Disentangling Ancestral State Reconstruction in Historical Linguistics - comparing classic approaches and new methods with Oceanic grammar

Hedvig Skirgård[1]

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology. Leipzig, Germany.

June 19, 2023

Abstract

Ancestral State Reconstruction (ASR) is an essential part of historical linguistics (HL). Conventional ASR in HL relies on three core principles: fewest changes on the tree, plausibility of changes and plausibility of the resulting combinations of features in proto-languages. This approach has some problems, in particular the definition of what is plausible and the disregard of branch lengths. This study compares the classic approach of ASR to computational tools (Maximum Parsimony and Maximum Likelihood), conceptually and practically. Computational models have the advantage of being more transparent, consistent and replicable, and the disadvantage of lacking nuanced knowledge and context. Using data from the structural database Grambank, I compare reconstructions of the grammar of ancestral Oceanic languages from the historical linguistics literature to those achieved by computational means. The results show that there is a high degree of agreement between manual and computational approaches, with a tendency for classical HL to agree more with the approaches that ignore branch lengths. Taking into account branch lengths explicitly is more conceptually sound, as such the field of historical linguistics should engage in improving methods in this direction. A combination of computational methods and qualitative knowledge is possible in future and would be of great benefit.

Keywords: oceanic languages, ancestral state reconstruction, grammar

1 Introduction

Historical linguistics offers us a unique and insightful window into our human past. By reconstructing the paths languages take, we can learn about our history and infer the migration paths of people and cultures. By reconstructing the words, sounds and

grammar of ancient languages, we can learn about communities long gone. Historical linguistics is devoted to this endeavour and has made great strides in our understanding of human history since its inception. The field has established methods that have enabled us to classify languages into language families and reconstruct words and sounds of proto-languages (unobserved ancestors of observed languages). Conclusions from historical linguistics are also influential in other historical sciences, for example archaeology (c.f. **?**: S364).

Scholars who work in historical linguistics produce valuable and greatly inspiring work and they possess a wealth of knowledge not only of the languages themselves, but also the cultures, societies and history of the region. At times, it is difficult to be explicit about all the background information and context that goes into an analysis in historical linguistics — which makes it hard for someone else to replicate and examine the study thoroughly. When the information is made explicit in publications, it can be hard to aggregate across publications while keeping definitions and weights consistent.

In this study, we focus on one particular subset of the historical linguistics toolbox — the inference of earlier states of languages — and outline how computational approaches can be a complement that serves to increase speed, transparency and consistency. We discuss the underlying mechanisms of the conventional "manual" approaches to reconstruction in historical linguistics and compare the conceptual framework to computational alternatives. As a practical example, we compare reconstructions of Oceanic grammar. This study makes visible opportunities for methodological expansion that can be made by incorporating computational approaches into mainstream historical linguistics.

Historical linguists typically engage in three different tasks simultaneously: a) the identification of cognates and sound correspondences in languages, b) the inference of sub-grouping (networks/trees)[1] and c) the inference of sounds/forms/patterns in proto-languages (Ancestral State Reconstruction = ASR). In conventional approaches in historical linguistics, these three tasks are done at the same time and inform each other — they are necessarily interlinked. However, in historical analysis of biology and cultural evolution, these tasks are more separated out. Fig. **??** illustrates these three tasks for four different kinds of material: words (sounds & cognates), grammar, genes and biological features. The arrows indicate task workflow with information on words leading to the construction of trees, which in turn enables ASR on lexicon and grammar. This is mirrored in the biological sciences, with genome serving as the bases for the trees which then make ASR possible. The feedback loop between tree construction and ASR in the classical analysis of cognates and sound correspondences is illustrated with circular arrows forming a loop. The same is true of biological traits where biologist care about not predicting impossible ancestral states (**?**) and may therefore revise the tree-building if these occur. Furthermore, both linguists and biologist may return and re-examine their original classification of their data (task a; cognate coding, sample labelling, sequence alignment etc.) given the outcome of ASR (c.f. re-estimating sequence alignment in genetics while estimating trees (**?**). The three tasks are not

[1] C.f. how biological cladistics finds relationships between species based on shared derived characteristics from common ancestors (**?**: 16-17))

necessarily independent in the biological sciences, but it is possible to carry them out separately and the links between them are explicit.

It is clear that the depth of knowledge possessed by historical linguists greatly informs their work, and that there is indeed value in linking these three tasks to each other — for example by revising a tree when a reconstructed state does not make sense, classifying cognates of extant languages based on knowledge derived from elsewhere in the tree on how the changes can go etc. However, there are disadvantages as well. The first among these is the difficulty of providing a highly transparent methodology. This kind of labour involves a vast amount of knowledge and careful decisions, and it is not easy to make all of them explicit and accessible.

In this paper, we focus specifically on the task of ASR — we are not deriving new trees or re-coding structural data. In addition to increasing transparency, quantitative approaches to ASR also have the benefit of speed. If we can interrogate the conventional methods of ASR in historical linguistics and compare those principles to various computational approaches and evaluate the agreement, then we may be able to improve on transparency and offer historical linguists a convenient tool that can effectivize part of their labour.

We will be solely focusing on the third of these tasks: Ancestral State Reconstruction (ASR) and comparing the methodology and results from conventional historical linguistics to computational approaches using structural data from a large-scale typological database (Grambank v1.0, **?**).

One of the major differences between ASR in conventional historical linguistics and in biological and cultural evolution is the evaluation of appropriate data for phylogenetic analysis. Studies in historical linguistics typically require that the input data satisfies the Double Cognacy Criterion (**?**) — both for the construction of trees (task b in Fig. **??**) and ASR (task c). It is difficult to apply this test to non-vocabulary data because it is not clear what the correlates are to words and phonemes in structural data.

In cultural evolution and biology on the other hand, data is deemed appropriate for historical analysis if homoplasy can be excluded. Excluding homoplasy means that it is reasonable to assume that the tree in question estimates the history of the data etc (c.f. **?** and **?**). One of the most common approaches to test if data is valid to use for analysis with a particular tree is to test for statistically significant phylogenetic signal. Phylogenetic signal is the *tendency for related species to resemble each other more than they resemble species drawn at random* (**?**: 905). This concept is independent from measurements of conservatism of traits or species/languages. Tests of phylogenetic signal can be carried out for linguistic data as well as biological and cultural data, as we will learn more about in section **??**.

One of the drawbacks of conventional approaches to ASR in historical linguistics is that they typically involve a great deal of manual work and, as mentioned earlier, it can be difficult to be 100% transparent with all analytical decisions and their contexts. In particular, while there is often agreement on the presence of sound correspondences or cognate sets, there can often be conflicts regarding how to weight information and the plausibility of reconstructions. In contrast, computational phylogenetic methods are a set of tools that can be applied with great speed and all analysis is explicit and consis-

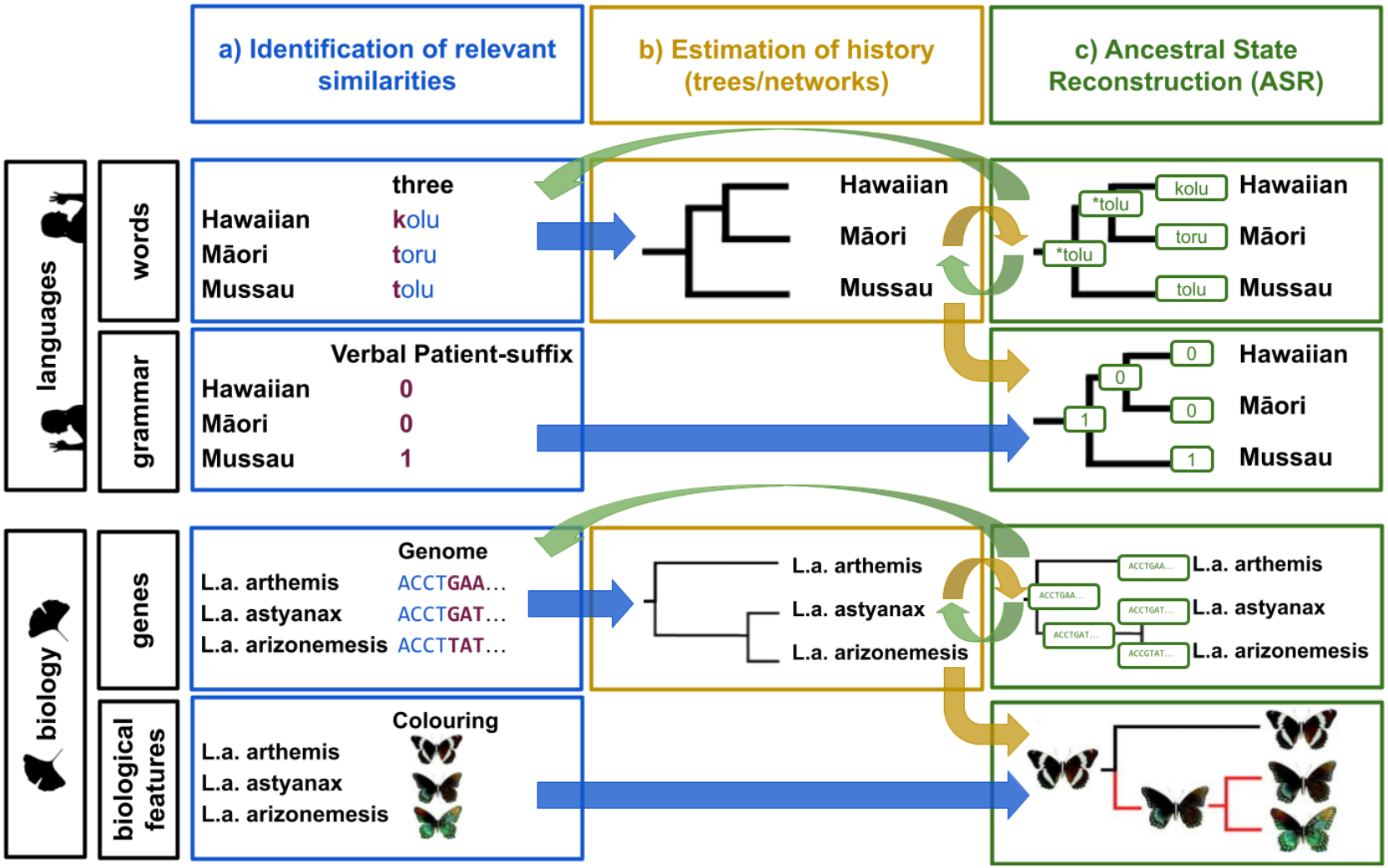


Figure 1: The three tasks involved in the historical reconstruction of linguistic matter (words & sounds), patterns (grammar) and biological traits. The tasks follow each other The butterfly illustrations are modified from **?**.

tent even over large amounts of data. Computational approaches are not intended to replace traditional historical linguistics, but rather to function as a complement — effectivizing parts of the process. In this paper, we compare the approaches conceptually and examine how often computational methods of ASR arrive at the same conclusions as traditional historical linguists. We will also investigate what the computational methods say when historical linguists disagree, and make new predictions about the grammar of proto-languages.

The practical example of this paper is the Oceanic language subgroup of the Austronesian family and the grammatical features of four of its proto-languages. We use information about the extant daughter languages from the Grambank dataset (**?**) to infer the structure of proto-languages given three trees: a) Glottolog 4.5 (**?**), b) **?** - MCCT and c) **?** - 100 random posterior trees aggregate (see more in section **??**).

Findings from the historical linguistics literature have been translated into data points in the Grambank format for four specific proto-languages: Proto-Oceanic, Proto-Central Pacific[2], Proto-Polynesian and Proto-Eastern Polynesian. The computational methods take as input the language-level data points in the Oceanic subgroups and then infer grammatical states of ancestral nodes in the trees (proto-languages). The structural features of the four proto-languages are extracted for each tree and method and compared to conclusions from classical historical linguistics.

The results are evaluated in terms of concordance between each method and the predictions from classical historical linguistics. We are evaluating how much they agree, not necessarily which one is correct. Which method is the most appropriate should be decided *a priori* based on the conceptual underpinnings and assumptions of the method and how plausible that model of change is (see more in sections **??**, **??** and **??**). Both traditional methods of ASR in historical linguistics and the particular computational approaches in this paper have advantages and disadvantages. Much of the conceptual infrastructure is similar, which is why we would assume a high degree of concurrence between the methods.

There is one area of Oceanic grammatical reconstruction where there is considerable disagreement. This concerns the nature of the alignment system of Proto-Polynesian and Proto-Central Pacific. These issues will be investigated and evaluated separately from the overall results of how much agreement there is between classical historical linguistics and computational approaches.

Finally, this study also yields predictions about grammatical features of the four proto-languages that were not addressed by the historical linguistics studies surveyed here.

[2]The concept of Central Pacific as a coherent subgroup is not uncontroversial. **?** and **?** made a case for a subgroup consisting of Fijian, Polynesian and Rotuman. Later **?** shows that the evidence for this stage is limited. We will be using it here in this study because it occurs enough frequently in the literature, but readers should be aware that it is less likely to have been a genuine coherent language of a community compared to the others. Thank you to Andrew Pawley for drawing this to my attention.

2 Background

2.1 The methods of ASR in traditional historical linguistics

This section lays out the fundamental principles of historical linguistics and how they relate to this paper.

The core method by which historical linguists reconstruct language history generally is known as the "Comparative Method". The Comparative Method is based on finding words or morphemes in different languages that have the same (or similar enough) meaning and that display non-trivial systematic phonological correspondences. By investigating these sets of words, it is possible to deduce which are inherited from a common ancestor, i.e. which ones are so-called "cognates". For example, **?**, **?** and many other scholars have analyzed Māori [maor1246] /toru/ (meaning 'three') as deriving from the same word as Hawai'ian [hawa1245] /kolu/ ('three'). These two words are "cognates" of each other and this information can be used to reconstruct a form for proto-Polynesian. Furthermore, many words that mean the same/similar thing in Māori and Hawai'ian show this pattern of t/k, e.g. Māori: /mate/ - Hawai'ian: /make/ 'to be dead' and Māori: /whitu/ - Hawai'ian: /hiku/ 'seven' (**?**). There is a systematic correspondence between these two sounds; regularly when there is a /t/ in Māori there is a /k/ in the corresponding position in Hawai'ian[3]. This is known as a *systematic sound correspondence*.

One crucial part of this approach is what **?** calls the "Double Cognacy Criterion" which states that both the part — the sound — and the context it occurs in — the word — need to be cognate in order to form valid data for ASR and sub-grouping in conventional historical linguistics. In the above example, the sounds /t/ and /k/ are cognates of each other, as are the words /toru/ and /kolu/.

The Double Cognacy Criterion is often more difficult to apply when reconstructing structural features, which has led some to say that reconstruction of grammar is impossible (c.f. **?**) and others to use a different approach which does without Double Cognacy, often labelled Syntactic Reconstruction (**?**: 17)[4]. This is relevant for the studies in historical linguistics (section **??**) that we will be comparing the computational results to. In this paper, we will compare methods of ASR in historical linguistics literature more broadly and not focus on only ASR in the specific sense of the traditional Comparative Method.

