to languages A and B than to language D, its true sibling, and that language C will therefore end up either falsely grouped with A and B, or falsely as

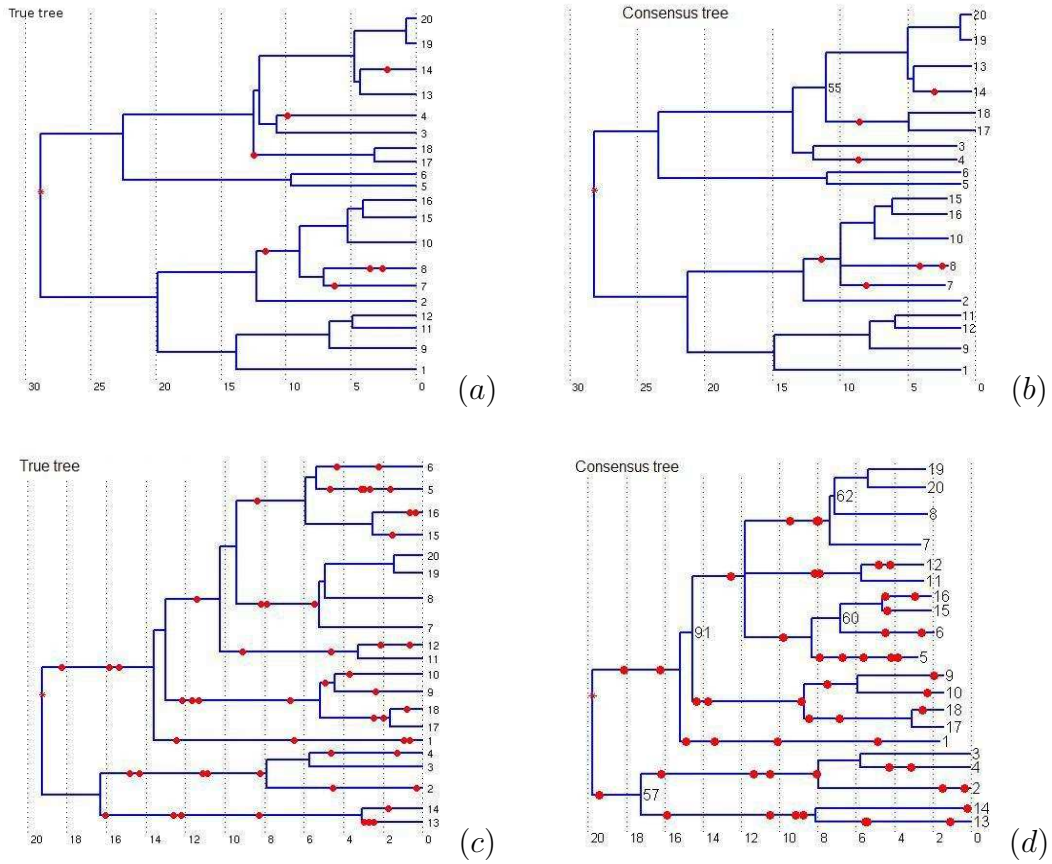

Figure 4.1: *Simulations of synthetic data show the robustness of the model: (a) and (c), true trees; (b) and (d), reconstructed consensus trees. The consensus trees are very close to the true trees and the reconstructed catastrophes are on the correct branches.*

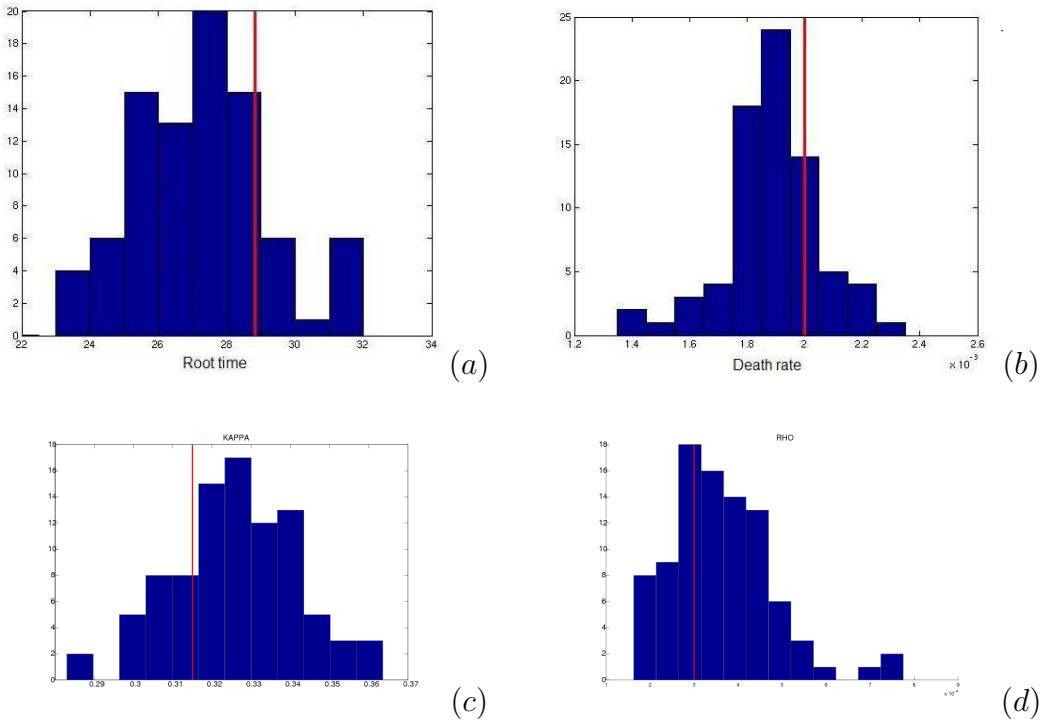

Figure 4.2: *Simulations of synthetic data show the robustness of the model: samples from the posterior for the root age t_r (a), the death rate μ (b), the death probability at a catastrophe κ (c) and the catastrophe rate ρ (d) are close to the true values, shown here in red.*

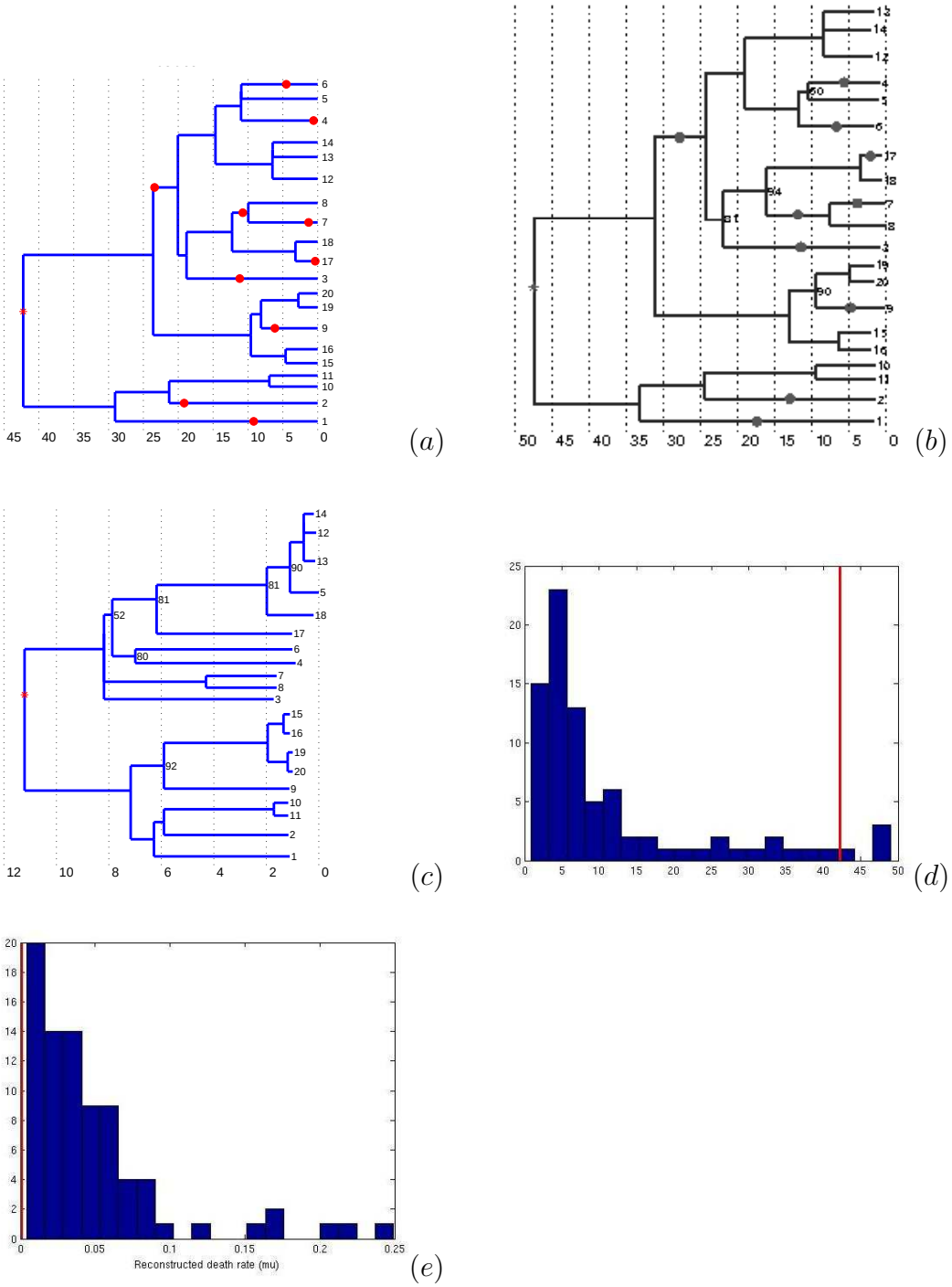

Figure 4.3: *Importance of including the catastrophes: given data synthesized under a true tree with catastrophes (a), which was well reconstructed by a model with catastrophes, as shown in the consensus tree (b), we tried to fit a model without catastrophes. The topology showon in the consensus tere (c), root age $t_r(d)$ and death rate μ (e) were all badly reconstructed.*

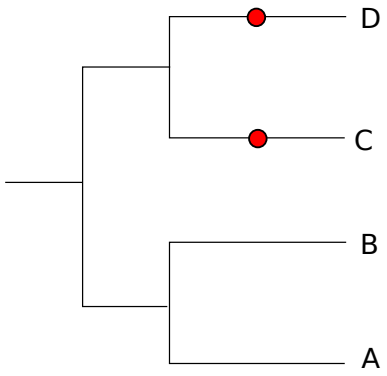

Figure 4.4: More change has happened on the branches leading to C and D than on the branches leading to A and B. Blust [2000] is concerned that such situations will lead to "subgroup splitting".

an isolate; Blust [2000] calls this "subgroup splitting". The reconstructions above show that rate heterogeneity can be detected and need not lead to false groupings. The worst that can happen is that if too much change occurs, there will be high uncertainty in the reconstruction, but this is of course true of any method. For example, in Figure 4.1 (a), the subtree containing languages 7, 8, 10, 15 and 16 resembles Figure 4.4, and is correctly reconstructed as uncertain in Figure 4.1 (b). On the other hand, in Figure 4.3 (a), the subtree containing languages 4, 5, 6, 12, 13 and 14 is incorrectly (and confidently) reconstructed in Figure 4.3 (c) when catastrophes are not taken into account.

4.1.2 Missing data

The data used in this section were synthetized using the model described in Chapter 2, with no catastrophes and with 20 leaves. The missing data parameters ξ_i $(i = 1 \ldots 20)$ were drawn from a $Beta(3, 1)$ distribution restricted to the range $[0.05, 1]$, since this corresponds roughly to the proportions of missing data we observe in real data sets.

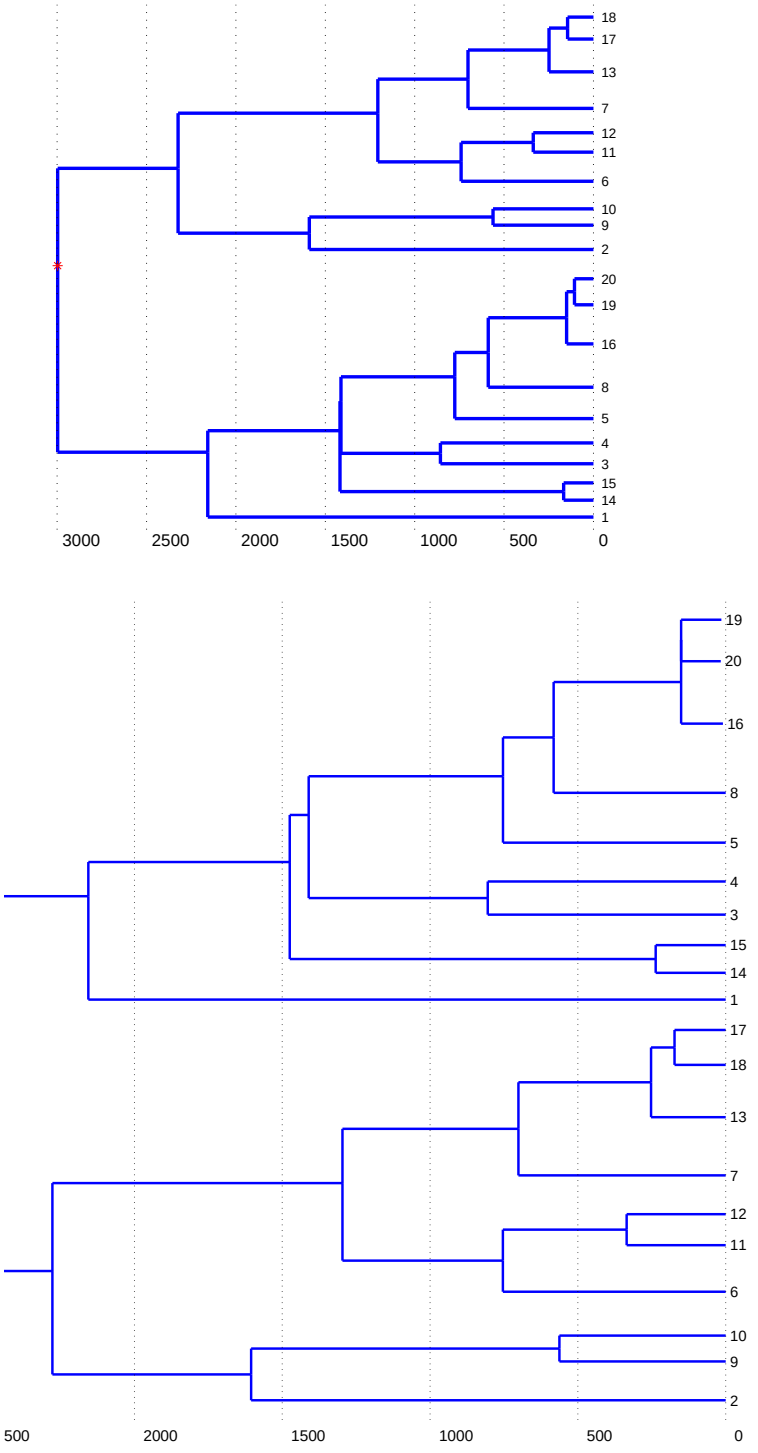

Figure 4.5: True and consensus tree for synthetic data with some data missing at the leaves.

Figure 4.6: Red: true values of the parameters ξ_i, $i = 1 \ldots 20$; Blue; 95% HPD intervals from the posterior distribution.