Besides finding cognates, traditional historical linguists also propose subgroups of languages as a way of modelling their history — typically in the form of trees. The estimation of historical relationships between languages in traditional historical linguistics is mainly/only focused on sub-grouping, which has the consequence that branch/edge lengths are typically not estimated at all. As one anonymous reviewer of this paper noted, branch length estimation is not a goal of the Comparative Method at all. This

[3]Further research into more Austronesian languages shows that Hawai'ian /k/ is more likely to be an innovation and Māori /t/ a retention from an older proto-language (c.f. in the Austronesian language Amis of Taiwan 'three' is /tulu/). Therefore, we can reconstruct that the change went from /t/ → /k/.

[4]Note that the term "Syntactic Reconstruction" is used for reconstruction of both morphology and syntax.

is important, because, as we will see, trees without branch lengths are most likely implausible and make for sub-optimal ASR. There is some work in traditional historical linguistics to establish branch-lengths through relative chronology of changes and archaeological anchor points (dating of clay tablets, texts etc, c.f. **?**), but that is mainly related to Indo-European. Note that estimating branch lengths need not be the same as precise chronological dating. For ASR, the key is the relative order of events rather than their precise timestamps (contrary to estimating urheimat, where dates matter more). It is also possible to incorporate uncertainty about splits - for example by positing many different trees with varying branches and splits that fit within a certain probability scope (e.g. a Bayesian posterior).

The processes of suggesting sub-grouping and ASR are done in tandem in historical linguistics; they are estimated simultaneously (c.f. **?**: 7). Subgroups are proposed based on **shared innovations**. In order to determine what is and what is not an innovation, a certain amount of reconstruction of proto-languages words and sounds (ASR) is necessary. In order to do ASR, some of the tree structure needs to be approximated. The tasks are done together.

This is different from analysis in biology and cultural evolution where ASR is typically carried out as a separate next step after a reliable model of history is constructed (c.f. **?** & **?**). In this paper, we are comparing specifically conventional historical approaches to ASR with computational approaches to ASR. The computational methods reviewed in this paper only does ASR, they do not suggest alternative sub-groups or trees. This makes them a bit different from classical historical linguistics approaches to ASR where sub-grouping and ASR are done simultaneously. Thankfully, this feedback loop between ASR and sub-grouping is primarily a factor in the classical historical linguistics analysis of linguistic matter (sounds and words) — and less relevant for linguistic patterns (structural features) which is the topic of this paper.

Historical linguistics has been primarily concerned with the reconstruction of sounds and words, but there is also work on the reconstruction of grammatical features such as morphemes or word order. **?**: 17-22 outlined three general principles for ASR in traditional historical linguistics that can be applied to structural data and vocabulary[5]:

(i) the number of changes posited (as few as possible, also known as Maximum Parsimony)

(ii) the plausibility of the changes posited

[5]It is also possible to include information on the history of particular regions and cultural factors when conducting linguistic ASR, but this is less often made explicit. In the case of Indo-European, which is the language family that has received the most attention so far in historical linguistics, there are even specific procedures involving particular parts of the tree. For example, **?**: 6 describes the "Anatolian Criterion" whereby "Proto-Indo-European ancestry can only be established for cognate sets that include an Anatolian language". This is because most historical linguists working on Indo-European see Anatolian languages as an early off-shoot branch of proto-Indo-European. In addition, work on the Indo-European language family can also make use of explicit information on ancient languages where we do have textual remains, such as Latin, Sanskrit etc. These kinds of procedures are currently not available for most language families in the world, historical linguists working on Indo-European represent a particularly privileged position informationally in relation to the field at large.

(iii) the plausibility of the reconstructed language as a human language (i.e. the degree to which the reconstructed traits work well in harmony with each other)

The first of these principles ("fewest amount of changes") is the same as what is known in the wider field of phylogenetics as "Maximum Parsimony" (**?**). The idea is to reconstruct states in proto-languages such that there are as few changes as possible between nodes in the entire tree. **?**: 17-22 explains how this works by positing an example of seven languages where there is a majority of one kind of value, X, and fewer of another, Y. Fig. **??** illustrates this example. If we only examine which feature is the most common, we should reconstruct X at the root of this tree (this is what **?** calls the "frequency heuristic"). However, this candidate solution would result in two changes (one each on the two paths from the root and to tips A and B respectively). If we instead reconstruct Y at the root, we would only need one change (between the root and PC-G). The solution where we reconstruct Y at the root results in fewer changes — it is the most parsimonious — and is therefore the preferred candidate.

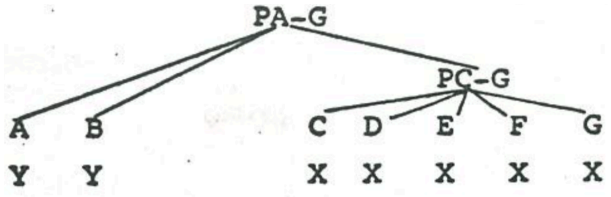


Figure 2: Tree from **?**: 19 illustrating Maximum Parsimony.

It is important to note that Maximum Parsimony does not take into account the length of branches, only the changes between each node of the tree (regardless of how far apart they are). It is of course possible that the true solution is *not* the one with the fewest changes. Maximum Likelihood (and other more recent ASR methods) improve upon this by taking branch lengths into account and actively estimating likely asymmetric rates (the rate of change X→Y can be different from Y→X) and actively applying the most likely rate to the inference of unknown ancestral states. ML attempts to find the most likely solution, MP can be said to find the solution with the slowest rate[6].

The next principle of ASR in historical linguistics concern the plausibility of changes — phonological, semantic or grammatical. For example, many historical linguists posit that /s/ is more likely to become /h/ than it is to become /k/[7] and this information is taken into account when doing ASR. Going to semantics next, in the earlier example from Māori and Hawai'ian, the words /toru/ and /kolu/ both mean 'three', but it

[6]It should be noted that my use here of "rate of change" in relation to Maximum Parsimony (changes per branch) is not directly comparable to rates of change estimated by other methods, such as ML. MP does not technically estimate a rate of change at all and does not model branches in a meaningful way, it is only concerned with changes between nodes. MP can however be said to *assume* the slowest rate of change, given the definition of the rate as "changes per branch".

[7]Historical linguists do concede that there are instances of irregular sound change (**??**) and that while they can often be explained by contact, analogy or avoidance of homophony, they sometimes remain unexplained.

is possible for cognates to have less similar meanings. For example, **?** reconstructs the proto-form **panua* as meaning 'land' or 'inhabited territory'. In various daughter languages, this has changed to 'place', 'community', 'village', 'house', 'people', 'world' and 'weather'. The meanings are related to each other, but not identical as in the 'three' - 'three' example earlier. Historical linguists aim to find plausible semantic connections between words that are proposed to stem from the same proto-form, they cannot be too dissimilar. The sound correspondences can be guiding here, if two forms have somewhat different meanings but convincing sound correspondences then they may still be cognates. This is difficult, as can be seen from this quote from **?**: 229: *there are no exact rules for handling semantic change; the final factor here is necessarily the common sense and the experience of the individual scholar.*

The plausibility of changes also comes into play when reconstructing structural traits. For example, a language going from having no marked dual number on nouns to having a trial number category would be taken as unusual by most linguists (c.f. **?**: 8) — it seems like the language has skipped over a necessary step, jumping from 'many' directly to 'three' without passing 'two'. Grammaticalization theories have given rise to a number of these plausible historical changes (**?**: 594-5, 598).

Lastly, ASR in historical linguistics deals with the plausibility of the whole of the reconstructed proto-language as a system (**?**: 1). For example, if we reconstruct a language with very uncommon combinations of features we should be wary and probably question the analysis. For example, it is rare to find a language that has a gender distinction in the first person, but not in the third (though not impossible; c.f. **?**). If the solution results in a proto-language with many rare features or unusual combinations of features, the work may require reconsideration. If something is rare in the languages that exist today, we would expect it to be relatively rare also in past languages. This is more relevant for phonology and linguistic structure where we have more worked-out theories of plausible combinations than in the lexicon. This principle has parallels in biology as well, where they avoid impossible ancestral states (c.f. (**?**)).

The procedure of ASR in historical linguistics that has been outlined in this section can and has been applied successfully to sounds and words (be they lexical or grammatical words). The application of this approach to abstract structural features is more controversial, and is not always included under a more strict application of the term "Comparative Method" *per se*. For this reason, we will refer to "traditional historical linguistics ASR" rather than the "Comparative Method" in relation to bodies of work that concern the reconstruction of the grammar of proto-languages (see section **??**). It is notable that most of the discussion regarding whether structural ASR is possible has been in relation to Indo-European languages (c.f. **?**). This paper offers a view from the Pacific Ocean, where the waves of the debate are more peaceful.

2.1.1 Disagreements in historical linguistics

As discussed, ASR in historical linguistics includes judgements of plausibility. This requires some assumptions about what are plausible features to co-occur in language, and which pathways of language change are more plausible than others.

Plausibility is important for ASR, both in linguistics, cultural studies and biology.

However, this principle is sensitive to differing assumptions and theories. What is more plausible as a reconstructed language, society or species may differ from scholar to scholar. Besides debates over precise sub-groupings, many arguments in historical linguistics boil down to disagreements about the plausibility of combinations of traits or of changes. This is also true of the different reconstructions of the alignment system of Proto-Polynesian.

? disagrees with **?**, **?**, and **?** on the case-marking systems of Proto-Polynesian on grounds of plausibility. Chung, Hale and Hohepa argue for a reconstruction that is technically less parsimonious on most trees of the languages (i.e. involves more changes), but which they say is more plausible. They posit that Proto-Polynesian had a nominative-accusative case marking system[8]. If this was the case, that would mean positing more changes along the tree than if we assumed, as **?** does, that the Proto-Polynesian language was ergative-absolutive. This is due to Sāmoan and Tongan both having ergative-absolutive marking and both splitting off early (in most accounts of the Polynesian tree) from Proto-Polynesian. Fig. **??** shows the Polynesian tree with Grambank feature GB409 values marked out[9].

I have summarised Chung's critique of Clark's proposal into three main points:

(a) the tree used is not an accurate representation of the language history (there was more interaction between Sāmoan and Tongan after splitting, and these interactions explains the situation)

(b) it is possible that the Proto-language contained variation and was undergoing change that was only fully realised in some of the daughters

(c) the morpho-syntactical historical process itself is less plausible

In a review of **?**, Chung writes:

> *Such an approach* [as Clark's] *relies on the assumption that the subgroups have developed quite independently once they split off from Proto-Polynesian, so that features shared by both must be attributed to the Proto-language. But in fact, both parts of this assumption are too strong. It is well known that the two primary subgroups of Polynesian did not develop totally separately; there was long-standing contact in pre-European times between speakers of Tongic and some Samoic-Outlier languages, as Clark himself notes (p. 27). Further, and more generally, it is simply not true that every feature shared by related languages must have existed in the Proto-language uniting them. Languages are constantly undergoing change; it is reasonable to suppose that Proto-languages were no different from real languages in this respect. But if this is so, then it is also reasonable that changes begun in a*

[8]Hale, Hohepa and Chung actually suggest three different theories which differ in specific details. For a summary of the differences between the proposals, see **?**: 247-249.

[9]Grambank feature GB409 asks if *any* ergative flagging is present. In some instances, the system is not wholly or primarily ergative, but ergative marking is present. It is possible that the scholars involved in the debate would not classify such languages as "ergative-absolute" languages *per se*.

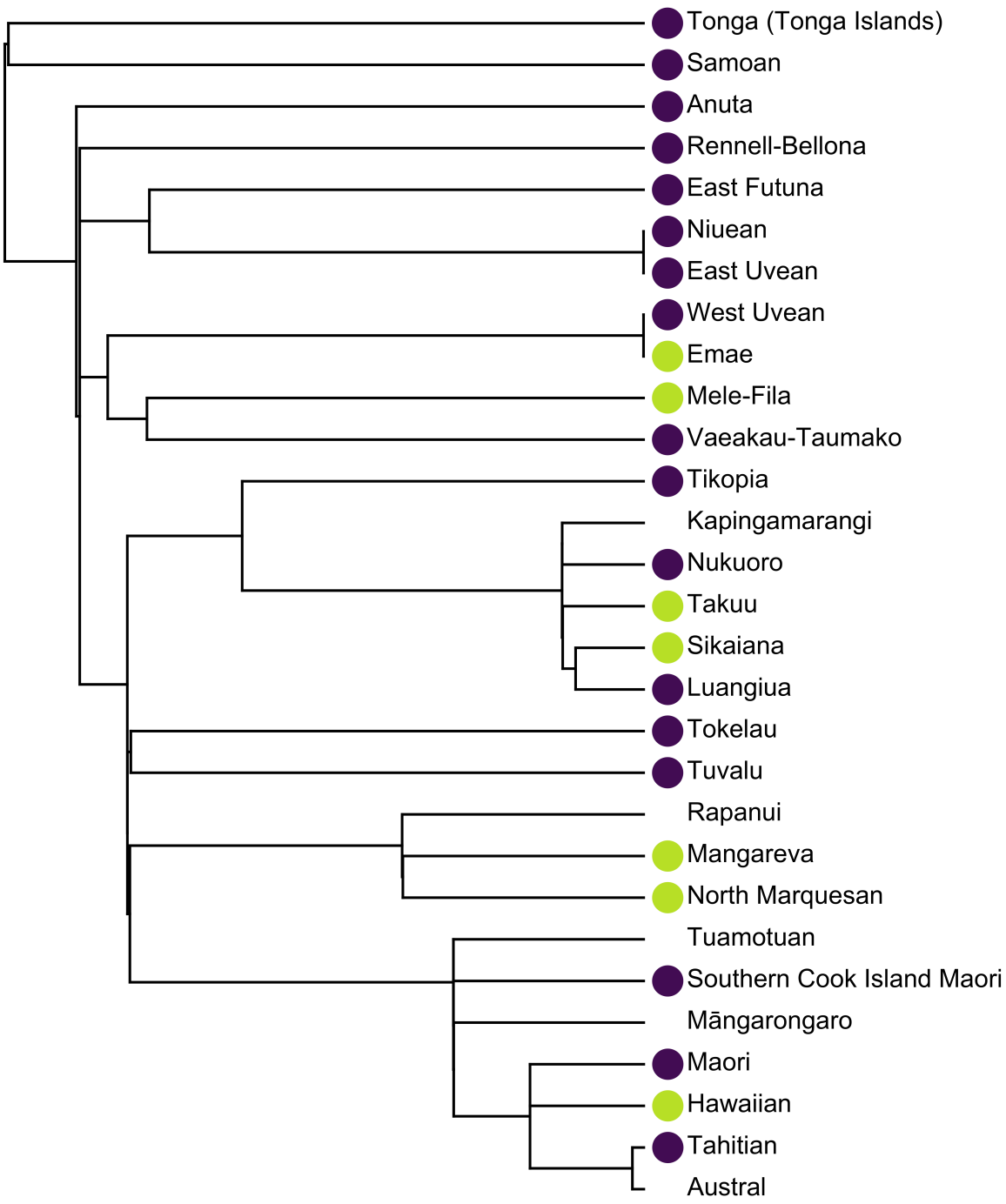


Figure 3: The Polynesian languages in the **?** Maximum Clade Credibility Tree-tree, with the coding of Grambank feature GB409 "Is there any ergative alignment of flagging?" marked out. Purple = Yes, Green = No and absence of dot = Not enough information/not clear.