Figure 4.5 shows the true tree used to synthetize the data, and the consensus tree from our analysis. The topology is almost perfectly reconstructed, the only issue being that the topology of the subtree with leaves 16, 19 and 20 is undecided. Figure 4.1.2 shows the true values of the ξ_i ($i = 1 \ldots 20$) and 95% HPD intervals from the posterior samples. Of the 20 parameters, 19 are covered by the HPD interval and one is not. Figure 4.1.2 shows that the root time and death parameter μ are also well reconstructed.

These analyses on in-model synthetic data show that data simulated following our model, and of the size that we typically observe in real data sets, contain a great deal of useful information about the topology, catastrophe placement, and model parameters, at least for the parameter regions we considered.

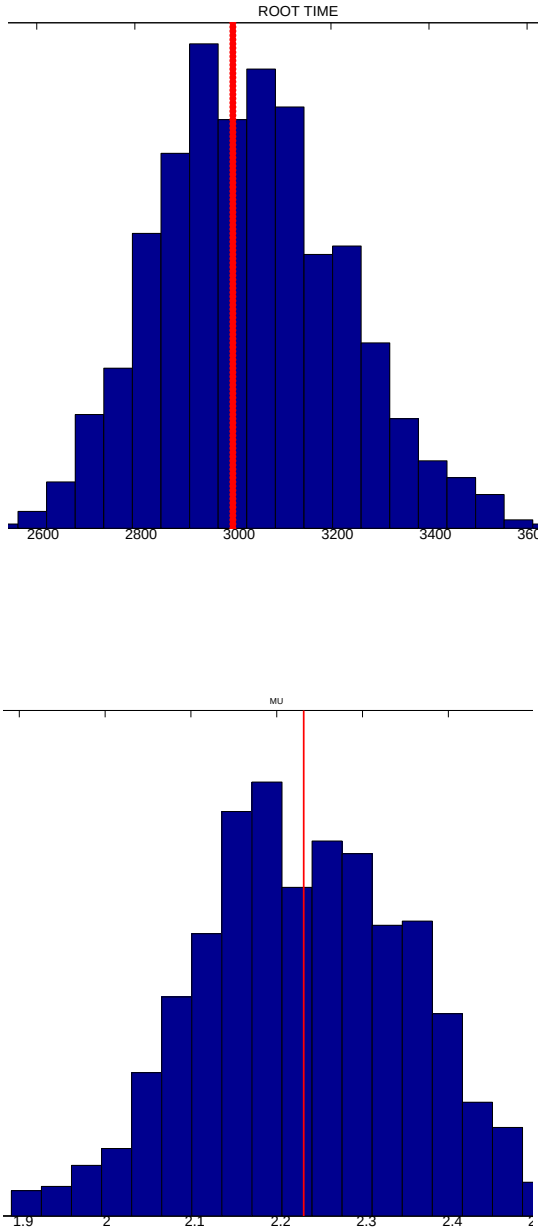

Figure 4.7: Red: true values of the root time (left) and of the death parameter μ; Blue: histograms from the posterior sample. This is not a "selected" example; it is simply the first one we generated.

4.2 Out-of-model testing

We fit out-of-model data also. These are synthetic data simulated under likely model-violation scenarios, and are used to identify sources of systematic bias.

4.2.1 Borrowing

Borrowing between languages is a frequent phenomenon, which we do not include in our model. Even though the levels of borrowing for core vocabulary are much lower than for general vocabulary [Embleton, 1986], we searched for potential systematic bias. We simulated data with different levels of borrowing under two different models of borrowing (global and local). The model of borrowing we use is as follows: borrowing events occur at rate $b\mu$ in each language, for some level of borrowing b. At a borrowing event at time t in language l_1, a language l_2 is chosen uniformly at random; one trait is chosen uniformly at random amongst those present in language l_1 at time t, and it is copied into l_2 - if the trait is already present in l_2 then there is no effect. Under the global model, borrowing can occur between any two languages. Under the local model, it can occur only between languages which split less than T_b years previously (we used $T_b = 1000$ years); this is a crude proxy for geographic proximity. This model of borrowing was specified by Nicholls and Gray [2008], who also checked borrowing for a simple model fit. We looked at low levels ($b = 0.1$) and high levels ($b = 0.5$) of borrowing. We did not consider "catastrophic" borrowing, in which one language would borrow many words from another language in a short amount of time, as did Warnow et al. [2004]. The results presented here are for the global model; results for the local

model are very similar.

Similar studies using this exact model have been performed by Greenhill et al. [2009], who also found that the levels of borrowing we would expect in core vocabulary are unlikely to introduce significant systematic bias in age and topology estimates. Nunn et al. [2006] also find evidence of bias in reconstructions for data with high levels of borrowing.

For $b = 0.1$, the topology is well reconstructed, with only minor differences between the true tree and the output (Figure 4.8 (a)-(b)). The dates, catastrophes and parameters are also correctly reconstructed. This is typical, so the effect of low levels of borrowing is negligible, under both global and local models of borrowing. For $b = 0.5$, the topology was surprisingly well reconstructed in the examples we looked at, given the amount of noise in the data (Figure 4.8 (c)-(d)). However, we found that for $b = 0.5$, we systematically underestimated the root age and overestimated the rate parameters by up to 75% (Figure 4.8 (e)-(f)). This is of little concern to us, since we have reason to believe that no such high levels of borrowing occurred in the data we are analysing. For example, English is often cited as a language which borrowed a lot of its lexicon. It is estimated that 50% of its lexicon was borrowed from Romance languages (mainly French and Latin), but our data only contains the "core vocabulary". Only 6% of the core vocabulary of English comes from borrowing [Embleton, 1986]; furthermore, these borrowings are easy to detect through phonological irregularities and are removed from the data.

On the other hand, very high levels of horizontal transmission can be expected in certain cultural data sets, such as the spread of horse-culture

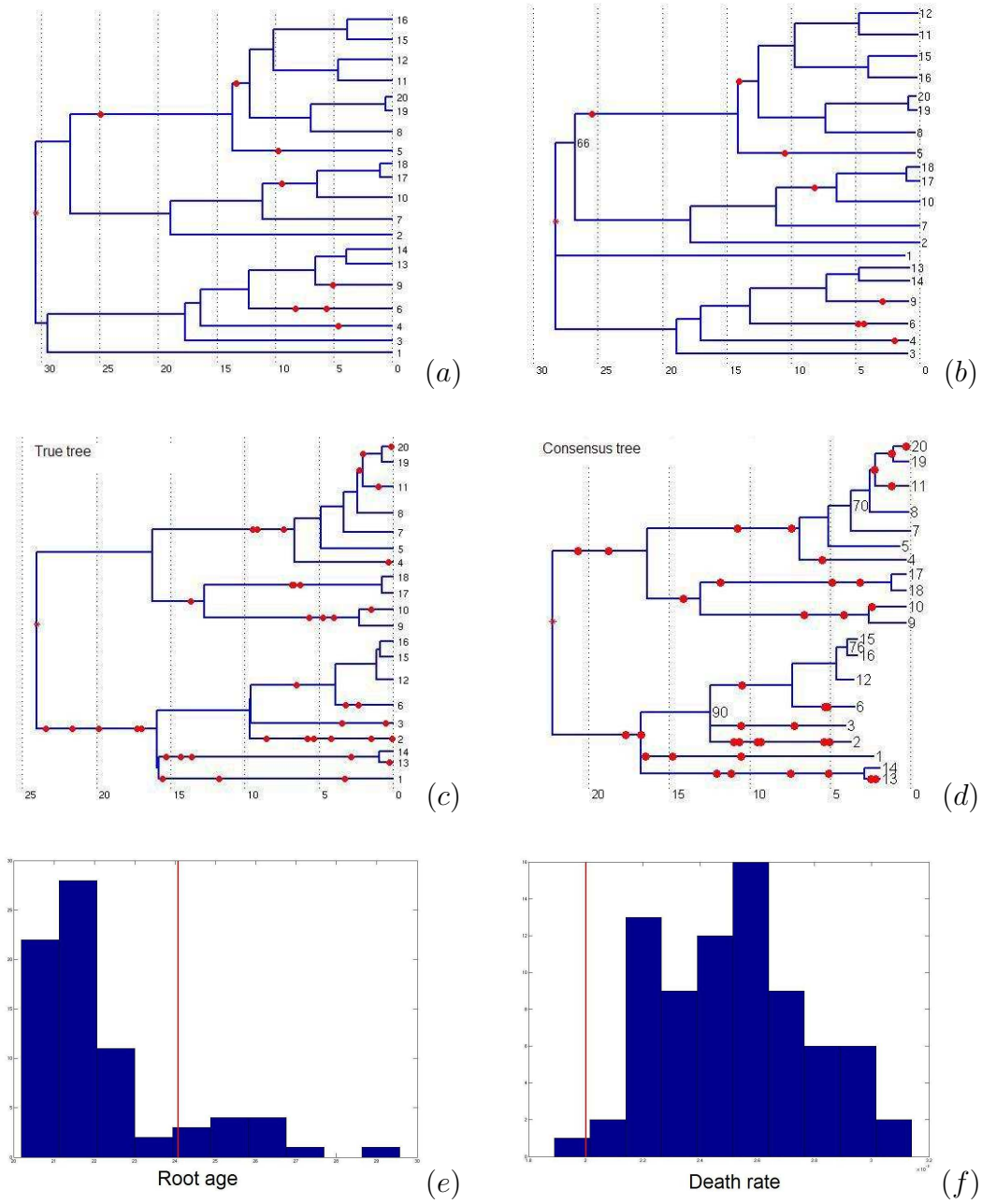

Figure 4.8: *Influence of borrowing. (a)-(b): low levels of borrowing (b=0.1) have negligible effects on the topology and the parameter estimates. (c)-(d): high levels of borrowing (b=0.5) still allow to reconstruct most of the topology, but the root age and parameter estimates, shown here for* $b = 0.5$*, are biased (e-f).*

in America [Roe, 1955][1] or of traits associated with Islam in East Africa [Ensminger, 1997][1]. We would therefore expect our methods to perform poorly on such datasets.

These results beg the question: how do we detect data sets with high levels of borrowing? Although we do not have a precise measure to suggest, the distribution of the number of languages per cognacy class seems helpful. We synthetized data with borrowing ($b = 0, 0.1, 0.5$ and 1) on a tree sample from the posterior for our analysis of the Ringe et al. [2002] data (24 languages) with parameters also taken from the posterior. Figure 4.9 gives the cumulative distribution of the number of languages at which a cognacy class appears; we also include the distribution for the Ringe et al. [2002] data. As one would expect, higher levels of borrowing make cognacy classes appear at more leaves. There is a clear difference between the various graphs, which can allow us to detect data with borrowing. The graph for Ringe et al. [2002] data is very close to the graphs for synthetic with little or no borrowing, in line with linguists' estimates of borrowing in the Indo-European core vocabulary [Embleton, 1986].

It should be noted that the distribution of the number of languages per cognacy class depends not only on the level of borrowing, but also on the topology and other model parameters. In order to detect borrowing in a data set, we therefore suggest to first analyse the data, then construct synthetic data sets using a tree and parameter values from the posterior and with various degrees of borrowing.

[1] Cited by Nunn et al. [2006].

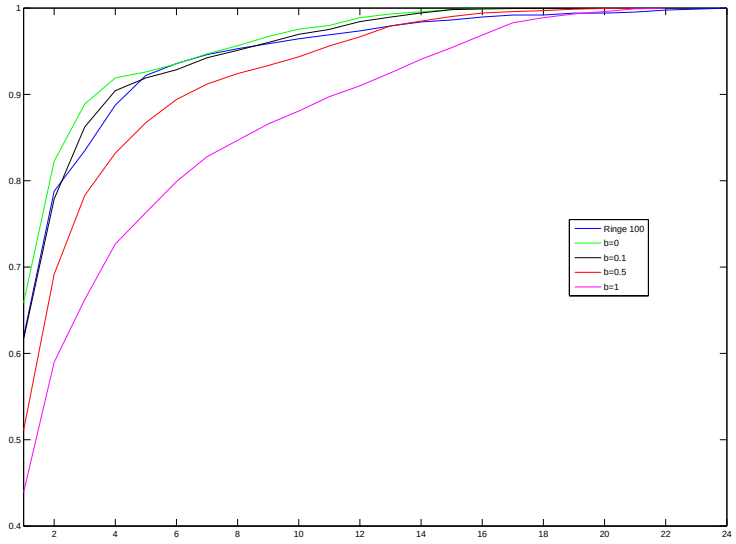

Figure 4.9: *Cumulative proportion of the number of languages at which a cognacy class is displayed for synthetic data with various values of the borrowing parameter b and for the real data from Ringe et al. [2002].*

4.2.2 Meaning categories

Our model ignores the fact that the data are grouped in meaning categories. This leads to several issues:

1. We treat all cognate classes as being independent from each other. However, it is likely that cognate classes within one meaning category are not independent. In particular, the probability for a word to die out is presumably higher if a synonym exists in the language.

2. Our model allows for a language to not have any word in a given meaning category, which is unlikely to happen since we are dealing with core vocabulary.

3. We assume that data go missing uniformly at random. Actually, they go missing in blocks: if a language has a question mark for a given cognate class, it usually means that we do not know the word(s) for that meaning category in that language, hence the language will usually also have question marks for all other cognate classes in that meaning category.

We did not test for systematic bias arising from the assumption that meaning categories are independent. To test for systematic bias arising from the assumption that cognacy classes within a meaning category are independent, we simulated synthetic data taking into account the meaning category structure, and attempted to reconstruct the parameters and tree; this expands on an analysis by Nicholls and Gray [2008].

Our model for diversification with meaning categories is similar to the model described in Chapter 2, with two modifications. First, we define 100 meaning categories. When a word in a language is to die, we check whether it is the last word in that language for that meaning category; if it is, we simply ignore the death event. This means that at any point on the tree, there is at least one word per meaning category. It also means that the effective death rate is lower than the death rate we define, so we should not expect our estimate of the parameter μ to be correct. Second, we simulate data missing in blocks: given a leaf i and a meaning category, all data for that meaning category at that leaf go missing simultaneously with probability ξ_i, and all data for that meaning category at that leaf are correctly registered with probability $1 - \xi_i$; this differs to the model we fit since there we assume that the different cognacy

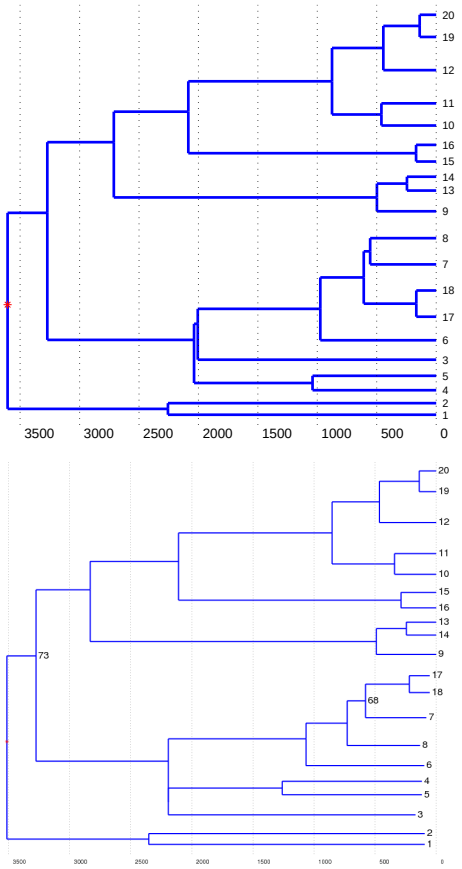

Figure 4.10: True tree and consensus tree for synthetic data with meaning categories imposed.

classes in the same meaning category go missing independently of each other. Note that we still expect to see reasonable estimates of ξ. We simulated data on a tree with 20 leaves, 5 constraints, no catastrophes, and with $\mu = 2.2 \cdot 10^{-4}$, taken from the posterior distribution of the analysis of the Dyen et al. [1997] presented in Chapter 5.