Proto-language may have continued even after its separation into daughter languages. In this way, related languages may come to share a feature that existed only in embryonic form, or not at all, in their common ancestor.

?: 539

This debate contains more twists and turns, with each side arguing for the plausibility of their account. In our analysis, we will be using trees that represent the history of the languages in a similar way to Clark, which means the results are sensitive to the same critique by Chung (i.e. not taking into account contact between Sāmoan and Tongan). We are also not able to use plausibility in our computational reconstructions since we do not have access to formalised data on what plausible language profiles or changes are. This is a key difference between computational reconstruction and traditional approaches to reconstruction. Knowledge of plausibility and how to weigh different kinds of evidence against each other is not formalised and therefore cannot be taken into account.

In this study, any instances of conflicting data from historical linguists concerning proto-languages are evaluated separately from the overall results and will be reported in a separate section. There are three instances of this: two features related to the alignment of Proto-Polynesian (GB408 Is there any accusative alignment of flagging? and GB409: Is there any ergative alignment of flagging?) and one feature for Proto-Central Pacific, where **?** and **?** disagree on the alignment as well.

2.2 Evaluating if the data is valid for phylogenetic analysis: the Double Cognacy Criterion and phylogenetic signal

There is considerable debate within historical linguistics regarding if patterns (grammar) can indeed be analysed with the Comparative Method at all. One if the primary sources of disagreement is the criteria whereby similarities are judged to be valid for historical study. The Comparative Method is built on the recognition of the importance of cognates and sound correspondences — two concepts that are difficult to translate into the world of morphology and syntax (see for example **?** & **?**). In order to establish shared inheritance, two languages need to exhibit pairs of words where the words themselves can with great certainty be said to be related and where there is also a correspondence between the sounds *within* the words. This is what **?** calls the "Double Cognacy Condition". What does this mean for grammar? Are morphological patterns within sentences similar to sounds within words? The answer is not clear, and most likely varies depending on what kind of structural data we are dealing with (word order vs organisation of pronoun paradigms vs presence of certain markers etc). For this study, we will not delve too far into this debate but instead, use a quantitative test of phylogenetic signal to estimate if the data is suitable for ASR.

We take a leaf out of the books of cultural evolution and biology in terms of evaluating if the data is appropriate for phylogenetic analysis. In these fields, the three tasks outlined earlier are separated out; appropriate data for analysis is collected (this

can differ for tree/network construction and ASR) and trees[10] are constructed (usually with carefully chosen model approaches and priors). Once a reliable tree exists, ASR is carried out as a separate next step (c.f. **?** & **?**). There is a veritable smörgåsbord of methods that a scientist can choose to apply to each task. For example, you can use a plain distances-based approach to making a tree (**?**) or engage with more sophisticated Bayesian tools like BEAST (**?**). Similarly, for ASR there are different approaches with different pros and cons (see **?** for an overview).

The input data for ASR can differ from what was originally underlying the construction of the tree. If we believe that the tree is likely to be a good estimation of the history also for other data besides what it was directly based on, we may be able to carry out analysis with that tree on different data. For example, **?** analyzes evolutionary dynamics of societal variables such as ritual human sacrifice and social stratification in the Pacific using trees that are based on basic vocabulary and archaeological priors of island settlement (**?**). Besides arguing that it is reasonable to assume that the history of a community's sociopolitical past is similar to its linguistic past, we can also test the strength of the phylogenetic signal statistically and use this as a guide. If the data has a reasonable phylogenetic signal, we assume that it is likely that it was generated by the tree and that we can proceed with further analysis. This is what **?** did for their data, and they found that it was possible to carry out the analysis. We can do this as well for our Oceanic trees and structural data.

Phylogenetic signal is the degree to which it can be assumed that a particular tree is likely to have given rise to the data in question, the tendency of related tips to resemble each other more for a particular variable than they would if randomly rearranged[11](**?**: 905). There exist several different tests of phylogenetic signal, among others: Pagel's λ (**?**), Bloomberg's K (**?**), Borges et al's δ (**?**) and Ives & Garland's alpha (**?**). In this study we will use a common and conceptually simple measure: **?**'s *D*-estimation. This metric has been used in language studies on sounds of languages (**?**) and grammatical features (**?**) and is relatively straightforward. The D-algorithm takes a tree and a binary trait (in this case structural linguistic features) and simulates what the distribution of values would be if the data was a) generated by Brownian evolution or b) randomly generated. Both scenarios are simulated with the same prevalence of tip states as the real data. The algorithm produces a D-estimate for each trait and tree, which represents the similarity to these two scenarios. If this value is close to 1, the data is similar to what would happen if the data was randomly generated (if the D-metric is higher than 1, this represents it being over-dispersed[12]) and if it is near 0 then it is more similar to Brownian evolution. The algorithm also produces kinds of p-values which show how likely it is that the data is dissimilar from 0 (Brownian) and 1 (random).

[10]History of organisms and culture can be understood as trees, waves and networks. For the sake of space, we will write "tree" since this is most common but waves and networks are not excluded *a priori*.

[11]Nota bene: this is *not* the same as stability/conservatism, phylogenetic signal is a separate concept.

[12]Over-dispersed here means that the trait is spread out over the tips in a way that shows no clusters, even less clustered than one might expect by chance. For example, sister-pairs may have opposite values to each other all throughout the tree. See table 1 in (**?**: 1044) for illustrative figures.

In this study, we are primarily concerned with 84 unique Grambank features [13] and three trees (Glottolog 4.5 (**?**), **?**-MCCT and an aggregate of 100 random trees in the **?** posterior). We carried out the D-estimate analysis on all of these features over all trees using the function phylo.d in the R package caper (**?**). The results are summarised in table **??**. The second column of the table shows the mean D-estimate value over all 84 relevant Grambank features for each phylogeny. The third column shows the percentage of features with p-values that indicate that they are Brownian or clumped (pval0 > 0.05). The fourth column shows the number of features for which it was not possible to carry out the D-estimate calculation because they do not meet the rigours of the model. In all of these cases, it was because there was a very skewed distribution of values over the tips (e.g. 4 tips with "absence" and 118 with "presence"[14]) and this is not suitable for the analysis (for more technical details see Supplementary Material **??**). We can consider all of these to be a kind of "super-conservative" (the feature rarely evolves), but we cannot derive a measurement of phylogenetic signal *per se*. Lastly, some features were excluded because there was too much missing data which caused the pruned trees to have too few tips for analysis.

tree	D-estimate (mean)	Proportion of features not significantly dissimilar to 0	features unfit for D-estimate	Too few tips altogether
Glottolog	0.34	47%	8	0
Gray - MCCT	0.28	58%	17	1
Gray - posteriors	-0.01	81%	22	1

Table 1: Table showing D-estimate (phylogenetic signal) of Grambank features that map onto research in traditional historical linguistics (n = 84). Posterios values are mean values over all 100 trees and features. Data unfit for D-estimates excluded.

Most features under study have a D-estimate similar to 0, meaning that they have phylogenetic signal. There is however many features that are not similar to a D-estimate of 0. The results and conclusions further discuss the relationship between this principle and agreement with conventional HL.

2.3 Computational phylogenetic methods

In recent years, linguists have begun to apply computational phylogenetic methods from biology to the reconstruction of linguistic history. Biologists have, similarly to

[13] Sometimes we are interested in the same feature for more than one proto-language. The total amount of data points we are interested in for comparison to conventional historical linguistics is 115 over four proto-languages, which reduces to 84 specific unique features.

[14] It does not matter here whether it is the presence or absence of a trait that is rare, this has no effect on the measurement of phylogenetic signal.

linguists, been interested in inferring trees of the genetic relationship between species[15], ancestral states and the tempo and mode of evolution (**?**). Biologists and linguists may have inspired each other, but methodologically the fields progressed separately for a long time (**?**: 370). Both fields are interested in answering similar questions: how are these languages/species related?, what was the earlier state of a language/species?, which traits are changing slower/faster? etc. The two fields have developed different methodologies, with biologists leaning more towards quantitative computational methods for tree construction and ASR compared to linguists who have focused more on rigorous tests for which linguistic data is valid for analysis (see for example the Double Cognacy Criteria in **?**). It is possible that this is due to the material under study, it may be more difficult to tease our non-trivial similarities in languages than in species (especially since genome sequencing became more readily available).

Applying computational methods of ASR to linguistic data is becoming more common. **?** apply three different methods (Maximum Parsimony, Maximum Likelihood and Minimal Lateral Networks) to cognate class reconstruction in three different language families. The aim of that study was primarily to evaluate how often the methods reconstructed the same state as what the authors label "the Gold Standard" (reconstructions by traditional historical linguists using the classical Comparative Method). This is similar to the study at hand; one of the aims of this paper is to estimate the degree of concurrence between computational methods using typological database data and conventional approaches in historical linguistics. The data that serves as input to the computational machinery in **?** was annotated by "hand" for cognacy by historical linguists, meaning that the identification of cognate classes is still an entirely human affair (task one in Fig. **??**). This is also true for this study, the identification of structural features in languages is a human process. The overall result of **?** was that Maximum Likelihood performed the most similar to traditional historical linguistics ASR, but that there were still several shortcomings. Most notable of these were undetected borrowings, variation within languages and parallel independent shifts. In this paper, we address the potential for contact events by using sets of trees from a Bayesian posterior (as **?** also do), some of which may represent an alternative contact history (see section **??**).

There are also two recent studies of Indo-European grammatical history: **?** and **?**. **?** evaluate different theories of the history of the morphosyntax of Indo-European by comparing these to the product of computational Bayesian phylogenetic modelling. They find support for the "canonical" model of Indo-European syntax. Goldstein in his paper challenges a commonly applied principle in the reconstruction of Indo-European syntax; the "frequency heuristic" which holds that *if the number of homologous elements (e.g., lexical cognates) in the daughter languages meets a minimum threshold (canonically three), their ancestor is reconstructed to the root of the tree* (**?**: 1/71). This is done because scholars argue that the true tree is unknown and that this is an appropriate method in the absence of the true tree. Goldstein argues that the

[15]Interestingly, the use of trees in linguistics and biology first occurred in publications just one year apart with **?** publishing a tree of languages and **?** a tree of species. However, as **?**: 370 notes, it was not until Darwin's publication of *The Origin of Species* in **?** that the concept of species trees in biology truly took off.

appropriate action is instead to carry out reconstruction on many different trees that represent possible histories — a Bayesian posterior tree sample. He argues that this is methodologically more sound and because the results of his approach are in accord with the consensus in historical linguistics it strengthens their validity.

Both **?** and **?** use a Bayesian method of ASR within the Continuous-Time Markov Chain (CTMC)-framework[16]. This approach comes with certain important assumptions, to quote from **?**: 77:

> *CTMCs model language change as a stochastic phenomenon with rate parameters that govern the amount of time between transition events. It is worth highlighting the assumptions that these models bring with them. First, character states at the nodes of a tree are assumed to depend only on the state of their immediate ancestors and the length of the branch along which they evolved (Cathcart 2018:4). Second, the probability of a transition depends only on the current state of a language. Its previous history is irrelevant. This is known as the markov property. Finally, rates of gain and loss are assumed not to vary across the tree.*

It is always important to be up-front and explicit about the assumptions an approach takes and evaluate if they make sense for the given situation. For linguistic data, these assumptions do seem to hold. For more details on the methods, please see **?**, **?**, **?** and **?**.

Another popular method of ancestral state reconstruction is Stochastic Character Mapping (SCM, **?**). SCM is a procedure that simulates character histories using Continuous Time Markov-rates. These rates are usually estimated on the basis of the tree topology and the data attested at tree tips before SCM is carried out, but can also be defined in other ways. SCM can follow the same CTMC approach employed by **?** and **?**, but not necessarily[17].

Computational approaches to reconstruction not only allow us to effectivize the process by inferring the prior states of hundreds of traits in a short span of time, but they also allow us to apply exactly the same principles in exactly the same way to all pieces of data. This is much harder to do manually, since different scholars may use slightly different assumptions and judgement when conducting ASR. One could say that what we lose in deep human insight, we gain in consistency and speed. Furthermore, if the deep human insight of historical linguistics could be quantified into priors that can be fed into computational models — we may not need to lose anything. Unfortunately, this is not the case currently, but it may be possible in the future.

[16]The main difference between the methods of **?** and **?** is that **?** use a tree structure informed by **?** and comparative-historical *communis opinio* and vary the branch lengths 10,000 times in a principled and informed manner to generate 10,000 different trees while **?** takes 100 random samples directly from the posterior of **?**.

[17]Thank you to one of the anonymous reviewers of Diachronica for highlighting this.

3 Material and methods

3.1 Methods: Maximum Parsimony, Maximum Likelihood and Most Common

In this study, we will be reconstructing the presence or absence of structural features in proto-languages of the Oceanic subgroup using Maximum Parsimony, Maximum Likelihood and Most Common. This section gives a brief overview of the three methods. Further technical details concerning the precise application can be found in the supplementary material **??**. For an extensive comparison of different methods of ancestral state reconstruction and their advantages, see **?**.

Maximum Parsimony finds the set of ancestral states that result in the fewest number of changes between nodes (also known as "lowest Parsimony cost"). If we think of the rate of change as the number of changes in the tree, then Maximum Parsimony selects the candidate solution with the slowest rate out of all possible solutions it can choose from. Maximum Parsimony is intuitively simple. While the principle of "Maximum Parsimony" is practised in traditional ASR in historical linguistics, it should be noted that they rarely use the term *per se*, but rather the description of "fewest number of changes along the tree".

Maximum Parsimony may be simple and intuitive, but it is not without its critics. Part of the critique is that it does not take into account branch lengths in the tree (the time between splitting events). Furthermore, Maximum Parsimony necessarily assumes that the solution that posits the fewest changes (slowest possible rate of change) is also the most probable one. This is not necessarily a valid assumption; some features may evolve at a faster rate than Maximum Parsimony predicts. Both of these disadvantages are addressed in the second method we will be applying: Maximum Likelihood.

Ancestral state reconstruction using **Maximum Likelihood** posits the most likely ancestral state distributions based on the overall probabilities given all the nodes in the tree and all branches. This approach does not assume that the slowest rate of change is the most probable one. If, for example, the distribution of values at the tips is very scattered, with sibling pairs frequently having different values, Maximum Likelihood will infer that the feature has a high rate of change and use that information when positing ancestral states as well. The Maximum Likelihood algorithm assigns probabilities of state changes and distributions based on branch lengths. A mutation along a shorter branch is given more weight in the likelihood calculations than if it occurred in a longer branch. Furthermore, reconstruction using Maximum Likelihood allows us to use a model of change where we do not assume that the rates for losses (1→0) are equal to the rate of gains (0→1). In this study, we use an "All Rates are Different" (ARD) model, which allows for the rate of loss and gain to be different[18]. It is possible to further specify the model, for example by specifying transition rates, specify certain nodes beforehand etc. For this study, this was not done since there is no information to base these decisions on. Specifically, we are also using a marginal Maximum Likelihood estimation, for more details see supplementary material **??**.