Figure 4.10 shows the true tree used to synthetize data as well as the consensus tree. The reconstruction is good, with two issues. The structure of the subtree made of leaves 7, 8, 17 and 18 is uncertain, but wrong in

most of the posterior samples (the correct structure appears in 25% of the posterior samples). The position of leaves 3, 4 and 5 relative to that subtree is undecided (the correct structure appears in 39% of the posterior samples). In both cases, this is presumably because the true tree contains a very short branch on which little or no change occurred. As expected, the death rate μ is not reconstructed, but the root age t_r and the missing data parameters ξ are still correctly estimated (Figures 4.11 and 4.12). Over all, these reconstructions with out-of-model data are very positive.

We are able to correctly estimate the root age (and other internal node times) despite our bad estimate of the death rate μ because the model misspecification is uniform over the tree: the effective death rate is still constant over the entire tree. It is the effective death rate that we reconstruct in Figure 4.11.

4.2.3 Reversibility

In our model, we have imposed the reversibility condition $\nu = \kappa\lambda/\mu$. In order to check for systematic bias arising from this condition, we simulated data with different values of ν and estimated all parameters under the reversibility condition. Here again, we do not expect to be able to correctly estimate all parameters, but we hope that no systematic bias will be introduced in the estimates of the topology and of the root age. The data we simulated used the parameter values $\mu = 2.23 \cdot 10^{-4}$, $\kappa = 0.2$, $\lambda = 4.46 \cdot 10^{-2}$ and we studied $\nu = 2\kappa\lambda/\mu$ and $\nu = \kappa\lambda/2\mu$. The data were simulated on a tree with 20 languages and 8 internal constraints.

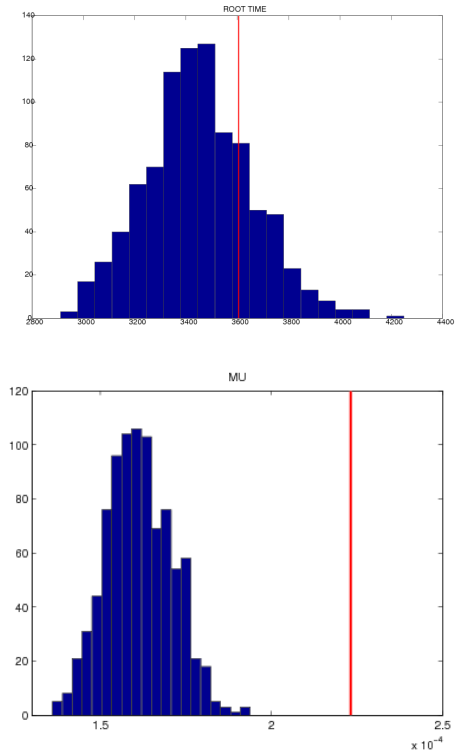

Figure 4.11: Analysis using synthetic data with meaning categories imposed. Red: true values of the root time t_r (left) and of the death parameter μ (right); Blue: histograms from the posterior sample.

Figure 4.12: Reconstruction of the missing data parameters ξ_i, $i = 1 \ldots 20$, for synthetic data with meaning categories imposed.

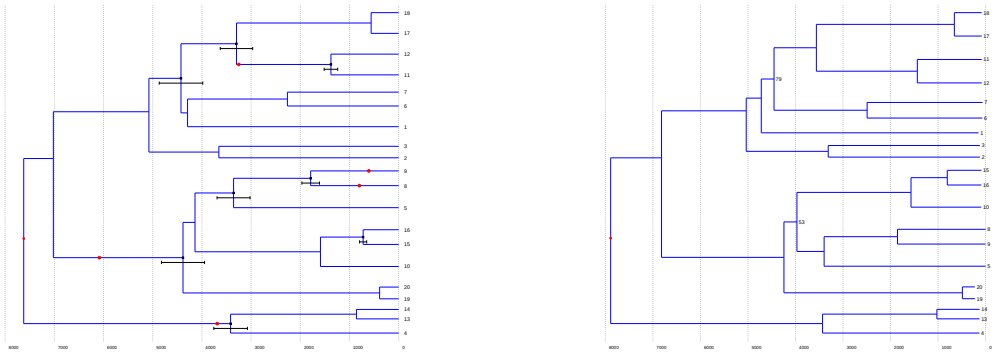

Figure 4.13: *Influence of the reversibility condition. Left: true tree under which data were simulated, under the condition* $\nu = \kappa\lambda/2\mu$*; right: reconstructed consensus tree.*

The reconstruction of the topology is not affected: the topology is still almost perfectly reconstructed, as shown in Figure 4.13. However, the position of catastrophes is much more uncertain: no catastrophes are supported in more than 50% of the posterior. The trees shown are for $\nu = \kappa\lambda/2\mu$; the situation is similar for $\nu = 2\kappa\lambda/\mu$.

The parameters μ and κ are not well reconstructed, as shown in figure 4.14: the posterior distribution is highly uninformative of κ and the death rate μ is systematically overestimated. However, the root age, which is the parameter of interest, is well reconstructed: for $\nu = \kappa\lambda/2\mu$, the true root age was 7622BP and the 95% HPD is 6932–8584BP; for $\nu = 2\kappa\lambda/\mu$, the true root age is 7664BP and the 95% HPD is 6472–7701BP. In both cases, the 95% HPD covers the true value.

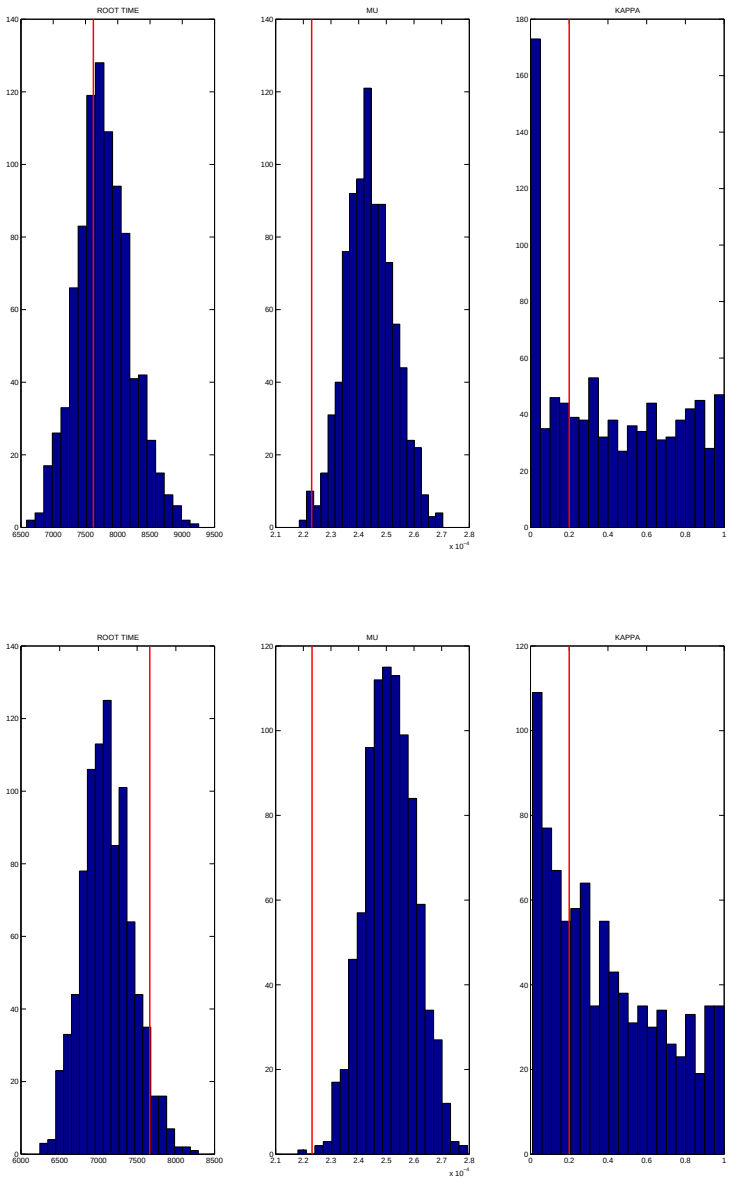

Figure 4.14: *Influence of the reversibility condition. Top: data simulated under the condition* $\nu = \kappa\lambda/2\mu$*; bottom: data simulated under the condition* $\nu = 2\kappa\lambda/\mu$*. Blue: posterior sample from the reconstruction under the reversibility condition* $\nu = \kappa\lambda/\mu$*; red: true value of the root age, death rate* μ *and probability of death at a catastrophe* κ*.*

4.3 Reconstruction of known ages

The calibration constraints described in Section 2.1 impose age constraints on some internal nodes and some leaves. In this section, we perform cross-validation test on all age constraints in the Ringe et al. [2002] data.

We remove each calibration constraint in turn and attempt to reconstruct its age from the data and the other calibration constraints. In the Germanic family, we have three constraints. Removing only one of these constraints would serve little purpose, since the age of the corresponding node would effectively still be constrained by the two other constraints and so we could expect to correctly reconstruct the node age in any case. We therefore removed these three constraints simultaneously and estimated the corresponding node ages jointly.

Our success at this exercise will be a good indicator to how much we can trust the age estimates for nodes for which there are no constraints, including the root. The topological constraints were all perfectly reconstructed. The bottom half of Figure 4.15 gives the constraint for each node, as well as a 95% HPD interval from the posterior sample for that node age. In 26 out of 30 tests, the 95% HPD overlaps the constrained age interval.

Some of the HPD intervals are very large and only barely cover the constraint, for example the one for Hittite. To quantify goodness of fit, and in particular to estimate the probability for us to make type II error, we use Bayes factors instead of p-values. Given calibration c ($c \in \{1 \dots C\}$), let

$$\Gamma^{C-c} = \bigcap_{\substack{c'=0 \\ c' \neq c}}^{C} \Gamma^{(c')}.$$

denote the enlarged tree space with the c'th constraint removed but all other constraints imposed. Let Γ_K^{C-c} be the enlarged tree space, extended to include catastrophes (as in Section (2.2.2)). We then perform a model comparison between the null model with all the constraints, $M^C : g \in \Gamma_K^C$, and the alternative model $M^{C-c} : g \in \Gamma_K^{C-c}$ with the constraint removed. The Bayes factor $B_{C,C-c}$ for the model comparison is the ratio of the posterior probabilities for these models with equal prior probability on the two models $P(M^{C-c}) = P(M^C) = 0.5$,

$$\begin{aligned}
B_{C,C-c} &= \frac{P(D|g \in \Gamma_K^C)}{P(D|g \in \Gamma_K^{C-c})} \\
&= \frac{P(D|g \in \Gamma_K^C, g \in \Gamma_K^{C-c})}{P(D|g \in \Gamma_K^{C-c})} \qquad (since\ \Gamma_K^C \subset \Gamma_K^{C-c}) \\
&= \frac{P(g \in \Gamma_K^C, g \in \Gamma_K^{C-c}|D)P(D)}{P(g \in \Gamma_K^C, g \in \Gamma_K^{C-c})} \times \frac{P(g \in \Gamma_K^{C-c})}{P(g \in \Gamma_K^{C-c}|D)P(D)} \\
&= \frac{P(g \in \Gamma_K^C|g \in \Gamma_K^{C-c}, D)P(g \in \Gamma_K^{C-c}|D)}{P(g \in \Gamma_K^C)} \times \frac{P(g \in \Gamma_K^{C-c})}{P(g \in \Gamma_K^{C-c}, D)} \\
&= \frac{P(g \in \Gamma_K^C|g \in \Gamma_K^{C-c}, D)}{P(g \in \Gamma_K^C|g \in \Gamma_K^{C-c})}
\end{aligned}$$

In the last fraction, the numerator $P(g \in \Gamma_K^C|g \in \Gamma_K^{C-c}, D)$ is the posterior probability of constraint c being respected given the data under the alternative model M^{C-c}; the denominator $P(g \in \Gamma_K^C|g \in \Gamma_K^{C-c})$ is the prior probability of the same event. We estimate these probabilities by simulating the prior and posterior distribution under the alternative model (with constraint c removed).

The estimates for the Bayes factors are plotted in the top half of Figure 4.15. The variance of these estimators is negligible.

Misfit in the cross-validation corresponds to strong evidence against the constraint. Following Raftery [1996], we take a Bayes factor exceeding 12 (*i.e.* $2\log(B_{C,C-c}) \gtrsim 5$) as strong evidence against the constraint. We have conflict for three of the thirty constraints: the ages of two leaves (Old Irish and Avestan), and the age of one clade (Balto-Slav). As our analysis in Chapter 5 shows, there is a high posterior probability that a catastrophe event occurred on the branch between Old Irish and Welsh, and another between Old Persian and Avestan. The evidence for rate heterogeneity in rest of the tree is so slight, that when we try to predict these calibrations we are predicting atypical events.

Note that our handling of missing data was instrumental in improving our predictions. The calibration interval for the Hittite vocabulary in these data is 3200–3700BP. If we ignore missing data (so replace ?'s with 0's), our prediction for the age of Hittite is 60–2010BP, well outside of the constraints. With missing data included in the model, the he 95% HPD interval for the age of Hittite in our model is 430–3250BP, which just overlaps the constraint and has a higher mean (so we are not only getting greater uncertainty, but also improving the fit). The Bayes factor gives odds less than 2:1 against, so the evidence against the constraint is "hardly worth mentioning" [Raftery, 1996].

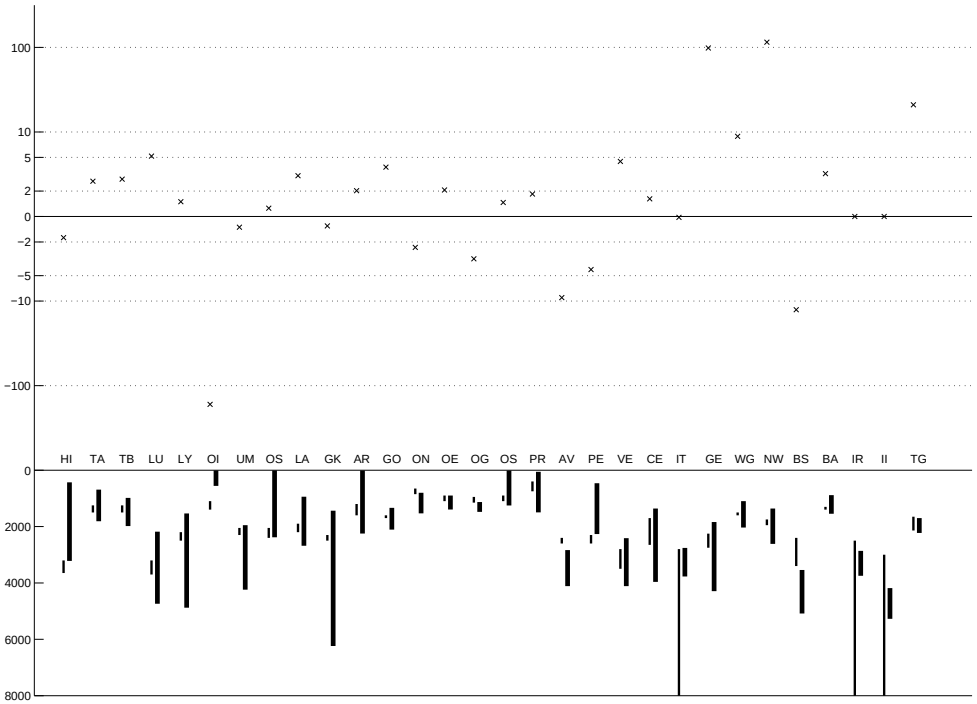

Figure 4.15: *Reconstruction of known node ages: top, logarithm of Bayes factors* $\log(B_{C,C-c})$ *for* $c = 1, 2, ..., C$*; bottom, thin lines show age constraints for different nodes, thick lines show 95% posterior HPD interval for the reconstructed dates when the constraint is removed. HI: Hittite; TA: Tocharian A; TB: Tocharian B; LU: Luvian; LY: Lycian; OI: Old Irish; UM: Umbrian; OS: Oscan; LA: Latin; GK: Greek; AR: Old Armenian; GO: Gothic; ON: Old Norse; OE: Old English; OG: Old High German; OS: Old Church Slavonic; PR: Prussian; AV: Avestan; PE: Old Persian; VE: Vedic; CE: Celtic group; IT: Italic group; GE: Germanic group; WG: West Germanic group; NW: Nort-West Germanic group; BS: Balto-Slav group; BA: Baltic group; IR: Iranian group; II: Indo-Iranian group; TG: Tocharian group*

Chapter 5

Analysis of Indo-European data sets

In this chapter, we analyse two lexical data sets of Indo-European languages and estimate the age of Proto-Indo-European using the model presented in Chapter 2 and with the implemention described in Chapter 3.

The consensus trees displayed in this chapter were built using the same method as in Chapter 4: in a tree, an edge corresponds to a split partitioning the leaves into two sets. A consensus tree displays just those splits present in at least 50% of the posterior sample. Splits which receive less than 95% support are labeled. Where no split is present in 50% of the posterior sample, the consensus tree is multifurcating. The length shown for an edge is the average posterior length given the existence of the split; similarly, the number of catastrophes shown on an edge is the average posterior number of catastrophes on that edge given the existence of the split, rounded to the nearest integer. Since all the parameters have marginal posterior distribution close to normal,

we give our estimates as 95% highest probability density intervals in the normal approximation, using the mean and twice the standard deviation in the posterior sample. For the root age, we give 95% highest probability density intervals.

5.1 Analysis of the Ringe et al. [2002] dataset

As mentioned in Section 2.2, several topological and age constraints are imposed on the trees we reconstruct, along with a range of ages for ancestral nodes and ancient languages. The tree in Figure 5.1 shows the consensus tree from a reconstruction which ignores these constraints, in an attempt to reconstruct these known topological facts. Without any internal constraints, there is no information about the rate parameters, hence the age estimates are meaningless in this analysis: the likelihood would be unchanged by multiplying all branch lengths by an arbitrary factor ν and dividing all the rates by ν. Similarly, the absence of age constraints implies that there is no signal for catastrophes and so no branch bears catastrophes in more than 50% of the sample trees. There are therefore no catastrophes in Figure 5.1 and we do not show any time scale. However, the reconstruction of the topology is interesting: nine of the ten known topological features are supported with probability at least 95% in the posterior distribution. The only exception is the North-West Germanic clade (formed by Old English, Old High German and Old Norse), which appears in 45% of the posterior sample.

We show a consensus tree in figure 5.3 for the complete data analysed with all clade constraints included. For the results described here, our prior on

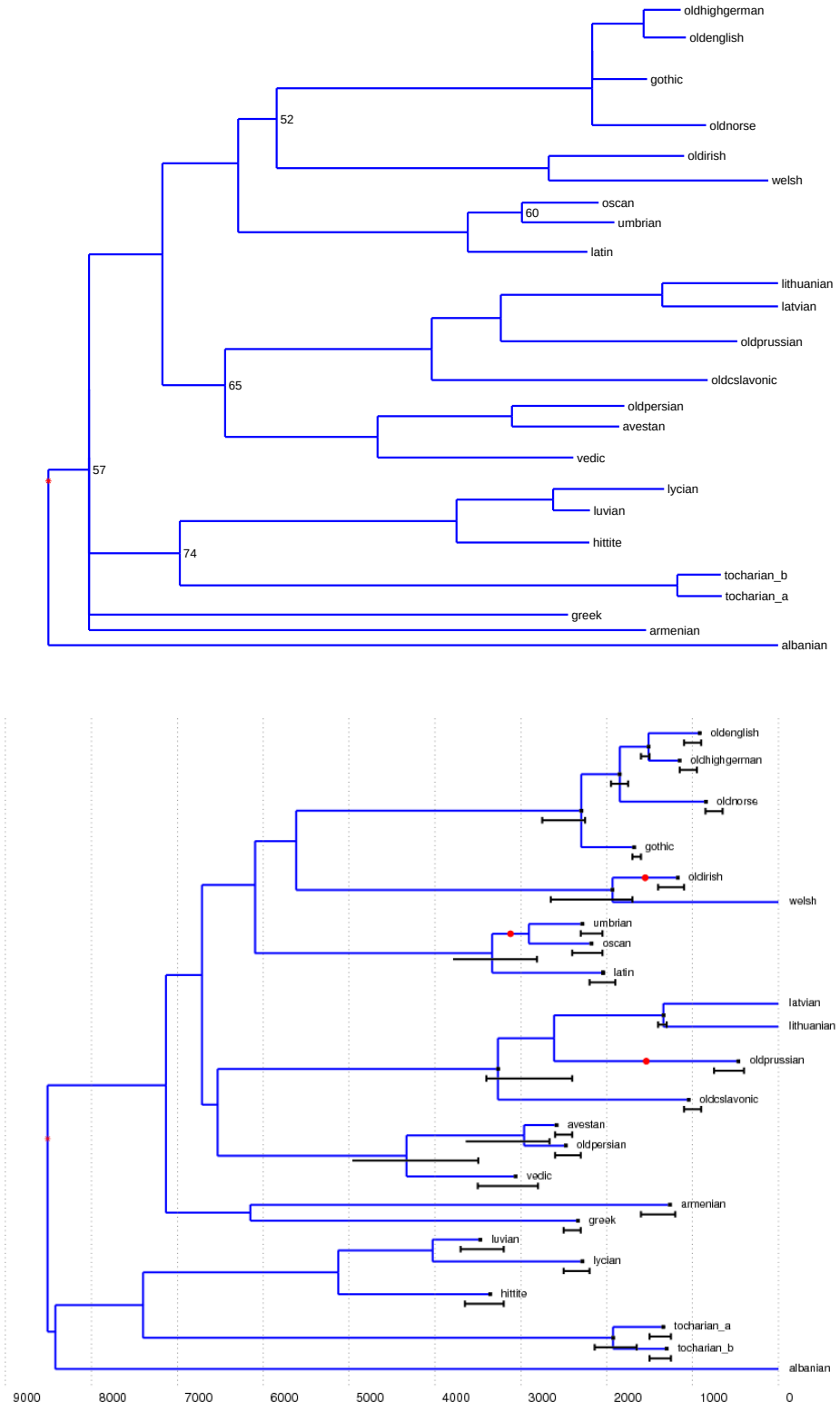

Figure 5.1: *Top: Consensus tree for the Ringe et al. [2002] dataset, without any internal constraints. The dates are meaningless and are therefore not shown, but the topology is well reconstructed. Bottom: A sample tree showing the age topological constraints in black; the bars on internal nodes are topological constraints.*

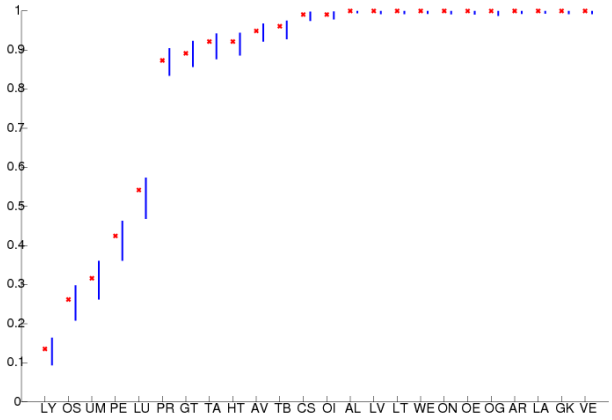

Figure 5.2: *Reconstructed values of the missing data parameters* $1 - \xi_i$ *for the Ringe et al. [2002] data. Blue: 95% highest probability density intervals at each language; Red: proportion of missing data for the language.*

topologies is uniform on labeled histories, and the prior on the catastrophe rate ρ is a $\Gamma(1.5, 0.0002)$; we get similar results with a uniform prior on topologies and a prior on the catastrophe rate $p(\rho) \propto 1/\rho$. The prior on other parameters of the model is described in Chapter 2. Our estimates for the parameters are as follows: $\mu = 1.86 \cdot 10^{-4} \pm 3.94 \cdot 10^{-5}$ deaths/year; $\kappa = 0.361 \pm 0.11$; $\rho = 1.3 \cdot 10^{-4} \pm 3.3 \cdot 10^{-5}$ catastrophes/year (corresponding to large but rare catastrophes: about 1 catastrophe every 15,000 years, or an average of 3.4 on the tree, with each catastrophe corresponding to 2400 years of change). The "life expectancy" of a cognacy class on a branch is then $1/(\mu+\kappa\rho) = 4800$ years. This is greater than the average branch length of 1160 years, so we expect an explosion of the number of languages a registered cognacy class appears at. Our estimates of the missing data parameters ξ_i are shown in Figure 5.2; for any language i, the posterior distribution of ξ_i is quite tight and close to the proportion of missing data in that language.

Figure 5.3: *Consensus tree for the Ringe et al. [2002] dataset. Red dots show catastrophes supported with probability above one half.*

The analysis reconstructs some well-known features of the Indo-European tree which were not part of our constraints. Linguists generally agree that the Germanic, Celtic and Italic families form a subtree, although the relative positions of the three families in this subtree is subject to debate. Our analysis reconstructs this subtree, but no particular configuration of the relative positions is favored. The Indo-Persian group can fall outside the Balto-Slav group but the relative position of these two is uncertain. The deep topology of the tree is left quite uncertain by these data, especially the position of Albanian. We find evidence for catastrophic rate heterogeneity in three positions: on the edges leading to Old Irish, Old Persian, and in the Umbrian-Oscan clade.

Our posterior 95% highest probability density (HPD) interval for the root age of the Indo-European family is 7110 - 9750 years BP. The posterior distribution of this key statistic is close to normal.

The registration rule used by Ringe et al. [2002] when collecting their data was rule R_1 from Section 2.2.3: all data which are observed at at least one leaf are recorded. To validate our results, we removed all singletons (traits which are observed at exactly one leaf), giving the data which would have been obtained had registration rule R_2 been followed. A consensus tree is displayed in Figure 5.4.

In general, the reconstruction of the topology is similar to our previous analysis. The 95% HPD interval for the root age is 6650 - 9380 BP, consistent with our previous analysis. The parameter estimates are as follows: $\mu = 1.56 \cdot 10^{-4} \pm 2.8 \cdot 10^{-5}$ deaths/year; $\kappa = 0.19 \pm 0.11$; $\rho = 1.01 \cdot 10^{-4} \pm 8.72 \cdot 10^{-5}$ (corresponding one catastrophe every 9,500 years, or an average of 5.4 on the

tree, with each catastrophe equivalent to $T_C = 1350$ years of change). The life expectancy of a registered word is 5800 years. The only noteworthy difference is the position of catastrophes, which are in very different places (the only conserved catastrophe is the one on the branch leading to Old Irish). This is not surprising: in Figure 5.3, the catastrophes are all close to the leaves. Presumably, the signal for these catastrophes comes in large part from an unusual number of singletons at those leaves, which is best explained by a catastrophical event. With singletons removed, the signal is much weaker. As a consequence, the reconstructed catastrophes are smaller and are placed in positions where there is rate heterogeneity, but not as strong as the rate heterogeneity detected in Figure 5.3. These catastrophes are present in some of the posterior samples for the analysis with singletons included, but they appeared in less than half the samples and are therefore not displayed in the consensus tree of Figure 5.3.

Another validation method of our results is to exclude part of the data, either by removing some cognacy classes or by removing some languages. The results for these analyses are in line with those for the complete data: on the one hand, removing some cognacy classes makes no significant difference in the reconstructed topology, and slightly increases the variance of the reconstructed ages; on the other hand, removing some languages increases the uncertainty of the topology and of the reconstructed ages. As an example, Figure 5.5 presents a consensus tree with half the languages *and* half the cognacy classes excluded. The excluded traits were chosen at random; the excluded languages were chosen in a way that ensures that most subfamilies were represented (although the Baltic family is not represented by any language in this analysis).

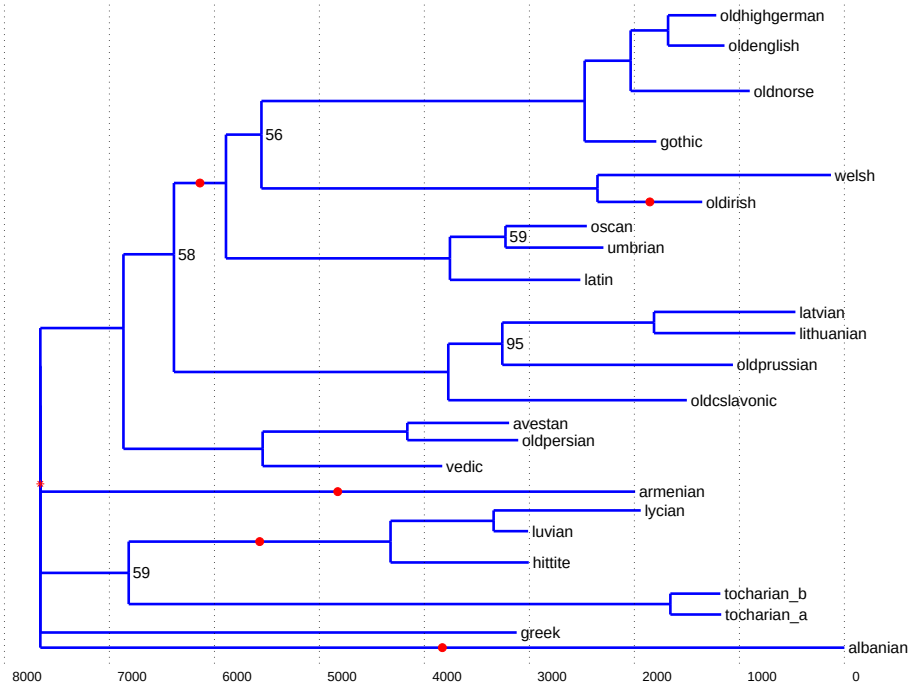

Figure 5.4: *Consensus tree for the analysis of the Ringe et al. [2002] data with singletons excluded.*

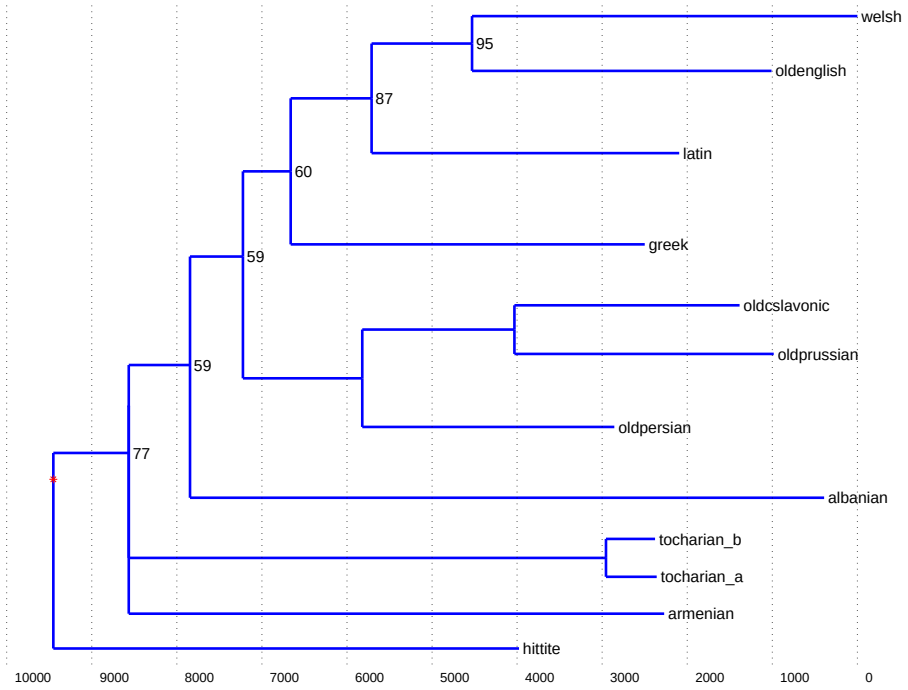

Figure 5.5: *Consensus tree for a subset of the Ringe et al. [2002] data.*

It is worth noting that in the end, more than half of the cognacy classes are discarded, since some of the remaining classes were only observed at leaves which are excluded from the data.