While Maximum Parsimony and conventional HL-ASR are similar, they have dif-

[18]Similarly to the studies by **?** and **?**, rates cannot vary within the tree in this study.

ferences as well. It is impossible for Maximum Parsimony to take into account branch lengths, nor can it assume anything but the slowest rate of change or posit different rates for losses and gains. It is, however, possible that historical linguists estimate something similar when they take into account the "plausibility of the changes posited". It is possible that scholars of historical linguistics take the length of time into account or assume that a loss is more likely than a gain for a given feature. In this study, we compare Maximum Parsimony and Maximum Likelihood reconstructions to the conventional ASR in historical linguistics. If the results from conventional historical linguistics are more similar to that of Maximum Likelihood, a potential explanation would be that the "plausibility of changes posited" is indeed operating along similar lines as Maximum Likelihood by taking branch length into account and assuming varying rates of change.

We will also compare the predictions of historical linguists with a "dummy-model" which is based solely on which value is the **Most Common** in the daughter languages of a given proto-language — entirely disregarding the tree structure[19]. In the toy example in Fig. **??**, this approach would reconstruct that the root had feature value "X". Whether we prefer Maximum Parsimony, Maximum Likelihood or another approach to reconstruction, it should be the case that actually taking the tree structure into account is the sounder methodology.

All analysis have been calculated in `R` (**?**) using the packages `castor` (Parsimony, **?**), `phangorn` (Parsimony, **?**) and `corHMM` (Maximum Likelihood, **?**). The packages `ape` (**?**), `adephylo` (**?**), phytools (**?**), psych (**?**), reshape2 (**?**) and tidyverse (**?**) were also used for data wrangling, analysis, summarising and visualising[20].

All of the R-code and data necessary for the analysis in this paper is published alongside the paper, in Supplementary Materials and in archived web-storage (Zenodo).

https://doi.org/10.5281/zenodo.8056616

https://github.com/HedvigS/Oceanic_computational_ASR/releases/tag/v0.1

3.2 Calculation of similarity between predictions from conventional HL and computational approaches

We calculate the similarity of the predictions of historical linguists and computational methods with a measure of concordance[21]. Concordance measures how closely the computational reconstruction matches historical linguists' reconstruction — how much they concur. It is measured as the number of agreements about grammatical features (i.e. Grambank binary questions) of predicted proto-languages, divided by the total number of grammatical features predicted.

For each feature, the methods predict a distribution of the two states (presence and absence) for every ancestral node. If the distribution is majority presence (more than

[19] This is similar to the frequency heuristic described in **?**.

[20] For a complete record of all R-packages used, see the supplementary material **??**.

[21] This metric is also known as *accuracy* in machine learning, but we do not use that term because we wish to avoid the connotation that what is being measured is the real-world accuracy of the reconstruction as opposed to the agreement between methods.

60% of the ancestral state is "1") it is registered in the results as "Presence"; if less than 40% presence it is registered as "Absence". If the ancestral state is between 40-60% of either state, the prediction is registered as "Half/Half". This was done to highlight the amount of uncertainty the results sometimes contain, while at the same time making it a fair comparison between Maximum Parsimony and Maximum Likelihood. Comparing the raw distributions themselves is not a fair comparison because Maximum Parsimony is always more likely to suggest 0, 0.5 or 1 results (because the majority of the splits in the tree are binary) whereas Maximum Likelihood rarely produces exactly 0 or 1. Rounding into these bins also makes it possible to derive the number of "True Negatives", "True Positives" etc which allows for the calculation of concurrence scores.

If the reconstruction of a feature by HL experts for an ancestral node was "Presence" and the algorithm did predict presence with over 60%, it is counted as a "True Positive", and so on[22]. Table **??** illustrates how the results are summarised.

Table 2: Table illustrating how the results of ancestral node predictions are calculated.

Finding in historical linguistics	**Prediction by Computational Method**	**Result**
Absence	>60% Absence	True Negative
Absence	>60% Presence	False Positive (type 1-error)
Presence	>60% Presence	True Positive
Presence	>60% Absence	False Negative (type 2-error)
Absence	40-60% Presence/Absence	Half
Presence	40-60% Presence/Absence	Half

For each method, a plain concordance score (Eq (**??**)) is then calculated. The score is calculated between all computational methods and the conventional historical linguists' prediction and also between all computational methods themselves [23].

$$\frac{\text{True Negative} + \text{True Positive}}{\text{True Negative} + \text{True Positive} + \text{False Negative} + \text{False Positive}} \tag{1}$$

It is also important to take into account the Half-results. This count represents instances where the method was not able to say with strong confidence that something was present or absent. The reason it is interesting to separate these out is that while they may indicate a majority result in one direction, it is not far from suggesting the direct opposite. For example, if one of the methods reconstructs Proto-Oceanic as having a 51% chance of having ergative marking — it is not far away from suggesting that this marking is absent. In order to take these types of cases into account the cut-off

[22]The terms "True" and "False" are used here in accordance with terminology in machine learning. In this instance, they are indicating whether the results from the computational method and historical linguists agree (True) or not (False). It should not be interpreted as a measure of empirical "Truth" necessarily.

[23]When comparing one computational method result to another, "half" - "half" count as a True pair. Otherwise the scoring is the same.

of 40%-60% was set and summarised as "Half" results. We can apply the concordance score to this summary statistic as well, as shown in Eq. (**??**).

$$\frac{\text{True Negative} + \text{True Positive} + \frac{\text{Half}}{2}}{\text{True Negative} + \text{True Positive} + \text{False Negative} + \text{False Positive} + \text{Half}} \tag{2}$$

Both scores will be reported, but we will rely mainly on the concordance score with the inclusion of the Half-results. This is because this approach takes into account the possible uncertainty of the half-scores which can be valuable information.

In a similar study of ancestral states of cognate classes, **?** compared three different methods of ancestral state reconstruction for lexical data (cognate classes): Maximum Parsimony, Maximum Likelihood and Minimal Lateral Networks. They found that reconstructions using Maximum Likelihood performed the most like the predictions by historical linguists. However, **?** describe the general performance of all the computational reconstruction methods they used as "poor". **?** evaluated the methods using F1-scores which is the harmonic mean of Precision and Recall (**?**). This way of evaluating performance focused on True Positives and ignores True Negatives altogether. It was suitable for the study by **?** because they were primarily interested in the presence of cognate classes, which makes disregarding True Negatives admissible. This is not the case here, True Negatives for structural features are meaningful in a different way than the absence of cognate classes. Because of this, we will not report F1-scores in the main text but only in the supplementary material (see Supplementary Material **??** and **??** [24]).

In addition, we also tested the strength of correlations between agreement with HL and measurements of phylogenetic signal on one hand and the distribution of tips in each state on the other. For this comparison, we are comparing each method separately against the HL-agreement and phylogenetic signal/distribution of states. For this reason, we did not use the above approach of binning results into Presence, Half or Absence but used the predicted values from the method directly instead (see Supplementary Materials **??** and **??**.

3.3 Data

3.3.1 The Grambank-dataset

The data for the study is taken from the Grambank-project (**?**). The Grambank dataset consists of 195 structural features which have been coded by a large group of research assistants for over 2,400 languages. This dataset includes 280 Oceanic languages.

The questionnaire's 195 questions cover what are often called the "core domains" of traditional grammatical description: word order, possession, negation, tense, aspect, mood, deixis, interrogation, comparatives and more. Features are included in the questionnaire if it is likely that it is possible to code them for the majority of the world's languages which have been described grammatically (approx 4,000 languages,

[24] I am very grateful for mathematical assistance from Stephen Mann in regards to the F1-score including half results calculation.

see **?**). This means that rarer features are not included such as family or region-specific ones. The full questionnaire is found in appendix **??**.

The Grambank dataset is coded by student, research assistants and other collaborators under the supervision of expert linguists. Each feature is accompanied by documentation guiding coders so that the questionnaire is applied as consistently as possible across different languages. For more details on the coding workflow of Grambank, see **?**.

There are differences between how grammatical structures are described in the historical linguistics literature and how they are defined in Grambank, for more on this see section **??**.

3.3.2 Data coverage

This study is focused on the Oceanic subgroup of the Austronesian language family. The Oceanic subgroup covers almost all languages in Remote Oceania (with the exceptions of Chamorro and Palauan) and large parts of Near Oceania. Fig. **??** from **?**: 2 shows the extent of the major subgroups of the Austronesian language family, with Oceanic covering the largest surface area. Following the language classification of Glottolog 4.5 (**?**), there are 522 languages in total in the Oceanic subgroup.

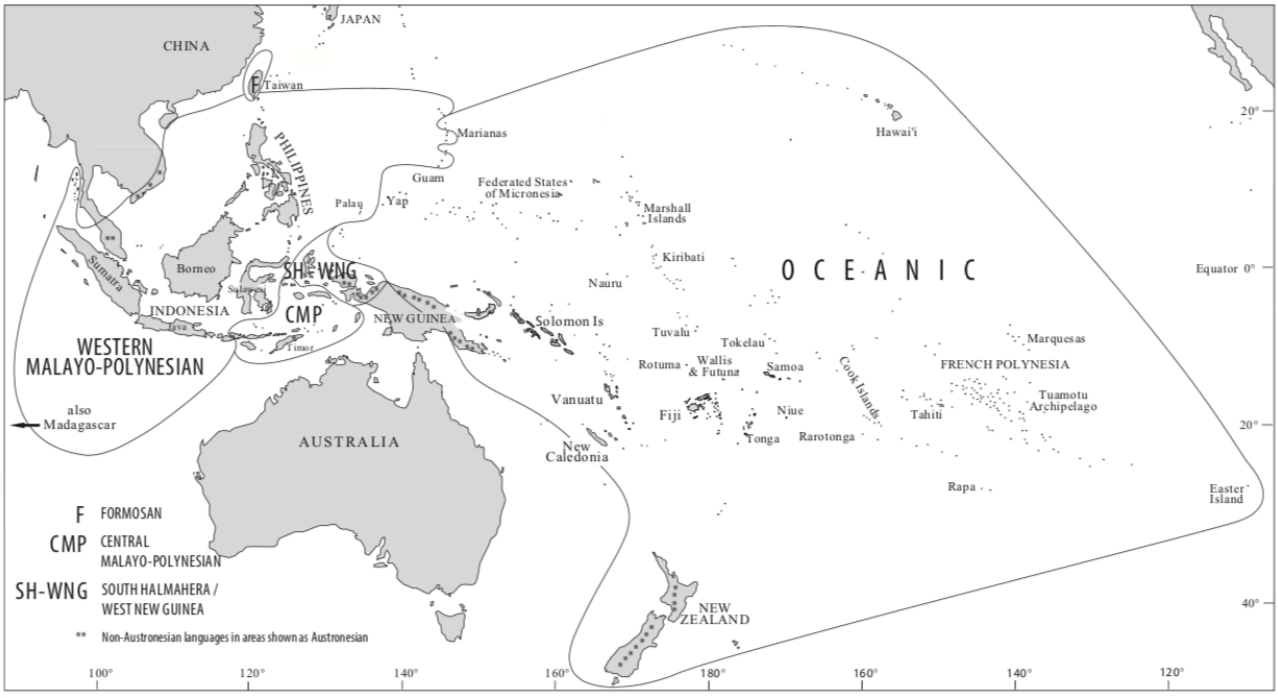


Figure 4: Map of the Austronesian language family and major subgroups, from **?**: 2.

Not all languages of the Oceanic subgroup have a grammatical description, but out of those that have it, nearly all are included in Grambank. Table **??** shows the coverage of Oceanic languages in the entire dataset. According to Glottolog, there are 289 Oceanic languages that have a grammar or a grammar sketch. Out of these 180

are included in grambank. The map in Fig. **??** shows the same coverage information, with languages coded for their data coverage status.

Island group	**More than half of the features covered in Grambank**	**Less than half of the features covered in Grambank**	**Grammar exists, but language not in Grambank (yet)**	**No grammar**
Bismarck	42	7	0	5
Central Pacific	33	1	1	10
Central Vanuatu	48	1	0	42
Interior New Guinea	4	0	0	11
Micronesia	16	1	0	6
N Coast New Guinea	19	3	2	76
New Caledonia	14	0	3	16
Northern Vanuatu	5	0	0	9
S New Guinea	26	1	4	35
Solomons and Bougainville	30	4	1	25
Southern Vanuatu	8	0	0	1
Temotu	5	2	0	3
Total	250	20	11	239

Table 3: Table showing coverage of Oceanic languages in Grambank per island group.

The coverage of Grambank data for the Oceanic subgroup is in general better in the east than in the west. However, since we control for genealogical relatedness in our ASR with trees directly, this is less of a problem for our methodology than if we were using traditional probability sampling (c.f. **?**).

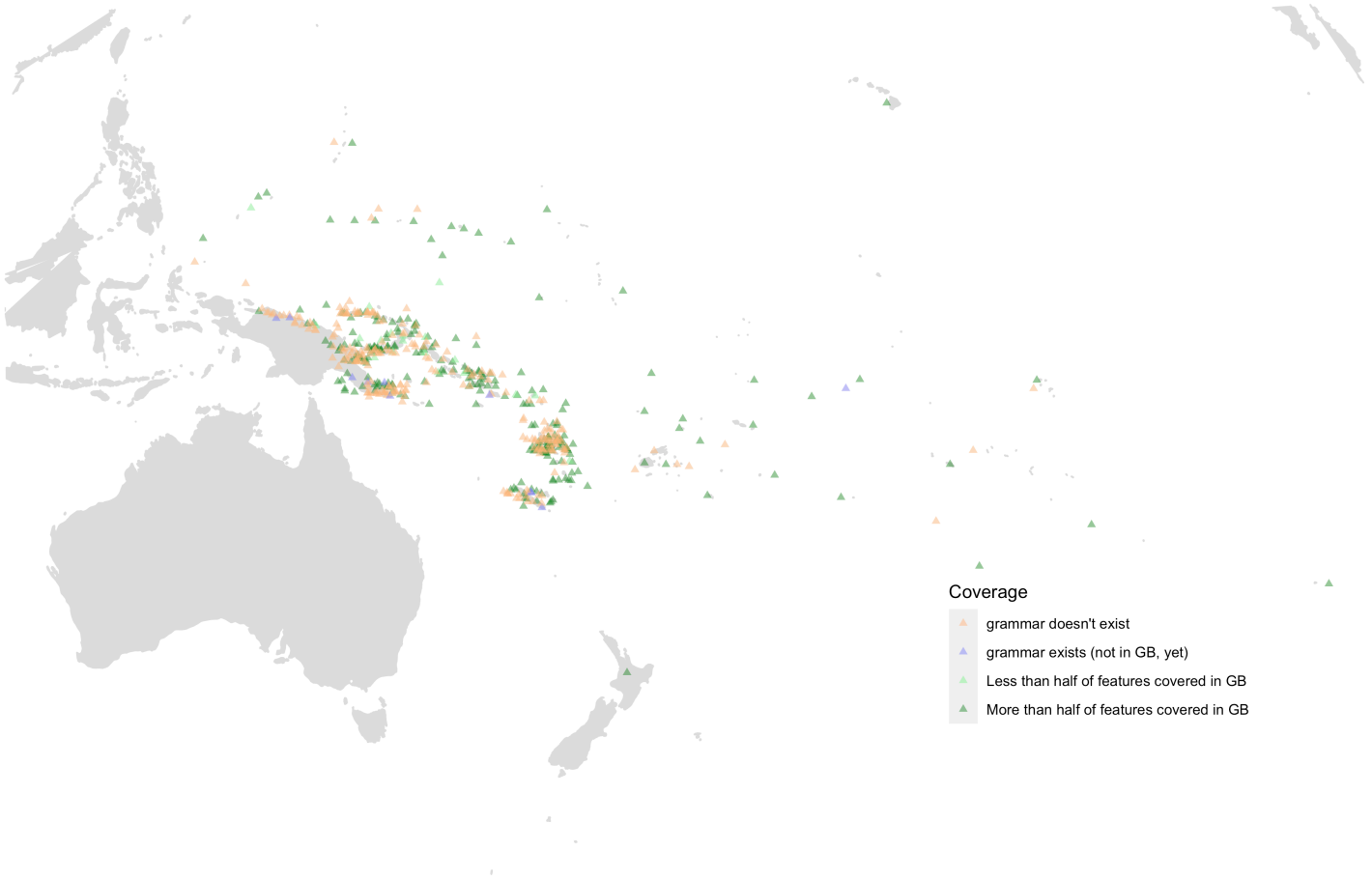


Figure 5: Map of Oceania, with Oceanic languages coloured for their coverage in Grambank.