With this reduced data set, there is no signal for catastrophes, and the mode for the catastrophe death probability κ is effectively at 0. Unsurprisingly, the variance of all reconstructed statistics is very high: the 95% HPD interval for the root age is 5910 - 12950 BP, but the mode for this statistic is close to the mode in analyses with the complete data. Similarly, the death parameter μ is reconstructed as $1.35 \cdot 10^{-4} \pm 9.2 \cdot 10^{-5}$, a much greater variance than in previous analyses. The results are similar when only cognacy classes or only languages are excluded, although the variance is not as large.

5.2 Analysis of the Dyen et al. [1997] data

Figure 5.6 presents a consensus tree for the analysis of the Dyen et al. [1997] data under our stochastic-Dollo model. Our estimates for the parameters are as follows: $\mu = 2.37 \cdot 10^{-4} \pm 2.16 \cdot 10^{-5}$ deaths/year; $\kappa = 0.121 \pm 0.096$; $\rho = 2.17 \cdot 10^{-4} \pm 1.2 \cdot 10^{-4}$ catastrophes/year (corresponding to smaller but more common catastrophes than for the Ringe et al. [2002] data: about 1 catastrophe every 4600 years, or an average of 31.3 on the tree, with each catastrophe corresponding to 550 years of change). The life expctancy of a registered word is 3800 years, again greater than the average branch length of 840 years.

The analysis of the Dyen et al. [1997] data strongly supports Indo-Iranian as an outgroup. It also supports the Germanic and Italic subfamilies being siblings, with Celtic as the next closest cousin, though the configurations Germanic-Celtic and Italic-Celtic are also present in the posterior (with about 15% posterior probability each). On the other hand, the analysis of the Ringe et al. [2002] data does not support any particular outgroup, and it shows a preference for a Germanic-Celtic subgrouping. Here again, the other configurations also appear in the posterior sample in non-negligible frequencies. In both cases, the position of Albanian is very unclear. There is agreement between the analyses for the other topological features; these also correspond to the results linguists have obtained through the comparative method.

There is rate heterogeneity in a number of positions. Superficially, some of these positions could be expected. For example, French Creoles, Pennsylvania Dutch and the Gypsy language of Greece all went through some

rate heterogeneity, which could be linked to the large geographical distance from their parent language. We do not have an explanation for the other catastrophes, although they are strongly supported in our analysis. We cannot compare the position of catastrophes in Figures 5.3 and 5.6, because the catastrophes in Figure 5.3 occur close to the leaves and the languages we use in the two analyses are different. However, there is one catastrophe deep in the tree in our analysis of the Ringe et al. [2002] data with singletons removed (Figure 5.4), on the branch leading to the subtree containing the Germanic, Celtic and Italic languages; interestingly, this is also one of only two catastrophes deep in the tree in our analysis of the Dyen et al. [1997] data.

The analysis of the Dyen et al. [1997] data gives a 95% highest probability density interval for the root age of 7080 – 8350 BP, with significant overlap of our estimates using the Ringe et al. [2002] data.

Dyen et al. [1997]'s data were registered by Gray and Atkinson [2003] following registration rule R_2 (*i.e.* they discarded singletons), so it is not possible to validate our results by analysing the data under a different registration rule. However, as previously, we validated our results by using only subsets of our data, and as previously, this led to similar results as with the complete data, only with greater uncertainty.

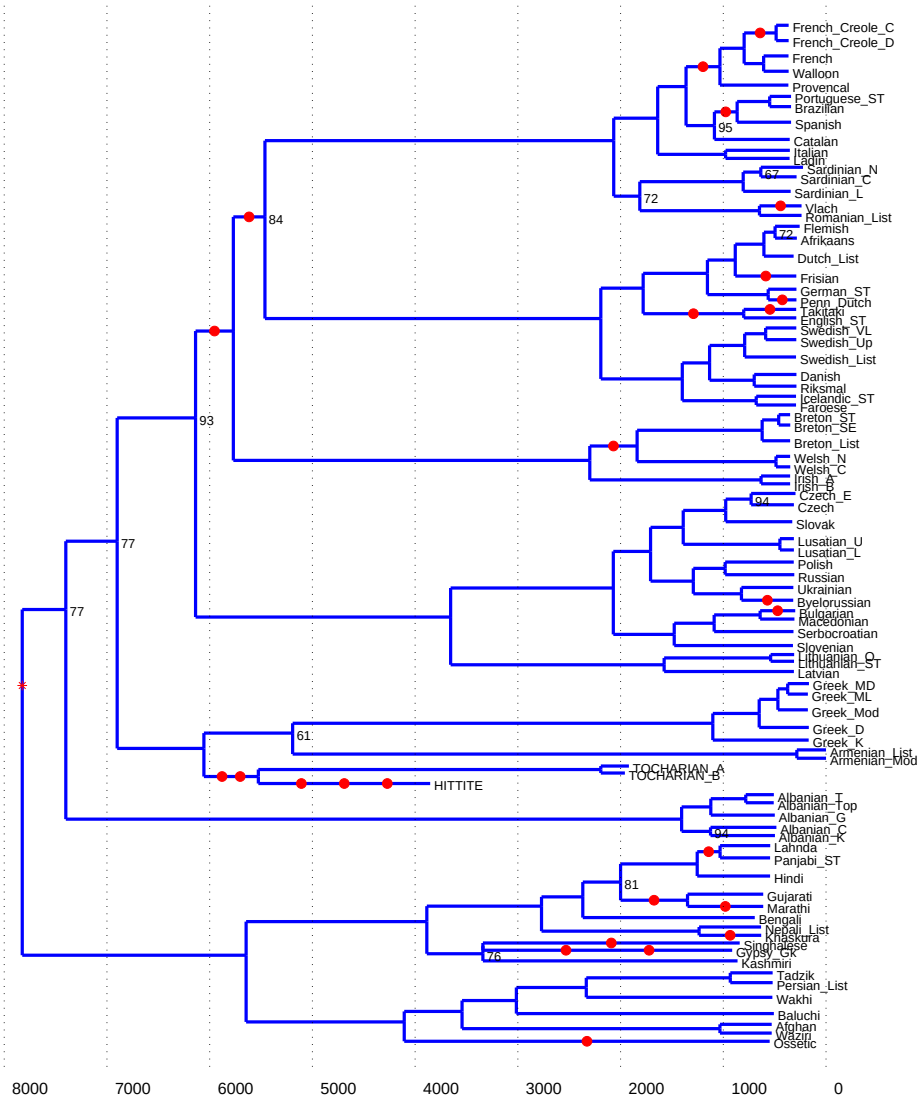

Figure 5.6: Consensus tree for the Dyen et al. [1997] data set.

Chapter 6

Conclusions and extensions

Several attempts at dating Proto-Indo-European using statistical methods have been made previously, as mentioned in Chapter 1. However, some of the models used were clearly not adapted to lexical data, and none of these models had been fully validated. The model we have described in Chapter 2 and implemented as descibed in Chapter 3 is specifically tailored to lexical data. Plausible sources for systematic bias are either incorporated in the model or have been tested for using synthetic data, either by us in Chapter 4 or in previous work. Most significantly, we have showed that borrowing at the levels that we expect in our data does not bias our age estimates.

The reconstruction of known ages presented in Section 4.3 further validates our ability to predict time depths. After several analyses of two data sets (Chapter 5), all our results agree with the Anatolian hypothesis that the spread of the Indo-European family started around 8000 BP. None of our analyses agree with the Kurgan theory that the spread started between 6000 and 65000 BP.

The results presented in this thesis show how promising statistical models of language diversification are. Although much of the effort so far has been in developing models of lexical data, it would be equally interesting to be able to correctly handle all the very diverse aspects of linguistic data: models of grammatical and phonological data, as well as models of contact-induced change rather than descent with modification, would be very useful to help further our understanding of language history.

A few such models already exist, and have been mentioned in Section 1.2. It would be of particular interest to be able to reconcile the various aspects of language change: despite the different mechanisms at play, these aspects are not independent of each other and all must be studied simultaneously if we are to understand the global picture of language diversification.

These models, or extensions of these models, could also be used to gain knowledge on the ancestral languages themselves. Even though the nodes on the trees we reconstruct do not correspond exactly to any languages of interest, they are closely related to such languages. Reconstructing ancestral sequences has been the subject of much research in molecular phylogenetics [Fitch, 1971, Yang et al., 1995]. In particular, many linguists are interested in reconstructing features of the language spoken by the Proto-Indo-Europeans [Schleicher, 1868]. Estimating which cognates are present at the root of the trees we sample would provide information about this Proto-language, and it would of great interest to compare these reconstructions to those obtained by linguists using the comparative method [Beekes, 1995].

The results presented in Chapter 5 indicate that our modelling of catastrophic rate heterogeneity is satisfactory in the setting we studied. To apply it to other

settings, however, it may be necessary to expand on it. For example, it may be interesting to impose that a catastrophe occur simulatenously on all branches at the same time, corresponding to an event that impacts all lineages. It may also be worthy to allow "negative catastrophes", or periods of time during which no change occurs. The most obvous extension is to allow catastrophes to have different sizes. Some of the work in molecular phylogenetics mentioned in Section 1.3 may be relevant.

In the following sections, we give three possible extensions of our methods: we discuss the objections famously raised by Bergsland and Vogt [1962] against glottochronology, reconstruct the spread of Swabian dialects and examine the hypothesis that "punctuational bursts" at language splitting events account for a large amount of linguistic change. Further work is needed to better understand and validate the ideas presented in this chapter.

6.1 Revisiting some extreme examples listed by Bergsland & Vogt [1962]

Swadesh [1952] developed a method known as *glottochronology*, which he claimed could be used to date the most recent common ancestor of any two related languages, by calculating the percentage of shared cognates in the core vocabulary. This was the first attempt at dating ancient languages with mathematical methods. Swadesh [1952] assumed that the core vocabulary evolved at a constant rate. It then follows immediately that if C is the percentage of shared cognates and r the retention rate, then the age of the

most recent common ancestor is $t = \frac{\log C}{2 \log r}$. The constant r was estimated using a pair of languages for which the age of the most recent common ancestor is known. Glottochronology had many shortcomings; from a statistical point of view, the main issues are that there was never any attempt to evaluate the uncertainty of the estimators, and that large amounts of data were not used.

Bergsland and Vogt [1962] strongly criticized glottochronology, and their research is at the origin of the strong disbelief amongst linguists of any statistical method for dating ancient languages. Bergsland and Vogt [1962] used three sets of related languages: Icelandic and three Norwegian dialects, which they compared to several versions of the common ancestor Old Norse; Modern Georgian, Old Georgian and Mingrelian, three Kartvelian languages; and Old and Modern Armenian. They estimated the retention rate r for each set and obtained very different values. This was taken to show that there is no universal constant retention rate and that attempts at dating are therefore pointless.

The methods described in this thesis and in other recent works on dating ancient languages present several advantages when compared to glottochronology:

1. It is the very nature of Bayesian estimation to compute uncertainties;

2. Phylogenetic methods allow us to include data from many languages, rather than only two;

3. The method described by Swadesh [1952] and used by Bergsland and Vogt [1962] only allowed one cognacy class per meaning category. Where there is polymorphism, some (and occasionally most) of the data had to

be ignored.

The research by Bergsland and Vogt [1962] is still viewed by many linguists as an issue that no dating method so far has managed to circumvent. For example, Nakhleh et al. [2005] state that "none of [Bergsland and Vogt's] objections have been effectively met by recent work".

The points made by Bergsland and Vogt [1962] helped a great deal in putting forward the flaws of glottochronology, but their methods also suffer from drawbacks which make their point less forceful. In this chapter, we show that most of the issues raised by Bergsland and Vogt [1962] do not apply to our methods. We calculate dates "Before the Present" (BP); by "present", we mean the time at which the data were recorded, which is in fact approximately 1962.

The first set of languages analysed by Bergsland and Vogt [1962] includes Modern Icelandic, Norwegian Riksmal and the Norwegian dialects of Gjestal and Sandnes, as well as five versions of Old Norse: 10^{th} , 11^{th} , 12^{th} and 13^{th} century Old Norse, and "Legal" Old Norse. We ignored Legal Old Norse in this analysis, because the age of this language is not clear. The known topology of the other languages is shown in Figure 6.1. Note that unlike the ancient languages in other chapters of this thesis, the versions of Old Norse are not leaves, nor are they the most recent common ancestor of a set of languages. Rather, they are assumed to lie on the Adam-Root branch. This is equivalent to coding these languages as leaves with the branch above them having length 0, as in Figure 6.2. In fact, 13^{th} century Old Norse is very close to the most recent common ancester of Modern Icelandic, Riskmal, Gjestal and Sandnes.

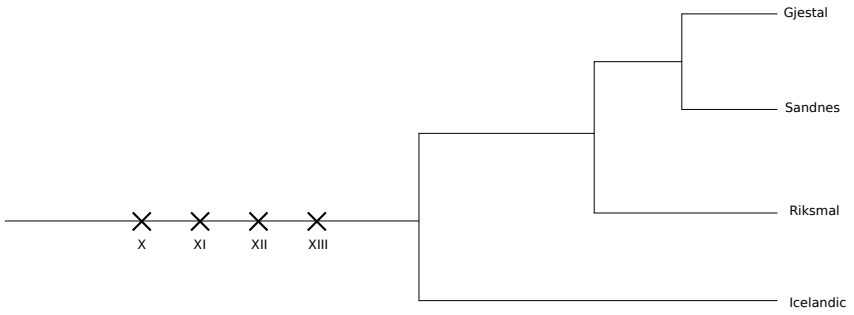

Figure 6.1: *Known topology of the eight studied languages of the Norse family. X, XI, XII, XIII: 10^{th} , 11^{th} , 12^{th} and 13^{th} century Old Norse.*

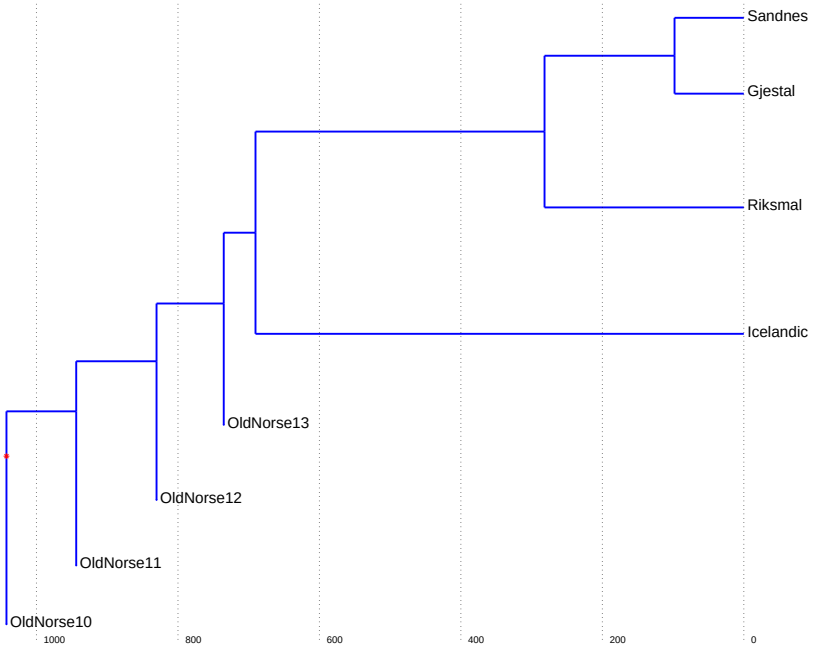

Figure 6.2: *Representation in TraitLab of the known topology of the Norse family. Note that the branches leading to the ancestral languages have length 0.*

Bergsland and Vogt [1962] claim that Icelandic was subject to almost no change between the 13^{th} and 20^{th} centuries. In particular, they claim that when the data are restricted to the shorter Swadesh-100 list (given in Appendix A), the set of cognacy classes observed in Modern Icelandic is exactly the set observed in 13^{th} century Old Norse. This is true for the data they collected, but only because their word list for Modern Icelandic includes words which they describe as "rare" and "literary" and which are in fact not used anymore in spoken Icelandic. This goes against best practice in linguistic data collection [Slaska, 2005]. In our analysis, words which only exist in literary Icelandic are treated as absent in that language. We also exclude literary words in the other modern languages, but these are much less frequent, presumably because these languages have less of a literary tradition.

Glottochronology as developed by Swadesh [1952] and used by Bergsland and Vogt [1962] did not allow for polymorphism (several words for one meaning): if synonyms were in use in a language, Bergsland and Vogt [1962] followed Hymes [1960]'s advice to toss a coin to choose which one to keep. Even though Bergsland and Vogt [1962] were not able to handle polymorphism, they still listed all words for a given meaning category in every language in their word lists. When we coded up the data in binary format, we were therefore able to include all words in a given meaning category for every language. Wherever Bergsland and Vogt [1962] expressed doubt on the presence of a cognacy class in a language, we listed that data point as missing. We excluded all known borrowings, and treated borrowed words as absent. In doing this, we followed to the best of our ability the standard procedure for registration data used by Ringe et al. [2002].

Even with these data excluded, we should expect Icelandic to have experienced less change than average. Indeed, there is a selection bias in the data analysed by Bergsland and Vogt [1962]: the languages they studied were chosen because they were expected to be exceptions to Swadesh [1952]'s retention constant; this selection bias has already been pointed out by Sankoff [1970]. The important question is whether this rate heterogeneity is so strong as to prevent us from correctly estimating the age of internal nodes.

In the analyses presented in this section, we did not include catastrophical rate heterogeneity.

Our main interest in Chapter 5 is in dating ancestral nodes. The main question with these data is therefore whether we are able to reconstruct the known ages of ancient languages. Since the different versions of Old Norse are in close succession, removing the age constraint on one language and trying to reconstruct its age (as we did in Section 4.3 for the Ringe et al. [2002] data) serves little purpose: the reconstruction will always cover the constraint, because there are constraints on the other languages which effectively still constrain the language's age. Removing all constraints at the same time (as we did for the Germanic clade and its subclades in Section 4.3) is not possible, since we would then have no age constraints left and therefore no means of estimating the parameters and ages. Instead, we analysed a data set restricted to only 13^{th} century Old Norse and the 4 modern languages, and fixed $\mu = 1.86 \cdot 10^{-4}$, the mean value from our main analysis of the Ringe et al. [2002] data presented in Chapter 5 (a better though less practical approach would have been to draw samples from the posterior distribution for μ which we sample in Section 5.1; this would have increased the standard deviation of

our estimate of the age of Old Norse). We ignored the known age constraint on Old Norse and attempted to reconstruct the age of this ancient language. The age constraint is 660-760BP (assuming the data were collected in 1960), and the 95% HPD interval for the reconstructed age is 615-872BP, completely covering the constraint. This shows well that even if one language changes at a very different rate, the presence of several languages still allows us to correctly reconstruct the age of ancient languages. Rather than use a value of μ taken from the posterior sample of another analysis, it would be interesting to get an age constraint for another internal node of the tree (such as the most recent common ancestor to Riksmal, Sandnes and Gjestal) and use this constraint to jointly estimate μ and reconstruct the age of Old Norse; unfortunately, we have no such temporal information at our disposal.

Bergsland and Vogt [1962]'s point remains valid when we try to reconstruct the age of a leaf. The branch between 13^{th} century Old Norse and Modern Icelandic should be of length between 660 and 760 years. We analysed the data with Riksmal, Sandnes and Gjestal constrained to be modern leaves, and with the century-wide time constraints on the differents flavours of Old Norse, but allowed the age of Icelandic to vary. The 95% HPD interval for the length of the branch between 13^{th} century Old Norse and Modern Icelandic is 100 – 221 years, far from the correct range. Had we tried to predict the age of Icelandic, we would have failed. In this aspect, the findings of Bergsland and Vogt [1962] are confirmed: Icelandic did change at a very low rate. Nonetheless, this low rate on one branch does not impact significantly our estimates of over ages and parameters: the leaf age estimation is exposed to error on the leaf branch, whereas the clade root age is more robust.

Finally, we analysed the complete data with all age constraints included. Since the topology and all the node ages are known, the only parameter of interest is the death rate μ, which we estimate at $\mu = 2.47 \cdot 10^{-4} \pm 4.0 \cdot 10^{-5}$ deaths/year. This is in line with the estimates for the Ringe et al. [2002] and Dyen et al. [1997] data sets presented in Chapter 5 (mean values $\mu = 1.86 \cdot 10^{-4}$ and $\mu = 2.37 \cdot 10^{-4}$, respectively). Remember that Icelandic and Old Norse were chosen by Bergsland and Vogt [1962] because they were the most extreme example they could find; yet our estimate of μ is in the same range as for other data sets, and we were able to reconstruct the age of 13^{th} century Old Norse. This is a good example of the main advantage of modern phylogenetic methods against glottochronology: the ability to include many languages in an analysis means that the natural variance in rates gets averaged out and we can still compute viable estimates.

The second data set studied by Bergsland and Vogt [1962] comprises Old Georgian (5^{th} century), Modern Georgian and Mingrelian, a language spoken in West Georgia. Modern Georgian is assumed to descend directly from Old Georgian. The age of its most recent common ancestor with Mingrelian is not known, but it "probably has to placed in the last millenium B.C.", according to Bergsland and Vogt [1962], and anyway before the 5^{th} century A.D.. Yet when they used glottochronology on the pair Modern Georgian-Mingrelian, Bergsland and Vogt [1962] reconstructed an age of about 1300BP, *i.e.* in the 7^{th} century A.D., which they call "much too young".

In this case, it is easy to see why the reconstructed age is off the mark: through is tormented history, Mingrelian has borrowed many words from

Georgian. As is usually the case, these words can be detected thanks to their irregular phonological characteristics. Helpfully, Bergsland and Vogt [1962] specify which words have been borrowed, but they do not discard them from their data, so we were able to register the data using the same standard procedure as Ringe et al. [2002], excluding borrowings from our data. The argument for excluding borrowed words is quite clear: when Mingrelian borrowed a word from Georgian, the orginal word must have already died out in Mingrelian; treating borrowed words as present in the data would therefore systematically underestimate the deat rate μ. We once again fixed the death parameter μ to the mean value from our analysis of the Ringe et al. [2002] data ($\mu = 1.86 \cdot 10^{-4}$). The 95% HPD interval for the age of the most recent common ancestor to Georgian and Mingrelian is 2065–3170 BP, which coincides almost exactly with the last millenium B.C.

We were not able to use the third set of languages studied by Bergsland and Vogt [1962], which is the couple Old Armenian and Modern Armenian, because the word lists were not clear enough to be transcribed into binary data. Bergsland and Vogt [1962] claim that this another case where languages changed at a rate slower than usual. However, the argument for including many languages rather than only two still holds, and it would be interesting to see whether their results are identical with other related languages in the same analysis.

We certainly do not disagree with Bergsland and Vogt [1962] that there is variability in the number of changes that occur per millenium in different languages. However, this variability is taken into account by stochastic models

and in general, it does not hinder our ability to date ancestral languages or to reconstruct the model parameters. It exposes us more to errors on leaf ages, but that is not an estimation problem of great interest to us.

In their conclusions, Bergsland and Vogt [1962] claim that the rate heterogeneity they have exhibited prevents statistical methods from correctly estimating dates, and that "it follows" that the topology cannot be correctly reconstructed either. This ipse-dixitism is in fact far from obvious: even if it were the case that our methods could not correctly estimate dates, they may well be able to reconstruct the topology, just like uncertainty in the topology does not imply that we cannot reconstruct dates. In fact, linguists have so far tended to accept with more ease that statistical methods are able to infer topologies than dates [McMahon and McMahon, 2005].

6.2 Swabian dialects

This section presents analyses of a data set of linguistic features of 14 Swabian dialects (spoken in Baden-Würtemberg and Bavaria). The data were initially collected by Konig [1989] on 2,400 maps, which were then transcribed into binary data by Rother [pers. comm.]. The data cover lexical, grammatical and phonological features and are divided into 14 categories: human body, community and clothing; farming, weather, wild fauna and flora; food, housework, time and adverbs; cattle and pets; ; crops; woodwork and transportation (lexical); vowel quantities; short vowels ; long vowels and diphtongs; plosive consonants; other consonants (phonological); verb forms; noun and article forms; pronoun and adjective forms and syntax

(grammatical).

The temporal and geographical scales of these data are much smaller than for the Indo-European family. As such, we might expect large amounts of borrowing and a tree model will probably not capture the entire linguistic history of these dialects. Nonetheless, it is certainly interesting to see whether reconstructed trees on these data using our model corresponds to known historical facts.

The rough geographical position of the 14 dialects in Bavaria and Swabia is shown in Figure 6.3. The river Lech, also shown in Figure 6.3, is a strong physical boundary and is expected to split the dialects into two subgroups on either side of it. According to Rother [pers. comm.], we should expect to see two subtrees, corresponding to the two sides of the river. On both sides, the Swabians progressed from North to South, so we should also hope to see this in the reconstructed trees.

Figure 6.4 shows a consensus tree for an analysis of the Household lexicon data. The only constraint we imposed was that the root lie between 1200 and 1500 BP. The consensus tree corresponds closely to the geographical features: we observe a clear East-West cut and on both sides, a progression from North to South. This corresponds to the path along which it is believed the Swabians colonized the region, although it is also possible that the data contain a signal for this topology because lexical items were borrowed along this path. Many catastrophes are found on the branch leading to the Ostfranken dialect. Ostfranken is known to be problematic because the data may not correctly reflect this dialect.

Our model was developed with lexical data in mind; in particular, the

Figure 6.3: *Rough geographical position of 14 Swabian dialects. The thick blue line represents the Lech river, a strong physical boundary between the East and West of this region.*

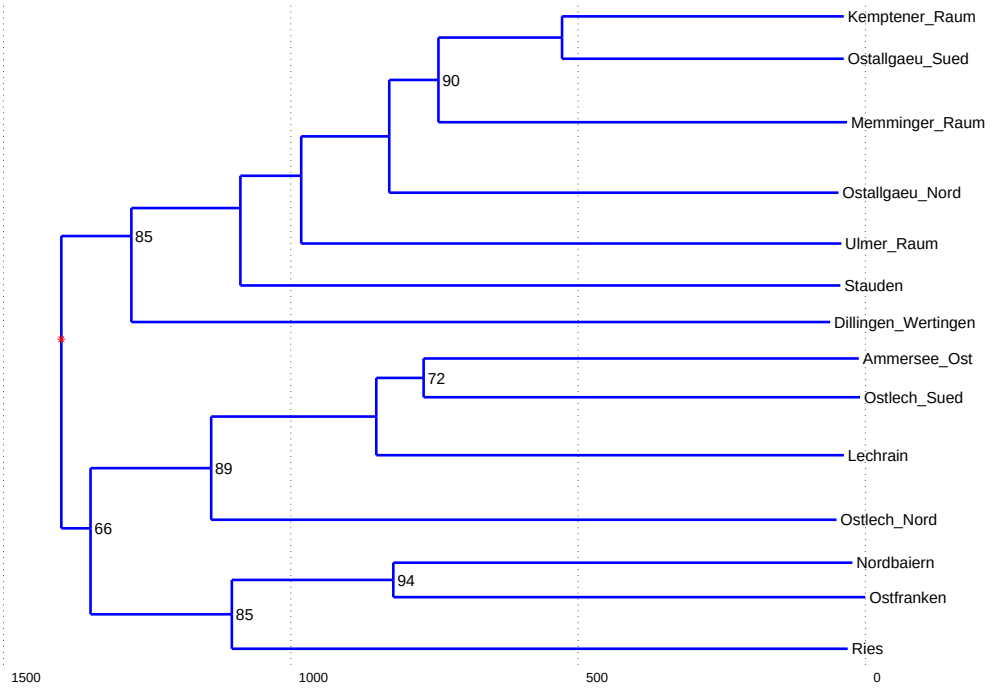

Figure 6.4: *Consensus tree for an analysis of the Household lexicon data of Swabian dialects.*

feature that a word can only be born once on the tree does not apply to syntactical or phonological data, where a finite site approach such as the one described by Lewis [2001] may be appropriate. As mentioned in Chapter 1, it is also debatable whether a tree model is appropriate for such data. Indeed, our analyses of phonological and syntactical data produced much less convincing results as the lexical data.

6.3 Punctuational bursts

Catastrophes were initially introduced in our model to take into account possible "catastrophic" rate heterogeneity. However, another possible use of this feature is to analyse what Atkinson et al. [2008] call "punctuational bursts". Atkinson et al. [2008] wished to test the hypothesis that languages diversify at two rates: a basic, slow rate, along the entire tree (the anagenic process); and a faster rate whenever two languages split (punctuational bursts). There are several plausible explanations for punctuational bursts. When a language is founded by the migration of a small subset of a population, the "founder effect" leads to fast change over a short period of time. Language splitting is also associated with other events which can lead to fast change, such as war.