3.3.3 The trees

The tree phylogenies used in this study are:

(a) the Maximum Clade Credibility Tree (MCCT) from **?**

(b) a random sample of 100 posterior trees from **?**

(c) the tree from Glottolog 4.5[25]

Figure **??** and Figure **??** show the Grambank coverage of languages over the phylogenies from the Gray et al 2009-MCC-tree and the Glottolog-tree respectively.

One of the major differences between the trees is that the Glottolog-tree does not contain *any* information on branch lengths. All the branches in the Glottolog-tree are of the same length (1), whereas the branches in the Gray et al 2009-trees (the MCCT and the posteriors) have meaningful branch lengths based on rates of change in the underlying data (basic vocabulary) and calibration points (archaeological dates). This can be seen in the visualisations in Figures **??** and **??** where the first has varying lengths of branches but the latter all have a uniform length (1). This has the consequence that some tips in the Glottolog-tree are *much further* from the root than others. This is a big disadvantage with this type of tree, since it suggests that different amounts of time has passed between the root and the languages at the tips.

In addition, the Glottolog trees contains more non-binary splits (polytomies) than the **?**-trees. Binary splits ought to be more plausible, since it is unlikely that a set of 3 or more languages are all *exactly* equally related to each other. Polytomies can be a way of signalling uncertainty, when it is not clear how to structure the group it may be preferably to suggest a polytomy than a less certain binary branching. In the Glottolog Oceanic tree (pruned for matches to Grambank), 10% of splits are not binary. In the **?** MCC-tree, only 3% are non-binary. Taking into account samples from the posterior is another way of accounting for uncertainty without needing polytomies as much. In the random sample of 100 trees from the **?** posterior 39 trees had binary splits, the mean percentage of non-binary splits across all 100 is 0.15%. Further technical details of the trees can be found in Supplementary Material **??**.

The Glottolog-tree contains all the languages in the Oceanic subgroup. Therefore the coverage per island group that is summarised in table **??** in the previous section applies to the Glottolog-tree as well. However, the **?**-trees do not contain all Oceanic languages, but rather 155. Out of these, 132 also occur in the Grambank dataset.

Finally, we are also using a sample of the posterior trees from **?**. Their study yielded 4,200 posterior trees. Tree topologies that are more probable occur more often. By using a set of possible trees instead of just one we may be able to include diverging historical accounts, which could for example estimate contact events as well as inheritance. Figure **??** shows a DensiTree-visualisation (**?**) of the 100 trees which are used in this study.

[25]The tree of Glottolog 4.5 (**?**) is based on work by **??** and **?**.

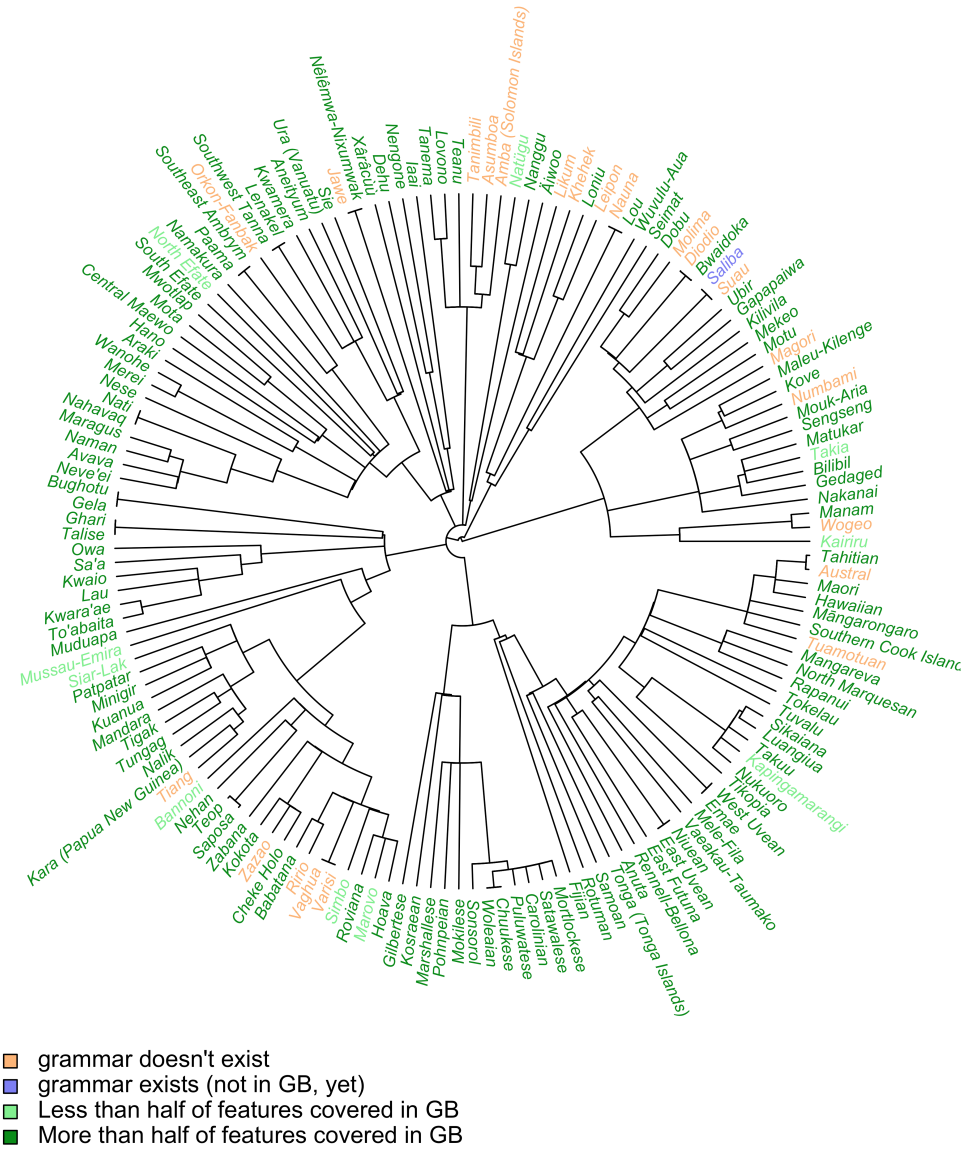


Figure 6: Maximum Clade Credibility Tree of Oceanic from **?**, with languages coloured for coverage in Grambank.

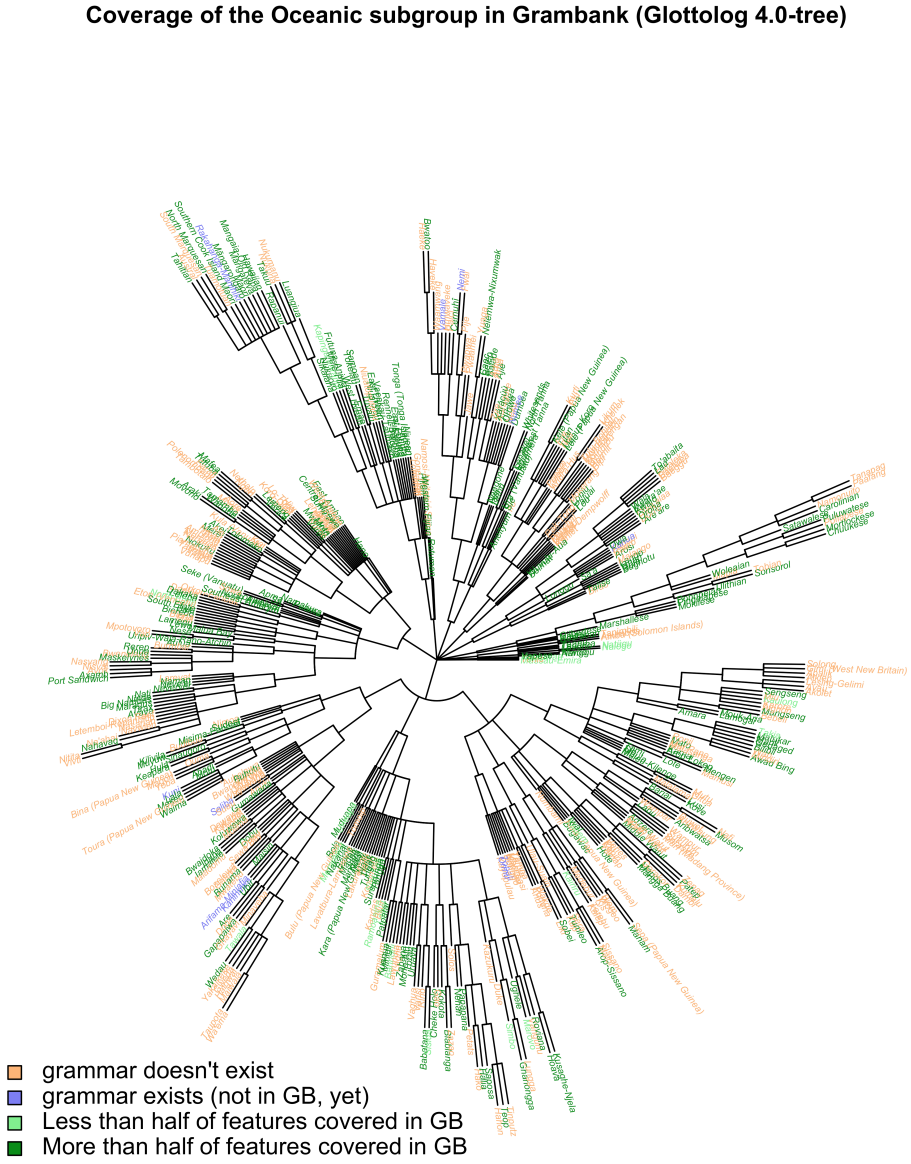


Figure 7: **Tree of Oceanic from Glottolog, with languages coloured for coverage in Grambank.**

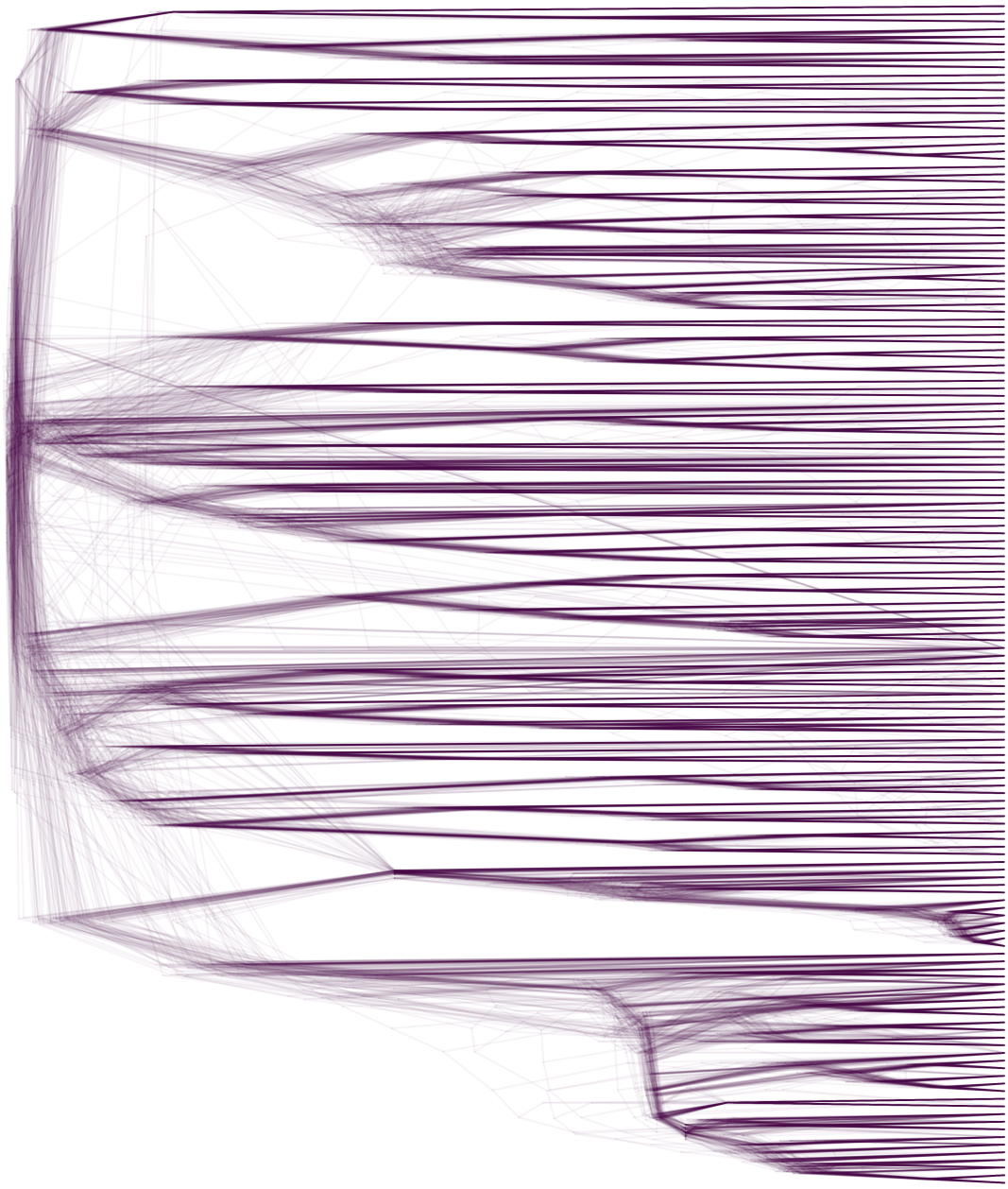


Figure 8: **DensiTree (?) visualisation of the 100 random sampled trees from the Gray et al 2009-posterior.** Made with the function densiTree() from the R-package Phangorn (**?**).

3.3.4 Data from historical linguistics on Oceanic proto-language grammar

Oceanic proto-languages are well-researched in terms of their lexicon and phonology compared to most languages in the world (see the book series on the Proto-Oceanic lexicon (**?????**), among other publications). There exists also substantial work done on the grammar of Proto-Oceanic using conventional methods in historical linguistics. We have summarised several major works in the field and distilled their research into predictions about Grambank variables in proto-languages. This section gives an overview of the works included and examples of how they have been incorporated into the study. Table **??** in the supplementary material **??** lists all of the publications used here as representations for the reconstruction of grammar in Oceanic linguists by conventional historical linguistics means.