Atkinson et al. [2008] study three language families: Indo-European (a subset of the Dyen et al. [1997] data), Austronesian and Bantu. In order to test for punctuational bursts, they construct language trees for each family using the software BayesTraits [Pagel et al., 2004]. For data in which all the leaves are modern (*i.e.* isochronous), they then plot the number of changes between each leaf and the root of the tree against the number of internal nodes

between the leaf and the root and find a positive correlation, which is a signal for the existence of punctuational bursts. They calculate that punctuational bursts account for 21% of changes in the history of the Indo-European family.

A more direct way to test for punctuational bursts is to include them directly in the model. Indeed, if punctuational bursts occur on the tree, they are equivalent to a catastrophe event occuring after every splitting event. Recall however that the position of a catastrophe on a branch does not matter. To include punctuational bursts in our model, we therefore simply need to impose the presence of exactly one catastrophe on each branch, as shown in Figure 6.5. With L languages, there will be $2L-2$ catastrophes on the tree g, each corresponding to an effective length T_C. If the total length of the tree is denoted by $|g|$, the proportion of change attributable to punctuational bursts is $\frac{(2L-2)T_C}{|g|+(2L-2)T_C}$.

We used this model to analyse the Ringe et al. [2002] data as well as the same subset of the Dyen et al. [1997] data as Atkinson et al. [2008]. The reason for taking only a subset of the data is that some of the languages listed by Dyen et al. [1997] could be deemed too close for their divergence to be truly regarded as a language splitting event. Therefore, 22 languages were removed from the data. It is worth remembering from Section 3.1 that with one catastrophe on every branch and with the improper prior distribution on the death rate $p(\mu) \propto 1/\mu$, the posterior distribution can also be improper when $\mu \to 0$. This is not an issue in this case, as the likelihood is very small for small values of μ and so the MCMC never visits this part of the state-space, so we can impose a very conservative cutoff, rendering the posterior distribution proper without altering the MCMC.

Our results are starkly different to those obtained by Atkinson et al. [2008]. In our analyses, the probability of death at a catastrophe κ is very close to 0, implying that punctuational bursts have close to no effect. For the Ringe et al. [2002] data, our model attributes 0.39% of all changes to punctuational bursts; for the Dyen et al. [1997] data, 0.23% of all changes are attributed to punctuational bursts.

Further work is needed to understand the discrepancy between our results and those of Atkinson et al. [2008]. Atkinson et al. [2008] show through simulation studies that the presence of borrowing in the data does not impact their model of punctuational bursts. It would however be interesting to see whether the presence of catastrophical rate heterogeneity, for which there is a signal in the data as shown in Section 5.1 and 5.2, introduces a bias for or against punctuational bursts in either methodology. Note also that our method assumes equal-sized catastrophes on all branches, and that unlike Atkinson et al. [2008], our model of language change is clock-like; this might explain the discrepancy.

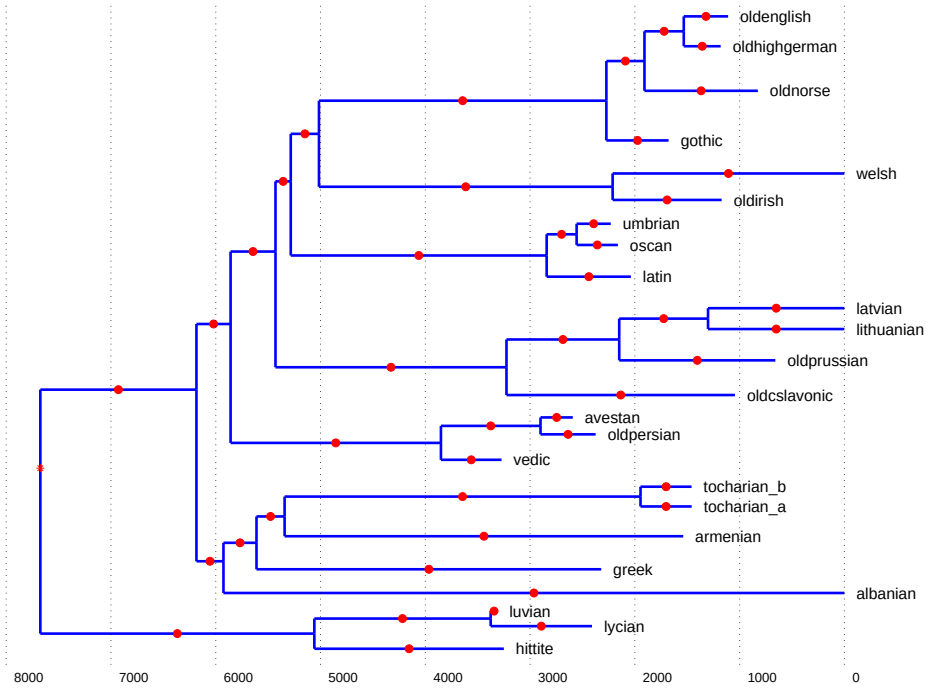

Figure 6.5: *Example tree with punctuational bursts: one catastrophe on each branch.*

Appendix A

Swadesh list

This appendix gives all 207 meaning categories in the standard Swadesh list, used by Ringe et al. [2002], Dyen et al. [1997] and Bergsland and Vogt [1962]. There have been a number of minor adjustements suggested for this list. The major refinement was when Swadesh shortened the list to 100 meaning categories. The meanings which are also included in the shorter list are indicated in bold.

1. **all**
2. and
3. animal
4. **ashes**
5. at
6. back
7. bad
8. **bark**
9. because
10. **belly**
11. **big**
12. **bird**
13. **bite**
14. **black**
15. **blood**
16. blow
17. **bone**
18. **breast**
19. breathe
20. **burn**
21. child
22. **claw**
23. **cloud**
24. **cold**
25. **come**
26. count
27. cut
28. day
29. die
30. **dig**
31. dirty
32. **dog**
33. **drink**
34. **dry**
35. dull
36. dust

37. **ear**
38. **earth**
39. **eat**
40. **egg**
41. **eye**
42. fall
43. far
44. **fat**
45. father
46. fear
47. **feather**
48. few
49. fight
50. **fire**
51. **fish**
52. five
53. float
54. flow
55. flower
56. **fly**
57. fog
58. **foot**
59. four
60. freeze
61. **full**
62. **give**
63. **good**
64. grass
65. **green**
66. guts
67. **hair**
68. **hand**
69. he
70. **head**
71. **hear**
72. **heart**
73. heavy
74. here
75. hit
76. hold
77. **horn**
78. how
79. hunt
80. husband
81. **I**
82. ice
83. if
84. in
85. **kill**
86. **knee**
87. **know**
88. lake
89. laugh
90. **leaf**
91. left
92. leg
93. **lie**
94. live
95. **liver**
96. **long**
97. **louse**
98. **man**
99. **many**
100. **meat**
101. **moon**
102. mother
103. **mountain**
104. **mouth**
105. **name**
106. narrow
107. near
108. **neck**
109. **new**
110. **night**
111. **nose**
112. **not**
113. old
114. **one**
115. other
116. **person**
117. play
118. pull
119. push
120. **rain**
121. **red**
122. right (correct)
123. right (side)
124. river
125. **road**
126. **root**
127. rope
128. rotten
129. **round**
130. rub
131. salt
132. **sand**

133. **say**
134. scratch
135. sea
136. **see**
137. **seed**
138. sew
139. sharp
140. short
141. sing
142. **sit**
143. **skin**
144. sky
145. **sleep**
146. **small**
147. smell
148. **smoke**
149. smooth
150. snake
151. snow
152. some
153. spit
154. split
155. squeeze
156. stab
157. **stand**
158. **star**
159. stick
160. **stone**
161. straight
162. suck
163. **sun**
164. swell
165. **swim**
166. **tail**
167. ten
168. **that**
169. there
170. they
171. thick
172. thin
173. think
174. **this**
175. **thou**
176. **three**
177. throw
178. tie
179. **tongue**
180. **tooth**
181. **tree**
182. turn
183. **two**
184. vomit
185. **walk**
186. **warm**
187. wash
188. **water**
189. **we**
190. wet
191. **what**
192. when
193. where
194. **white**
195. **who**
196. wide
197. wife
198. wind
199. wing
200. wipe
201. with
202. **woman**
203. woods
204. worm
205. ye
206. year
207. **yellow**

Table of notations

The following table gives a brief description of all the notations used in this thesis, as well as the page where the notation is introduced. Due to the limited size of the alphabets, some notations have several meanings, but the context always prevents any confusion.

Notation	Type	Description	Page
A	node	the Adam node of a tree, linked by a branch of infinite length to the root	33
a	integer	index of a cognacy class or a column of D; $1 \leq a \leq N$	26
B	matrix	any matrix (in practice, one of D, D^*, I)	27
B	set of edges	the set of edges with at least $L - d'$ descendants	64
$B_{C,C-c}$	real	a Bayes factor	96
b	real	the "level" of borrowing: the rate of borrowing is $b\mu$	84
C	integer	the number of constraints	33
C	real	a finite constant	58
C	real	the proportion of shared cognates between two languages	114
c	constraint	a constraint on the tree	33
c_1, c_2	node	the children nodes of node i	48
D	$L \times N$ matrix	data. Entries can be 0, 1 or ?	26
$\mathbf{D}^*$	$L \times N^*$ random matrix	notional full random binary data matrix	41
D^*	$L \times N^*$ matrix	realization of $\mathbf{D}^*$	41
$\mathcal{D}_a$	set of vectors	set of vectors d^* allowed in column a	27

Notation	Type	Description	Page
$\tilde{D}$	$L \times N$ matrix	masked version of D^*; corresponds to D^* with some entries replaced with ?, and to D with extra unregistered columns	42
d, d'	integers	integers involved in the registration conditions; usually equal to 0 or 1	62
d^*	binary vector	a column allowed by the data	27
E	set of edges	the set of all edges in a tree	33
$\mathcal{E}$	event	the event that a cognacy class is registered	45
$E_{\langle i,j \rangle}$	set of edges	the set of edges neighbouring edge $\langle i, j \rangle$	70
${}_2F_0$	function	a generalized hypergeometric function	54
f	function	a prior distribution	
f_G	function	the prior distribution on trees	34
g	tree	a binary rooted tree	33, 40
$\mathbf{I}^*$	$L \times N^*$ random matrix	random indicator matrix of observations	42
I^*	$L \times N^*$ matrix	realization of $\mathbf{I}^*$	42
i	integer	index of a language or node of a tree, or a row of the data D; $1 \leq i \leq L$	26
j	node	a node of a tree, often the parent of node i	
k	vector of integers	vector of number of catastrophes	40
k_i	integer	the number of catastrophes on edge $\{i, j\}$	40
L	integer	number of languages	26
$L(x)$	function	the likelihood function	
M	set of integers	a cognacy class; $M \subseteq \{1, 2, ..., L\}$	26
M^C	model	the null model with all constraints included	96
M^{C-c}	model	the model with constraint c removed	96
m	set of integers	list of languages displaying a cognacy class in the unobserved complete data	27
m_a	set of leaves	list of leaves known to display cognacy class a	49
N	integer	number of cognacy classes	26
$N*$	integer	number of cognacy classes born on the tree, including unregistered classes	41

Notation	Type	Description	Page
N_{samp}	integer	the number of samples in a MCMC run	71
p	function	a prior distribution	
Q	function	function counting the number of ?'s in a column	42
Q_a	integer	number of languages for which the data for cognacte a is ?	43
q	function	a MCMC proposal distribution	68
q_i	integer	the number of edges neighbouring edge $\langle i, j \rangle$	70
R	function	registration rule	42
r	node	the root of a tree	33
r	real	Swadesh's "retention rate"	115
$r_{i,j}$	real	a transition rate	53
r_S	function	the autocorrelation function of statistic S of a Markov Chain	71
S	set of nodes	the set of nodes having ages not bounded above by a constraint	34
s_i	integer	the number of leaves descended from node i	48
T	real	upper bound on the age of the root	33
T_b	real	the time after which local borrowing is no longer allowed	84
T_C	real	amount of time equivalent to a catastrophe	40
t_i	real	the age of node or leaf i	33
t_r	real	the age of the root r	33
t	vector of reals	the vector of ages of nodes in a tree; $t = (t_1, t_2, ..., t_A)$	33
$u_i^{(n)}$	real	the probability for a cognacte class present at node i to be registered at exactly n leaves below i	48
V	set of nodes	the set of all nodes of a tree	33
V_L	set of leaves	the set of all leaves of a tree; $V_L \subset V$	33
$V_L^{(i)}$	set of leaves	the set of leaves descended from node i	48
$v_i^{(n)}$	real	the probability for a cognacte class present at node i to be registered or missing at exactly n leaves below i	48
(X_n)	random process	a Markov chain	71
x	state	a MCMC state	67

Notation	Type	Description	Page
Y	function	function counting the number of 1's in a column	42
Y_a	integer	number of languages for which the data for cognacte a is 1	43
Z_D	point process	the point process of birth locations on the tree	44
z	(node, time) couple	a point on the tree	44
α	real	a MCMC acceptation probability	68
α, β	reals	the parameters of a Gamma distribution	57
Γ	set of trees	the set of all rooted binary trees with L distinguishable leaves	33
$\Gamma^{(c)}$	set of trees	the set of trees obeying constraint c	33
Γ^{C}	set of trees	the set of trees obeying all constraints	34
$\delta_{i,j}$	real	the probability for a cognacy class to survive down edge $\langle i, j \rangle$	48
ϵ	real	the length of time between deaths and births at a catastrophe; very close to 0	53
η	real	an arbitrary factor	51
θ	list of parameters	condensed notation for a list of parameters, for example $\theta = (g, \kappa, \rho, \xi)$	57
κ	real	probability of death of a cognate at a catastrophe	38
$\Lambda([g])$	real	normalizing constant: $\Lambda([g]) = \int_{[g]} \tilde{\lambda}(z)\, dz$	45
λ	real	birth rate of cognacy classes outside of catastrophes	38
$\tilde{\lambda}$	function	effective birth rate of cognacy classes at a point on the tree	45
μ	real	death rate of cognacy classes outside of catastrophes	38
ν	real	mean number of births at a catastrophe	38
ξ_i	real	probability for a cognate class to be missing at leaf i	40
ξ	vector of reals	the vector $(\xi_1, \ldots, \xi_L$	
π_i	real	the equilibrium probability of displaying i cognacy classes	54
ρ	real	rate of occurrence of catastrophes	38
$\sigma(t)$	permutation	order of the internal ages t	33
τ	real	a time on an edge of the tree	44

Notation	Type	Description	Page
τ_S	real	the autocorrelation time of statistic S	71
Ω_a	set of sets of integers	set of cognacy classes consistent with the data for cognacy class a	27
Ω	list of sets of sets of integers	the vector $(\Omega_1, \Omega_2, ..., \Omega_N)$	27
ω	set of integers	a possible value for the list of languages at which a cognacy class is displayed	27

Bibliography

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition, 1964. ISBN 0-486-61272-4.

A. Alekseyenko, C. Lee, and M. Suchard. Wagner and Dollo: A Stochastic Duet by Composing Two Parsimonious Solos. *Systematic Biology*, 57(5): 772–784, 2008.