For each of these publications on the grammar of proto-languages, findings have been extracted that support a certain coding in the Grambank Questionnaire for a given proto-language. For example, **?**: 4 writes that a causative prefix can be reconstructed for Proto-Polynesian (**faka-*). In the Grambank questionnaire, we have the feature GB155 '*Are causatives formed by affixes or clitics on verbs?*'. For Proto-Polynesian and GB155 the predicted state from HL is"1" (yes/presence). For simplicity, we are only considering four ancestral languages: Proto-Oceanic, Proto-Central Pacific, Proto-Polynesian and Proto-Eastern Polynesian. The choice to focus on these four, in particular, was based on the fact that they are the most well-researched proto-languages in the literature in terms of grammatical features that can be coded for in Grambank.

As evident by the example in the previous paragraph, the work on ASR of grammar in Oceanic languages typically concerns specific forms (e.g. **faka-*) while the Grambank questionnaire targets more abstract features. This means that the Grambank coding of the proto-languages based on conventional historical linguistics ASR is not a *precise* rendition of the literature, but a typological interpretation of the historical research. This task is the same as the coding of the extant daughter languages (Tikopia, Paluan etc), we read grammars that describe particular forms, paradigms etc and then translate that into Grambank-datapoints.

When doing ASR in conventional historical linguistics, scholars in this field also take into account fossilised forms (e.g. the common noun marker *-a* fusing to roots in Paamese (**?**: 141)) and related meanings (e.g. the hypothesis of *-Cia* changing from a transitivising suffix to a marker of passive voice (**????** and **?**). The Grambank dataset, however, (as many other typological surveys) only considers productive patterns and does not include information on specific formal expressions of grammatical phenomena or so-called "fossils" which no longer express the function productively.

As an example of what it means to consider fossils, let us consider markers of definiteness in Oceanic languages. **?** investigates "common noun phrase markers"[26] in Oceanic and finds that in many languages there is a reflex of what is taken to be proto-Oceanic **na/*a*, but that in some languages there is another marker with a different origin (Māori *te* for example). In Crowley's study, languages where there is no common noun phrase marking whatsoever and those with a marker that is not

[26]This term is more or less identical to a prenominal definite/specific article.

cognate with **na/*a*, are both included in type 1 (see Fig. **??**). These languages are contrasted with those that have retained some kind of reflex of **na/*a* (type 2-4 in Fig. **??**). This means that we can distinguish languages which have retained the proto-form from those that have not, but not languages which have a common noun phrase marker from those that do not.

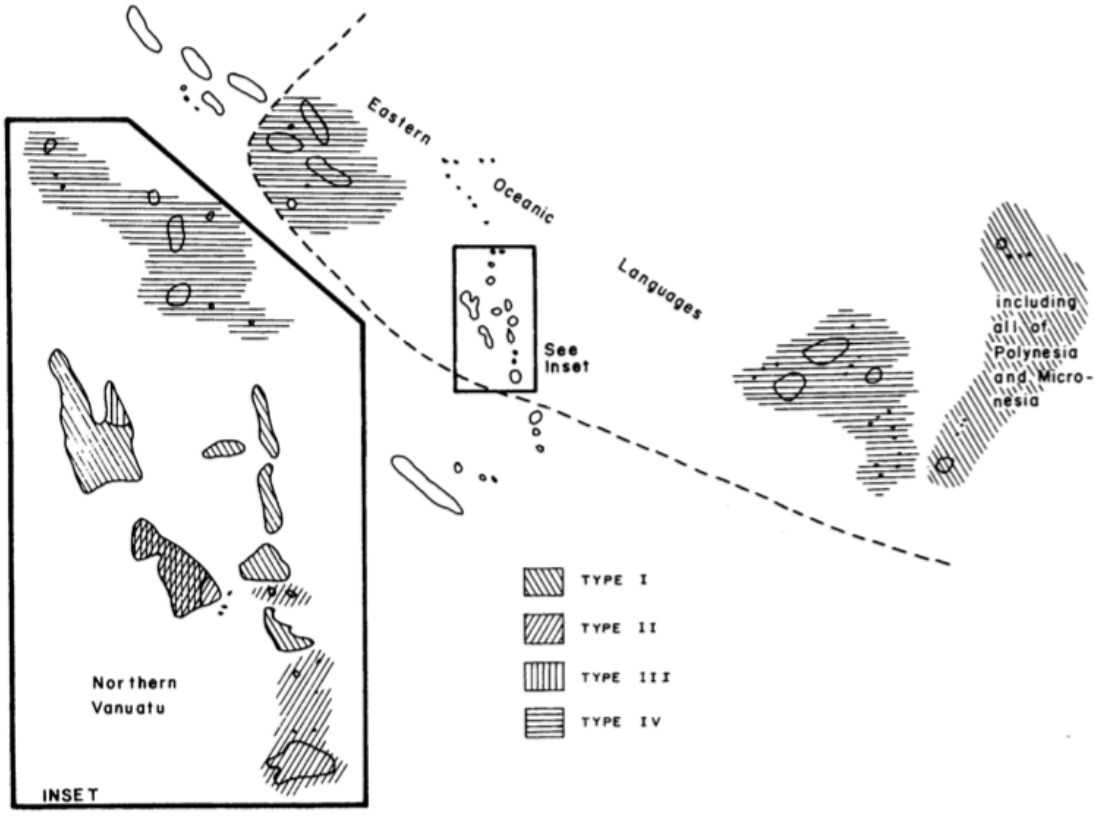


Figure 9: **Map of four different types of common noun phrase markers in Eastern Oceanic from ?: 162. Type 1: absence of common noun phrase marker or marker is not a reflex of **na /*a*, type 2: non-productive system involving a reflex of **na /*a*, type 3: productive marking involving **na /*a* as a prefix that is regularly separable from the noun and type 4: productive marking involving **na /*a* generally existing as a free-standing marker. Areas with cross-hatching show a distribution of both Type I and Type II systems, with definite areas being difficult to delineate on a map of this scale.**

In contrast, the corresponding feature in Grambank is 'GB022: *Are there prenominal articles?'* (see Fig. **??**). Languages that have *te* (like Māori) or reflexes of **na/*a* as articles before the noun both count as "yes" (1) for GB022 and those that have no prenominal marker as a "no" (0). This Grambank feature splits Crowley's type 1 into two categories and combines all the languages with reflexes of **na/*a* and *te* (or other markers) into one category with no distinction made for the form. We can now distinguish those that have a prenominal article from those that do not, but we cannot

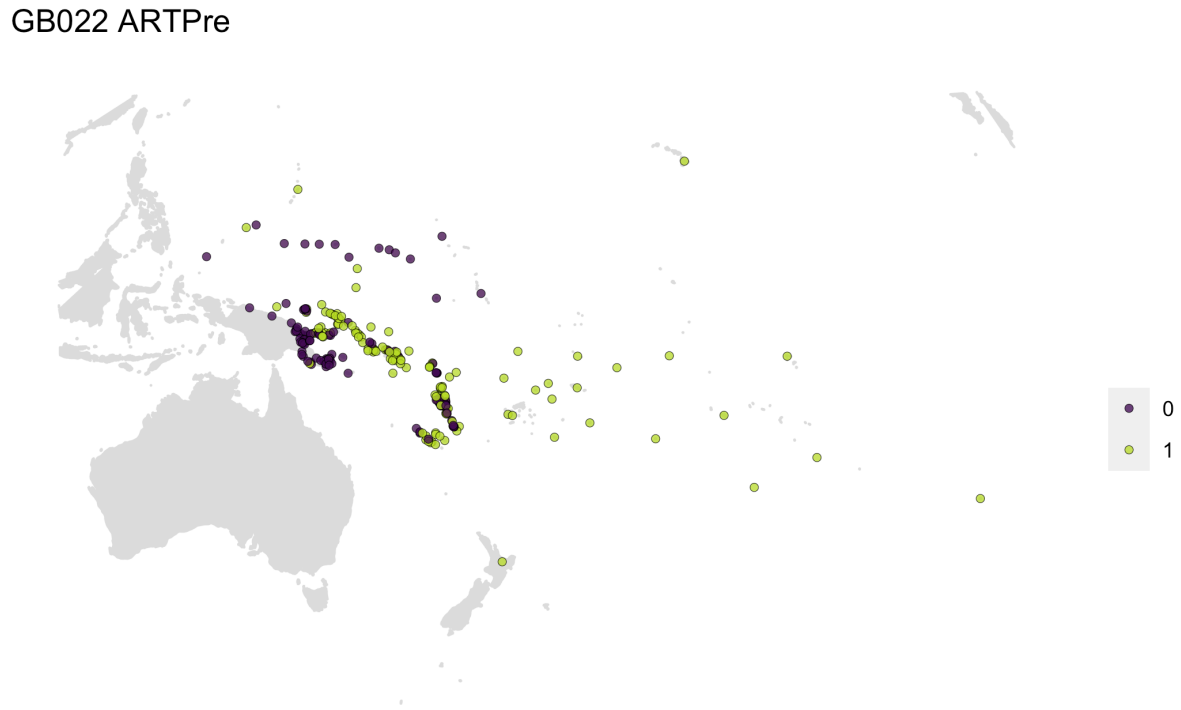


Figure 10: Map of Austronesian languages for GB022 *Are there prenominal articles?* Yellow = "yes", purple = "no".

tell apart those which have retained the proto-form **na /*a* from those which have not.

This is a difference in the kind of data that goes into the analysis, not a difference in the analytical methods themselves (compare with task a and c in fig **??** respectively).

While this principal difference is important, it should also be noted that Grambank feature GB022 have strong phylogenetic signal (negative D-estimates which are statistically similar to 0, see section **??** for details). This gives us confidence that we can move ahead.

As we have seen, Grambank data is composed of abstract features such as "is a grammatical distinction made between X and Y?". This makes it different from most ASR-studies of grammar in historical linguistics, which tend to be more focussed on particular grammatical expressions such as morphemes. Two languages can be coded alike in Grambank and many other typological surveys, but not share ancestry. It is also possible that such abstract features track inheritance beyond the particular forms. **?**: 503 notes that a particular structure of the pronominal system of Mokilese is maintained, despite the formal markers being continuously replaced. He argues that there are discourse-related reasons for maintaining this system and that the interaction between this construction and the rest of the grammar is such that the distinction is

maintained. When particular markers are lost in this system, new ones appear in their place[27]. This may be true of more features, and in such cases, languages can share a grammatical structure due to inheritance but not have the same particular forms. **?**: 400-401 also notes that while reconstructions based on lexical data are seen as more secure, they're not always possible or practical. It can be beneficial and necessary to aim to reconstruct patterns.

In the Grambank project, research assistants read published grammatical descriptions and extract information such that it fits with the definitions of our typological questionnaire (see Supplementary Material **??**). This survey of the literature on Proto-Oceanic grammar is essentially the same task. Just as with the literature on reconstructed languages, scholars sometimes disagree on the nature of contemporary languages and how they should best be analysed. It is up to the coder to make calls on which analysis to apply, what can be inferred from the literature and what should be left as unknown. It is possible to squeeze even more findings out of these publications; I have tended to be conservative in my interpretations the literature on Oceanic proto-languages. Out of the 201 (binarised) features in our questionnaire, 33% (67) were answerable for Proto-Oceanic given the existing studies in historical linguistics. The average completion per language in the whole of the Grambank data set is 85% (170).

Overall, the literature on ASR of grammar in Oceanic suggests that Proto-Oceanic was a language with a prenominal definite/specific article (**?**: 136), a distinction between inclusive and exclusive first person pronouns (**?**: 112, **?**: 184, **?**: 500, **?**: 67, 75), no gender distinctions in pronouns (**?**: 498), a dual number category in pronouns (**?**: 498, **?**: 69 and **?**: 173), a distinction between alienable and inalienable possession[28] (**?**: 69), prepositions (**?**: 167, **?**: 498), subject proclitics and object enclitics on the verb (**?**: 498-499, **?**: 83), possessive suffixes on the possessed noun (**?**: 495, **?**: 155) and a transitivising suffix on verbs (**?**: 352, **?**: 171, **?**: 80, 92). All studies cited are found in the table in Supplementary Material **??**,

Most of the time, the scholars of Proto-Oceanic are in agreement in their predictions. For example, **?**: 142, **?**: 292, **?**: xiii, 125 and **?**: 89 all propose that proto-Polynesian had a construction marking prohibitive that was different from declarative negatives. However, in some instances, there are disagreements as discussed in section **??**. In total there are 115 data points where there was either just one publication supporting the statement or if there were several they agreed. There were 3 data points where there is disagreement, these all concerned alignment of either proto-Polynesian or proto-Central Pacific.

4 Results

We are examining results from three approaches in total: a) Maximum Parsimony (MP), b) Maximum Likelihood (ML) and c) Most Common value in daughter languages

[27] **?** also notes that Goddard has observed similar patterns in Algonquian languages (**?**).

[28] A distinction can be made between three different kinds of possessive classification: alienable/inalienable, direct/indirect and dominant/inactive. For the purposes of Grambank and this study, these are treated as similar enough to be included in the same category.

(MC). For (a) and (b) we are also using three different trees: i) Glottolog, ii) **?** MCC-tree and ii) the mean values of reconstruction of a random selection of 100 (out of 4,200) trees in the Bayesian posterior of **?**. That gives 2 * 3 + 1 results, i.e. 7.

The results are divided up into three sections: 1) Concurrence with conventional historical linguistics, 2) new predictions and 3) disagreements among historical linguists.

4.1 Concordance between traditional historical linguistics and computational methods

Table **??** shows the number of False, Positive and Half-results for each method and tree[29]. Overall all methods have a large amount of True Negative/Positive results compared to False Negative/Positive, i.e. the vast majority of the time they reconstruct the same grammatical features as suggested by traditional historical linguistics literature. One of the most striking features in Table **??** is the large amounts of Half-results in the Most Common method — the method where we simply count directly what is most common in all daughters. This means that there were many instances where this approach would not confidently be able to predict a presence or absence. It is precisely in such instances that a reliable tree and more sophisticated methodology are worthwhile in order to construct the previous states well — looking at frequency alone is not sufficient.

Method	**False Negative**	**False Positive**	**Half**	**True Negative**	**True Positive**	**Total**
ML Glottolog	10	3	4	46	52	115
ML Gray et al (2009) - MCCT	9	2	9	43	51	114
ML Gray et al (2009) - posteriors	10	1	8	44	51	114
Most common	5	0	16	46	48	115
Parsimony Glottolog	8	2	4	46	55	115
Parsimony Gray et al (2009) - MCCT	6	5	10	42	52	115
Parsimony Gray et al (2009) - posteriors	7	6	4	43	55	115

Table 4: Table showing the amount of False Negative, False Positive, Half, True Negative and True Positive results.

[29]There was one feature for the ML method and the Gray et al 2009-trees where the computation could not be carried out because all the languages had the same value. In such cases, the function used (corHMM from the R-package corHMM (**?**)) gives an error because it cannot compute the rates matrix. This is why the total is 114 for ML + Gray et al 2009-trees.

Given these counts, we can calculate the concordance scores (see section **??**). These are displayed in Fig **??**. A score of 1 means identity with predictions of historical linguists and 0 means entirely dissimilar from them.

The inclusion of the half-results has the effect of evening out the differences between the performance of the different methods. The concordance scores which includes half-results for each method are more similar to each other.

The method that perform most similarly to historical linguists is Parsimony + the Glottolog 4.5 tree. The Glottolog 4.5 tree has a significant issue; it has no branch lengths and the topology is composed of a combination and compromise from several different sources as opposed to a principled and systematic investigation of data. Parts of the tree are suggested by different scholars, which means that different clades are not necessarily comparable. It does have an advantage though, and that is the sheer number of languages it includes. The overlap between languages included in Glottolog 4.5 and Grambank is greater than for the **?**-trees. It is possible that it is this sheer number of tips that gives it a greater concordance with historical linguists' predictions. The results also suggest that it is possible that historical linguists, in these specific studies, do not necessarily take into account branch lengths because there is higher agreement with the Glottog 4.5 tree (a tree without branch lengths) + the Maximum Parsimony method (a method that disregards branches altogether).

Overall, however, the methods perform similarly. There is very little that tells apart the different methods —- they are giving very similar results and all show a high degree of agreement with conventional historical linguistics. For a more detailed example of the few cases where they disagree, please see Appendix **??**.

We also carried out an analysis on whether phylogenetic signal (**?**) or the distribution of tips in either state predicts the level of agreement between each method and conventional historical linguistics (see Supplementary Materials **??** and **??**). The results show that there is no relationship between phylogenetic signal (as measured by the D-estimate (**?**)) and concurrence with HL, but that there is a weak to moderate correlation with prevalence in daughter languages. If almost all languages have a given feature, HL and the computational methods tend to both reconstruct the same state. However, if the feature values varied more, however, the agreement is reduced. This is not unexpected, if most of the languages have the same profile it is not surprising that the ancestral nodes are reconstructed the same despite differing methods.

In section **??** we suggested that features with D-estimates dissimilar to Brownian evolution according to the D-estimates p-values may not be suitable for ASR. The Grambank features in this category did not show any different behaviour from the rest in terms of agreement with HL (see appendix **??**), like the other features they did not show a significant correlation between agreement with HL and D-estimate. The vast majority agreed with HL. This may tell us that either a) there is no robust relationship between the agreement between methods and phylogenetic signal or b) that this particular measure of phylogenetic signal is faulty (for example that Brownian motion is not a reasonable assumption). Future studies should delve further into different kinds of measurements of phylogenetic signal and their potential applicability to linguistic data.

We can also compare the methods to each other. Fig. **??** shows the pairwise

concurrence (including half-results) scores between all of the methods. All of the computational methods agree more with each other than any of them do with conventional historical linguistics. The reason is most likely related to not only the difference in methodology but also the underlying data. All of the computational methods are using Grambank data and partially the same trees, whereas the underlying language-level data in conventional historical linguistics ASR is different and the tree-structure as well. For the purposes of this study, the historical linguistics literature has been translated into Grambank data-points, but there is likely to be some discrepancy in the definitions of grammatical concepts and how to apply them to each and every language in the data-set. Reconstruction in conventional historical linguistics does not always spell out the specifics of the tree structure in terms of particular splits and branches, but is rather based on broader subgroups — this can also be a contributing factor.

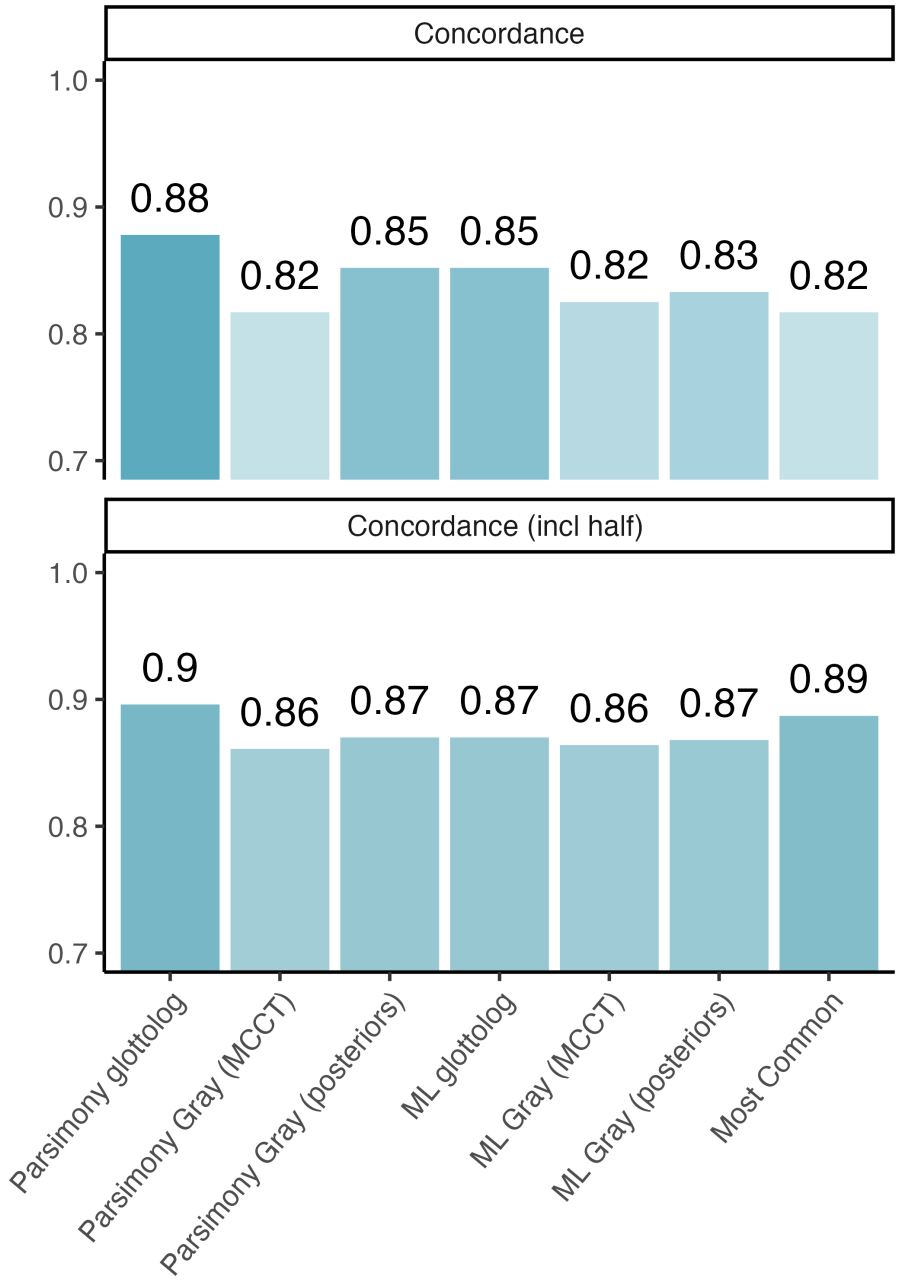


Figure 11: **Barplots of concordance scores of each method.**

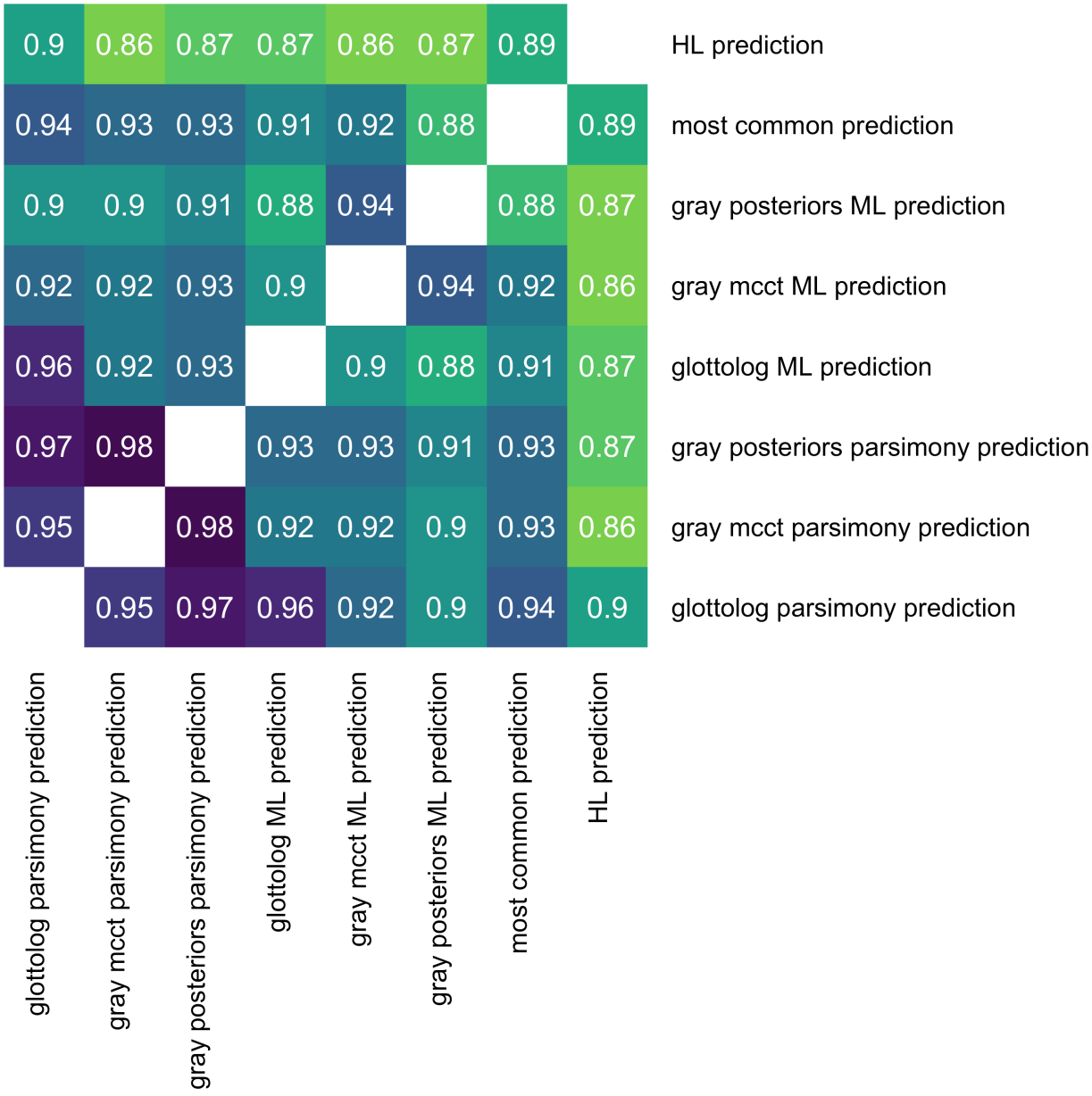


Figure 12: **Heatmap of accuracy score (including half) between reconstruction, per tree and method. Dark blue = high concurrence, lightgreen = low concurrence.**

4.2 New predictions

Besides the predictions made by historical linguists, we can also explore what else has strong support in our computational reconstructions that is not explicitly mentioned in the literature (see supplementary material **??**). There are 111 features that are predicted as present in the four proto-languages by the two methods (MP and ML) with all three trees (Glottolog, Gray et al 2009 MCCT and ditto posteriors); i.e. 6 times. For example, they propose that Proto-Oceanic has inclusory constructions, Proto-Central Pacific uses verbs for admoninal property attribution ("adjectives") and Proto-Polynesian has numeral classifiers. 107 of these 111 predictions were also the Most Common in the daughter languages, meaning that more than 60% of the languages possessed the trait. There is therefore perhaps little surprise that all the methods agree. However, there are 4 cases where both MP and ML (for all three trees) agree on the presence of a particular Proto-language structural feature despite it not being the most common (see table **??**). This is where the tree structure comes into play and adds information beyond frequencies.

Feature	**Proto-language**	**Name**
GB024b	Proto-Eastern Polynesian	Is the order of the numeral and noun N-Num?
GB093	Proto-Central Pacific	Can the P argument be indexed by a suffix/enclitic on the verb in the simple main clause?
GB421	Proto-Central Pacific	Is there a preposed complementizer in complements of verbs of thinking and/or knowing?
GB433	Proto-Central Pacific	Can adnominal possession be marked by a suffix on the possessed noun?

Table 5: Table showing the four Grambank features that were predicted as present by ML and MP in all three trees, but were not the most common feature in all languages.

4.3 Where the conflicts are: Ergativity

The nature of the alignment system of Proto-Polynesian and Proto-Central Pacific is contested (see section **??**).

Grambank has two features that pertain to these disagreements:

- GB408 *Is there any accusative alignment of flagging?*
- GB409 *Is there any ergative alignment of flagging?*

It is entirely possible for a language to be entered into the database as "yes" for both of these, i.e., from the perspective of Grambank languages are not wholly "ergative" or "accusative" — they can have both ergative and accusative flagging simultaneously.

This makes it possible for us to prove both Chung and Clark "right", the results can come out such that Proto-Polynesian had both accusative *and* ergative alignment flagging. Table **??** shows a summary of the predictions from the different historical linguists in regard to the alignment of Proto-Polynesian and Proto-Central Pacific.

Proto-language	**Feature ID**	**Prediction**	**Source**
Proto-Polynesian	GB408	Present	(**?**: 261-261), **?**
Proto-Polynesian	GB408	Absent	**?**: 106-107
Proto-Polynesian	GB409	Absent	(**?**: 261-261)
Proto-Polynesian	GB409	Present	**?**: 106-107
Proto-Central Pacific	GB409	Present	**?**: 1
Proto-Central Pacific	GB409	Absent	**?**

Table 6: Table showing the features where historical linguists disagree.

The results in fact come out strongly in favour of the proposal by Clark. Table **??** shows that MP, ML and MC all reconstruct presence for ergative flagging in Proto-Polynesian. On the matter of nominative-accusative marking, there is disagreement with the MP results all suggesting absence but the ML and MC giving a half-result.

Method	**GB408 Proto-Polynesian**	**GB409 Proto-Central Pacific**	**GB409 Proto-Polynesian**
parsimony Glottolog	Absent	Absent	Present
parsimony Gray et al (2009) - MCCT	Absent	Absent	Present
parsimony Gray et al (2009) - posteriors	Absent	Absent	Present
ML Glottolog	Absent	Absent	Present
ML Gray et al (2009) - MCCT	Half	Absent	Present
ML Gray et al (2009) - posteriors	Half	Absent	Present
Most common	Half	Present	Present

Table 7: Table showing the computational results for the features where historical linguists disagree.

As was noted earlier in section **??**, the computational reconstructions differ from those arrived at through the conventional ASR in historical linguistics primarily because the data used in this study is abstract presence or absence of structural features whereas historical linguists use specific concrete forms instead (c.f. **?**). Besides the parsimony principle (as laid out by **?**: 19 for example), expert historical linguists also take into account the plausibility of the proposed proto-language and the chain of changes

posited (**?**). It is not possible for the computational reconstructions to take these assumptions into account without having them formally described and introduced into the model, which is not possible at this time. This may be the reason for the lack of support for Chung's theory; the crucial information that underpins it is not accounted for in the analysis.