Q. Atkinson, G. Nicholls, D. Welch, and R. Gray. From words to dates: water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society*, 103(2):193–219, 2005.

Q. Atkinson, A. Meade, C. Venditti, S. Greenhill, and M. Pagel. Languages evolve in punctuational bursts. *Science*, 319(5863):588, 2008.

A. Barbrook, C. Howe, N. Blake, and P. Robinson. The phylogeny of the Canterbury Tales. *Nature*, 394(6696):839, 1998.

R. S. Beekes. *Comparative Indo-European Linguistics An introduction*. Amsterdam: Benjamins, 1995.

M. Ben Hamed, P. Darlu, and N. Vallée. On Cladistic Reconstruction of Linguistic Trees through Vowel Data. *Journal of Quantitative Linguistics*, 12(1):79–109, 2005.

K. Bergsland and H. Vogt. On the validity of glottochronology. *Current Anthropology*, 3(2):115, 1962.

R. Blust. Why lexicostatistics doesn't work: the 'universal' constant hypothesis and the Austronesian languages. *Time depth in historical linguistics*, 2:311–31, 2000.

A. Bouchard-Côté, P. Liang, T. Griffiths, and D. Klein. A Probabilistic Approach to Diachronic Phonology. *Empirical Methods in Natural*

Language Processing and Computational Natural Language Learning (EMNLP/CoNLL), 2007.

D. Bryant, F. Filimon, and R. Gray. Untangling our past: Languages, trees, splits and networks. *The Evolution of Cultural Diversity: a Phylogenetic Approach*, pages 69–85, 2005.

L. Campbell. Languages and genes in collaboration: some practical matters. In *Language and genes: An interdisciplinary conference, University of California Santa Barbara*, pages 8–10, 2006.

L. Cavalli-Sforza, A. Piazza, P. Menozzi, and J. Mountain. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences*, 85(16):6002–6006, 1988.

M. Collard and S. Shennan. Ethnogenesis versus phylogenesis in prehistoric culture change: a case study using European neolithic pottery and biological phylogenetic techniques. *Archaeogenetics: DNA and the population prehistory of Europe. Cambridge: McDonald Institute for Archaeological Research. p*, pages 89–97, 2000.

B. Comrie. Position paper for the Languages and Genes conference at UCSB, September 2006.

W. Croft. Evolutionary linguistics. *Annual Review of Anthropology*, 37:219–234, 2008.

C. Darwin. *The Descent of Man, and Selection in Relation to Sex*. 1871.

J. Diamond and P. Bellwood. Farmers and their languages: the first expansions. *Science*, 300(5619):597–603, 2003.

L. Dollo. Les lois de l'évolution. *Bulletin de la Société belge de Géologie, de Paléontologie et d'Hydrologie*, 7:164–166, 1893. Translated in Gould [1970].

A. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214, 2007.

A. Drummond, G. Nicholls, A. Rodrigo, and W. Solomon. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics*, 161(3):1307–1320, 2002.

M. Dryer, M. Haspelmath, D. Gil, and B. Comrie. World Atlas of Language Structures, 2003.

M. Dunn, A. Terrill, G. Reesink, R. Foley, and S. Levinson. Structural phylogenetics and the reconstruction of ancient language history, 2005.

M. Dunn, S. Levinson, E. Lindström, G. Reesink, and A. Terrill. Structural phylogeny in historical linguistics: methodological explorations applied in Island Melanesia. *Language*, 84(4), 2008.

I. Dyen, J. Kruskal, and B. Black. FILE IE-DATA1. Raw data available from http://www.ntu.edu.au/education/langs/ielex/IE-DATA1. Binary data available from http://www.psych.auckland.ac.nz/psych/research/RusselsData.htm., 1997.

S. Embleton. *Statistics in historical linguistics*. Brockmeyer, 1986.

J. Ensminger. Transaction costs and Islam: explaining conversion in Africa. *Journal of Institutional and Theoretical Economics*, 153:4–29, 1997.

S. Evans, D. Ringe, and T. Warnow. Inference of divergence times as a statistical inverse problem. *Phylogenetic Methods and the Prehistory of Languages. McDonald Institute Monographs*, pages 119–130, 2004.

J. Farris. Phylogenetic Analysis Under Dollo's Law. *Systematic Zoology*, 26 (1):77–88, 1977.

J. Felsenstein. *Inferring phylogenies*. Sinauer Associates Sunderland, Mass., USA, 1981.

W. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic zoology*, 20(4):406–416, 1971.

W. Fitch. LinguisticsAn invisible hand. *Nature*, 449(7163):665–667, 2007.

P. Forster and A. Toth. Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences*, 100(15):9079–9084, 2003.

L. Fortunato, C. Holden, and R. Mace. From Bridewealth to Dowry?: A Bayesian Estimation of Ancestral States of Marriage Transfers in Indo-European Groups. *Human Nature*, 17(4):355–376, 2006.

A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press New York, 2007.

C. Geyer. Practical markov chain monte carlo. *Statistical Science*, 7(4):473–483, 1992.

M. Gimbutas and H. Hencken. *The prehistory of eastern Europe*. Peabody Museum, 1956.

H. Glenner, A. Hansen, M. Sørensen, F. Ronquist, J. Huelsenbeck, and E. Willerslev. Bayesian Inference of the Metazoan Phylogeny A Combined Molecular and Morphological Approach. *Current Biology*, 14(18):1644–1649, 2004.

P. Goloboff and D. Pol. Parsimony and Bayesian phylogenetics. *Parsimony, Phylogeny, and Genomics*, pages 148–159, 2005.

S. Gould. Dollo on Dollo's law: irreversibility and the status of evolutionary laws. *Journal of the History of Biology*, 3(2):189–212, 1970.

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

R. Gray and F. Jordan. Language trees support the express-train sequence of Austronesian expansion. *Nature*, 405:1052–1055, 2000.

R. Gray, S. Greenhill, and R. Ross. The pleasures and perils of Darwinizing culture (with phylogenies). *Biological Theory*, 2(4):360–375, 2007.

J. Greenberg. *Language in the Americas*. Stanford University Press, 1987.

S. Greenhill, T. Currie, and R. Gray. Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society B*, 2009.

C. Holden. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proceedings of the Royal Society B: Biological Sciences*, 269(1493):793–799, 2002.

C. Holden and R. Mace. Spread of cattle led to the loss of matrilineal descent in Africa: A coevolutionary analysis. *Proceedings: Biological Sciences*, 270 (1532):2425–2433, 2003.

H. Holm. The new arboretum of Indo-European ï£¡treesï£¡. Can new algorithms reveal the phylogeny and even prehistory of Indo-European?*. *Journal of Quantitative Linguistics*, 14(2):167–214, 2007.

J. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees, 2001.

D. Huson and M. Steel. Phylogenetic trees based on gene content. *Bioinformatics*, 20(13):2044, 2004.

D. Hymes. Lexicostatistics so far. *Current Anthropology*, 1(1):3, 1960.

J. Inoue, P. Donoghue, and Z. Yang. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol*, 59(1):74–89, 2010.

D. Jones. Kinship and deep history: exploring connections between culture areas, genes, and languages. *American anthropologist*, 105(3):501–514, 2003.

P. Jordan and S. Shennan. Cultural transmission, language, and basketry traditions amongst the California Indians. *Journal of Anthropological Archaeology*, 22(1):42–74, 2003.

A. Kitchen, C. Ehret, S. Assefa, and C. Mulligan. Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proceedings of the Royal Society B*, 276(1668):2703, 2009.

W. Konig. Sprachatlas von Bayerisch-Schwaben., 1989.

J. Lansing, M. Cox, S. Downey, B. Gabler, B. Hallmark, T. Karafet, P. Norquest, J. Schoenfelder, H. Sudoyo, J. Watkins, et al. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proceedings of the National Academy of Sciences*, 104(41):16022, 2007.

B. Larget and D. Simon. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16:750–759, 1999.

W. Le Quesne. Further Studies Based on the Uniquely Derived Character Concept. *Systematic Zoology*, 21(3):281–288, 1972.

P. Lewis. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology*, 50(6):913–925, 2001.

E. Lieberman, J. Michel, J. Jackson, T. Tang, and M. Nowak. Quantifying the evolutionary dynamics of language. *Nature*, 449(7163):713–716, 2007.

R. Mace and C. Holden. A phylogenetic approach to cultural evolution. *Trends in Ecology & Evolution*, 20(3):116–121, 2005.

W. Mackay and G. Kondrak. Computing word similarity and identifying cognates with Pair Hidden Markov Models. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)*, pages 40–47, 2005.

J. Mallet. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20(5):229–237, 2005.

J. Mallory. *In search of the Indo-Europeans: language, archaeology and myth.* Thames and Hudson, 1989.

E. Marris. The Language Barrier. *Nature*, 453(7194):446–8, 2008.

A. McMahon and R. McMahon. *Language Classification by Numbers.* Oxford Linguistics, 2005.

A. McMahon and R. McMahon. Why linguists donï£¡t do dates: evidence from Indo-European and Australian languages. *Phylogenetic Methods and the Prehistory of Languages*, page 153, 2006.

A. Mesoudi, A. Whiten, K. Laland, and R. Harrison. Perspective: Is human cultural evolution Darwinian? Evidence reviewed from the perspective of The Origin of Species. *Evolution*, 58(1):1–11, 2004.

E. Moravcsik. Language contact. *Universals of human language*, 1:93–123, 1978.

M. Mulder, M. George-Cramer, J. Eshleman, and A. Ortolani. A study of East African kinship and marriage using a phylogenetically based comparative method. *American anthropologist*, pages 1059–1082, 2001.

L. Nakhleh, D. Ringe, and T. Warnow. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2):382–420, 2005.

G. Nicholls and R. Gray. Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):545–566, 2008.

J. Nichols. *Linguistic Diversity in Space and Time.* University of Chicago Press, 1999.

C. Nunn, M. Mulder, and S. Langley. Comparative methods for studying cultural trait evolution: A simulation study. *Cross-Cultural Research*, 40 (2):177, 2006.

M. O'Brien and R. Lyman. Evolutionary archeology: Current status and future prospects. *Evolutionary Anthropology*, 11(1):26–35, 2002.

M. O'Brien, J. Darwent, and R. Lyman. Cladistics is useful for reconstructing archaeological phylogenies: Palaeoindian points from the southeastern United States. *Journal of Archaeological Science*, 28(10):1115–1136, 2001.

M. O'Brien, R. Lyman, Y. Saab, E. Saab, J. Darwent, and D. Glover. Two issues in archaeological phylogenetics: taxon construction and outgroup selection. *Journal of theoretical biology*, 215(2):133–150, 2002.

M. Pagel. A method for the analysis of comparative data. *Journal of Theoretical Biology*, 156(4):431–442, 1992.

M. Pagel. Statistical analysis of comparative data. *Trends in Ecology and Evolution*, 15(10):418, 2000.

M. Pagel. Limpets break Dollo's law. *Trends in Ecology & Evolution*, 19(6): 278–280, 2004.

M. Pagel and A. Meade. Estimating rates of lexical replacement on phylogenetic trees of languages. In *Phylogenetic Methods and the Prehistory of Languages*, volume 1, page 173. McDonald Institute for Archaeological Research, 2006.

M. Pagel, A. Meade, and D. Barker. Bayesian estimation of ancestral character states on phylogenies. *Systematic biology*, 53(5):673–684, 2004.

M. Pagel, Q. Atkinson, and A. Meade. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449(7163): 717–720, 2007.

J. Parsons. *The Remains of Japhet, being historical enquiries into the affinity and origins of the European languages*. 1767.

A. Raftery. Hypothesis testing and model selection. In W. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*. Chapman & Hall / CRC, 1996.

B. Rannala and Z. Yang. Inferring speciation times under an episodic molecular clock. *Systematic biology*, 56(3):453–466, 2007.

C. Renfrew. Archaeology and Language. The Puzzle of Indo-European Origins. *Current Anthropology*, 29:437–441, 1987.

K. Rexova, D. Frynta, and J. Zrzavy. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics*, 19(2): 120–127, 2003.

D. Ringe, T. Warnow, and A. Taylor. Indo-European and Computational Cladistics. *Transactions of the Philological Society*, 100(1):59–129, 2002.

F. Roe. *The Indian and the horse*. University of Oklahoma Press, 1955.

F. Ronquist, J. Huelsenbeck, and P. van der Mark. MrBayes 3.1 Manual. *School of Computational Science, Florida State University*, 2005.

D. Sankoff. On the rate of replacement of word-meaning relationships. *Language*, pages 564–569, 1970.

A. Schleicher. *Linguistische Untersuchungen. 2. Teil: Die Sprachen Europas in systematischer Übersicht*. Bonn: HB König, 1850.

A. Schleicher. Eine fabel in indogermanischer ursprache. *Beitr "age zur vergleichenden Sprachforschung auf dem Gebiete der arischen, keltischen und slawischen Sprachen*, 5:206–208, 1868.

J. Schmidt. *Die Verwandtschaftsverhï£¡ltnisse der indogermanischen Sprachen*. Bonn, 1872.

M. Serva and F. Petroni. Indo-European languages tree by Levenshtein distance. *EPL-Europhysics Letters*, 81(6):68005–69000, 2008.

S. Shennan and K. Edinborough. Prehistoric population history: from the Late Glacial to the Late Neolithic in Central and Northern Europe. *Journal of Archaeological Science*, 34(8):1339–1345, 2007.

C. Skelton. Methods of Using Phylogenetic Systematics to Reconstruct the History of the Linear B Script. *Archaeometry*, 50(1):158–176, 2008.

N. Slaska. Lexicostatistics away from the armchair: handling people, props and problems. *Transactions of the Philological Society*, 103(2):221, 2005.

M. Spencer, B. Bordalejo, L. Wang, A. Barbrook, L. Mooney, P. Robinson, T. Warnow, and C. Howe. Analyzing the order of items in manuscripts of The Canterbury Tales. *Computers and the Humanities*, 37(1):97–109, 2003.

M. Spencer, E. Davidson, A. Barbrook, and C. Howe. Phylogenetics of artificial manuscripts. *Journal of theoretical biology*, 227(4):503–511, 2004.

M. Swadesh. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, pages 452–463, 1952.

S. Thomason. On the unpredictability of contact effects. *Sociolinguistic Studies*, 1(1):173, 2000.

S. Thomason. Contact-induced typological change. *Language typology and language universals: An international handbook*, pages 1640–1648, 2001.

J. Thorne and H. Kishino. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology*, 51(5):689–702, 2002.

J. Thorne, H. Kishino, and I. Painter. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 15(12):1647, 1998.

C. Turney and H. Brown. Catastrophic early Holocene sea level rise, human migration and the Neolithic transition in Europe. *Quaternary Science Reviews*, 26(17-18):2036–2041, 2007.

J. Velasco. The prior probabilities of phylogenetic trees. *Biology and Philosophy*, 23(4):455–473, 2008.

N. Wade. A Biological Dig for the Roots of Language. *The New York Times*, 153(52800):F1, 2004.

T. Warnow, S. Evans, D. Ringe, and L. Nakhleh. A Stochastic model of language evolution that incorporates homoplasy and borrowing. *Phylogenetic Methods and the Prehistory of Languages*, 2004.

Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Molecular Biology and Evolution*, 14(7):717, 1997.

Z. Yang and B. Rannala. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution*, 23(1):212, 2006.

Z. Yang, S. Kumar, and M. Nei. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141(4):1641, 1995.

G. Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87, 1925.