Given the topology of the trees used in this study, where the ergative flagging language Tongan is always attached to the Proto-Polynesian root at a higher level than Eastern Polynesian languages (c.f. Fig **??**), it is very likely that GB409 would be reconstructed as present for Proto-Polynesian. As Clark pointed out, it is the most parsimonious solution. However, it could still have been the case that GB408 (accusative) would have been reconstructed for Proto-Polynesian. The reasons for this may lie in different definitions of what counts as nominative-accusative or neutral in different descriptions, and/or plausibility of changes/states. As has been discussed earlier, it was not possible to include plausibility as a factor in this study.

The proposals of **?**, **??**, and **?** also involve the reconstruction of passive voice that relates to the development of the ergative systems. They suggest different pathways by which languages can develop from a nominative-accusative system to an ergative-absolutive one that relies on changes in the specifics of the passive voice construction that we, unfortunately, do not track. Given our data, which simply records the presence of a productive passive voice marker on the verb, we are not able to scrutinise the three precise theories in greater detail. The results largely support the hypothesis that Proto-Eastern Polynesian had a passive voice marker and that Proto-Oceanic and Proto-Polynesian did not. This can be seen as partial support for the proposals by **????**.

Concerning the alignment of Proto-Central Pacific, all the results, save the Most Common-model, predict an absence of ergative-marking. This is likely to be because Rotuman [rotu1241], Wester Fijian [west2519] and Fijian [fiji1243] are all coded as 0 for this feature and they split off early from the Proto-Central Pacific node. This supports the argument put forward by **?**. Similar to the Polynesian case, given the tree structure it is difficult for the computational approaches to produce another result in lieu of more information on the particulars of the development of alignment systems or possible contact.

5 Conclusions

We have investigated the history of structural features of Oceanic languages to examine how computational ASR-methods compare to reconstructions by historical linguists, including contributing to the debate on alignment in Oceanic proto-languages. This paper has compared different methodologies of ASR, both conceptually and practically (**??**). First we go through the conclusions based on the conceptually comparison, and then the results from the specific study of grammar in Oceanic languages.

Table **??** summarises the pros and cons of the different methods conceptually, as discussed in sections **??**, **??** and **??**).

Given the basic assumption that branch lengths matter (languages that are spoken at a similar time in history ought to have a similar distance to the shared ancestral

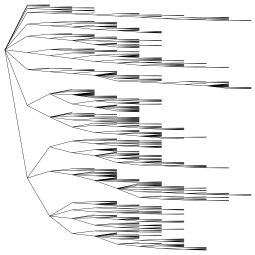
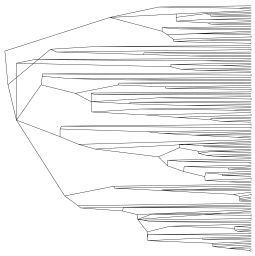
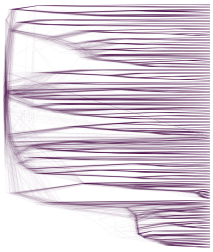
Table 8: Summary of conceptual pros and cons of the ASR-methods

ASR-Method	**Pros**	**Cons**
Conventional HL	widely used and attested; human-friendly ; takes into account complexities regarding item- and language-specific nuance and context	may ignore branch lengths; plausibility/rates of changes and plausibility of combined states are under-specified which leads to hard-to-resolve conflicts; possible: assumes slowest rate = most plausible rate
Maximum Parsimony	easy to understand; consistent; explicit	ignores branch lengths; assumes slowest rate = most plausible rate; does not allow asymmetric transition rates
Maximum Likelihood	consistent; explicit; takes into account branch lengths; dynamically estimates rates; can take further input such as priors on root state, rates etc	requires more knowledge of computational mathematics
Most Common	easy to understand	ignores the tree altogether; estimates no rates

language), we should choose methods that take that into account. If conventional HL-ASR does take branch lengths into account, it is often under-specified. It is desirable to be as consistent and explicit as possible. This enables others to interrogate the research and replicate. Computational methods allow us to be explicit about each analytical choice, which is difficult to do with conventional approaches. Given the importance of taking into account branches and the desirability of not assuming the slowest rate of change, Maximal Likelihood is the best approach out of these four.

Table **??** compares the pros and cons of the different phylogenies of this paper.

Table 9: Summary of conceptual pros and cons of the trees.

Tree		**Pros**	**Cons**
Glottolog 4.5		includes all Oceanic languages	has no branch lengths; possibly inconsistent subgrouping; many polytomies (10%); lowest proportion of D-estimates similar to 0
? - MCCT		has branch lengths; is based on explicit lexical data; transparent methodology at each step; fewer polytomies (3%)	includes fewer languages
? - random sample of 100 from posterior		has branch lengths; is based on explicit lexical data; transparent methodology at each step; much fewer polytomies (0.15%); encompasses more variation than MCCT; highest proportion of D-estimates similar to 0	includes fewer languages; takes longer time to calculate over

Once more, given that branch lengths matter if we want to understand the past we ought to go with the trees from **?**. After all, an equal amount of time has passed between the existence of a proto-community to today's extant languages, and we take our trees to estimate that history. It is possible that some languages are more conservative than others in their sounds, words or grammar, but in such cases, we should let the models figure that out rather than set all branches to the same length. When it comes to ancient languages, like Latin, Akkadian etc, it makes sense to place these at a closer distance to the root (as done for example in **?**). However, the differences in root-to-tip distances that a tree like Glottolog suggests for Oceanic seem extreme see Fig. **??** in Supplementary Material **??**).

While the MCCT is a practical summary of the 4,200 posterior trees, sampling over the actual set of posterior trees is preferable since it incorporates uncertainly in a better way and involves fewer polytomies.

Conceptually, the most reliable results *a priori* are those derived from Maximum Likelihood + random sample over posterior.

Now to the practical comparison, how does the computational approaches compare to conventional HL? Overall there is a high degree of concordance with reconstructions from expert historical linguists and all approaches. Reconstructions by both Maximum Parsimony and Maximum Likelihood agreed to a very large extent with the findings from historical linguistics. This suggests that the mechanisms at work in historical linguistic reconstruction may be similar to the concepts underlying the computational methods presented in this paper. The agreement was the highest when most of the languages had or lacked a feature (see Supplementary Material **??**), but it was generally high also when there was more variability.

The preferable method conceptually, Maximum Likelihood + random sample over **?** posterior, did not have the highest concordance (including half-results) score with HL - 0.87. The variation between the results was not large, however. The method that was the most similar to conventional HL, Parsimony + Glottolog, achieved a concordance of 0.9. The Glottolog-tree contained more matches to Grambank data-points, which is probably why it out-achieved the **?**-trees in concordance with conventional HL.

The methods which do not take into account branch lengths (Maximum Parsimony and Most Common) achieve a somewhat higher concordance with historical linguistics predictions. This is potentially troubling since it seems a sound principle that branch lengths in trees matter.

However, the general concordance between the outcomes of the different methods studied here gives us confidence that computational approaches are not so foreign to historical linguistics as they first may appear.

The classical historical linguistics toolkit is well-developed in terms of sub-grouping, but for future analysis, it would be beneficial to develop a framework regarding branch estimation as well. Pawley (personal correspondence) notes that most of the sub-grouping done in historical linguistics tends to be at the lower level, which suggests that further work on deeper relationships is also needed in order to improve the overall tree-structure (unless we have cause to believe in more community splitting events in recent time compared to long ago). Branch-estimation need not be the same as suggesting precise dates, with reasonable priors and constraints we can still produce a

result that signals uncertainty where it is prudent. While it is true that it is difficult to estimate rates of change, historical linguists do have knowledge that may reign in analysis so that it does not suggest fantastically slow or fast rates.

Computational methods need not be in conflict with conventional approaches, they can be complementary. There is certainly room for improvements in computational approaches based on knowledge from classical historical linguistics. When there were disagreements among linguists in regard to the structure of proto-languages, we saw more clearly the impact of the lack of information on the plausibility of changes and combinations, as well as contact-induced change. Currently, it is not possible to include information on these parameters directly into the computational ASR-models, because it has not been formalised in such a way that it can be included. If more work was dedicated to formalising such knowledge this may be possible in future. For example, it is possible to supply Maximum Likelihood ASR with a rates matrix that represents the plausibility of changes from one state to another (**?**: 8-9). It is also possible in other computational approaches to fix certain node states and study what the implications are.

The future of research on the history of languages probably lies in the combination of human and computational labour. Curating lexical cognate data (**?**) and constructing trees (**?**) still rely on teams of expert linguists annotating word-lists for cognacy. Methods are being developed for automatic cognate detection (c.f. **?**), but they are not yet ready to replace the vast human knowledge and experience of the experts in historical linguistics. However, once cognate classes, regular sound correspondences and structural features are identified, the work then turns to reconstructing history (sub-grouping or constructing trees/networks) and ASR (c.f. Fig. **??**). For these tasks, there are suitable computational methods that can be applied such as those in this paper - and others (c.f. **?**; **?** and **?**). Research into linguistic history can be greatly improved and effectivized by computational tools, which in turn can be given sensible priors and parameters to produce more reliable results in a future joint venture between classical and novel methods.

In order to improve these methods, we should attempt to include the knowledge that historical linguists have about plausibility of changes, harmonics of traits and contact events. Scholars of Oceanic languages have also acquired an immense knowledge of the languages, cultures and societies of the Pacific. This is why their research is so valuable and trusted. Some or all of this kind of information can be incorporated to guide computational methods, for example as priors in models. These priors should not be given the power to entirely constrain the outcomes, but guide the conclusions the method reaches given the data. In order for this to happen, more information needs to be made explicit in historical linguistics studies.

It is no doubt difficult to convey this wealth of contextual information in each and every academic paper. The task becomes further complex when we need to aggregate the knowledge and make it comparable and consistent across publications. Nevertheless, this is where I believe that the path of scientific discovery leads us next - computers and humans together.

Conversely, it is also desirable that computational methodologies and phylogenetics are made more accessible to the wider linguistics community and incorporated into

historical linguistics education. It is my perception that there is at times a disconnect between newer and classical approaches in this space, which is unnecessary and detrimental. It is my hope that this paper has made some advancements in both introducing historical linguists to some concepts in computational approaches to ASR and introducing non-linguists to ASR in historical linguistics more generally.

The more methodology and analytical choices are made explicit, the easier it is to assess the soundness of the study, replicate it and improve upon it. There are areas of this study that I look forward to receiving feedback on so that we can advance together as a field. This study aims at increasing the transparency of both the principles of reconstruction in classical historical linguistics and the corresponding computational approaches. Hopefully, this study (alongside **?** and **?**) can be a starting point for more joint ventures into our cultural past.

Acknowledgements

I am fortunate enough to have colleagues in academia who have been generous and helped me with technical matters and working through methodology conceptually, these are: Stephen Mann, Angela Chira, Jordan Brock, 華夏 Xià Huá, Cara Evans, Benedict King, Gerd Carling, Chundra Cathcart, Cristian Juárez, Viktor Martinović, Robert Tegethoff, Natalia Chousou-Polydouri, David Goldstein, Sandra Auderset, Russell Gray and Hannah Haynie. Mary Walworth has been very helpful in reviewing the Grambank coding of Oceanic proto-languages. I am also grateful to the two anonymous reviewers at Diachronica for their valuable feedback and to my former PhD supervisors Andrew Pawley, Nicholas Evans, Mark Ellison and Simon Greenhill who also provided valuable commentary. Any mistakes and misconceptions that remain are my own.

Abbreviations

HL	Historical linguistics
ASR	Ancestral State Reconstruction
ML	Maximum Likelihood
MP	Maximum Parsimony
MCCT	Maximum Clade Credibility Tree
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

Address of author

Name: Hedvig Skirgård
E-mail: hedvig_skirgard@eva.mpg.de
Postal address: Department of Linguistic and Cultural evolution
Max Planck Institute for Evolutionary Anthropology
Deutscher Pl. 6, 04103 Leipzig, Germany
ORCID: https://orcid.org/0000-0002-7748-2381

Summaries in German and French

Deutsch

Titel: Die Entwirrung der Ancestral State Reconstruction in der historischen Linguistik –Vergleich klassischer Ansätze und neuer Methoden anhand der ozeanischen Grammatik
Zusammenfassung: Ancestral State Reconstruction (ASR) ist ein wesentlicher Bestandteil der historischen Linguistik (HL). Konventionelle ASR in der HL basiert auf drei Grundprinzipien: möglichst wenige Änderungen sowie Plausibilität von Änderungen und der resultierenden Protosprachen. Dieser Ansatz weist einige Probleme auf, insbesondere die Definition dessen, was plausibel ist, und die Nichtberücksichtigung der Länge von Zweigen. Die vorliegende Studie vergleicht den klassischen Ansatz von ASR konzeptionell und praktisch mit computergestützten Werkzeugen (Maximum Parsimony und Maximum Likelihood). Computergestützte Modelle haben den Vorteil, dass sie transparenter, konsistenter und reproduzierbarer sind, und den Nachteil, dass differenziertes Wissen und Kontext nur begrenzt berücksichtigt werden. Anhand von Daten aus der Grambank Datenbank, die grammatische und strukturelle Merkmale beinhaltet, vergleiche ich Rekonstruktionen der Grammatik der ozeanischen Ursprungssprachen aus der historischen linguistischen Literatur mit solchen, die mit computergestützten Werkzeugen Mitteln erzielt wurden. Die Ergebnisse zeigen, dass es ein hohes Maß an Übereinstimmung zwischen Ergebnissen aus manuellen und computergestützten Ansätzen gibt, wobei die klassische HL tendenziell eher mit Ansätzen übereinstimmt, die die Länge von Zweigen ignorieren. Die explizite Berücksichtigung von Zweiglängen ist konzeptionell fundierter, daher sollte sich die HL mit der Verbesserung der Methoden in dieser Hinsicht befassen. Eine Kombination aus computergestützten Methoden und qualitativem Wissen ist künftig möglich und von großem Nutzen.

Français

Titre: Démêler la reconstruction de l'état ancestral dans la linguistique historique - comparaison des approches classiques et des nouvelles méthodes avec la grammaire océanienne
Résumé La reconstruction de l'état ancestral (ASR) est une partie essentielle de la linguistique historique (HL). L'ASR conventionnel dans HL repose sur trois principes fondamentaux : le moins de changements sur l'arbre, la plausibilité des changements et

la plausibilité des combinaisons de caractéristiques résultantes dans les proto-langues. Cette approche présente quelques problèmes, en particulier la définition de ce qui est plausible et l'ignorance des longueurs de branche. Cette étude compare l'approche classique de l'ASR aux outils informatiques (Maximum Parsimony et Maximum Likelihood), conceptuellement et pratiquement. Les modèles informatiques ont l'avantage d'être plus transparents, cohérents et reproductibles, et le désavantage de manquer des connaissances et des contextes nuancés. À l'aide de la base de données structurelle Grambank, je compare les reconstructions de la grammaire des langues océaniennes ancestrales de la littérature linguistique historique à celles réalisées par des moyens informatiques. Les résultats montrent qu'il existe un degré élevé d'accord entre les approches manuelles et informatiques, avec une tendance pour HL classique à s'accorder plus avec les approches qui ignorent les longueurs de branche. La prise en compte explicite des longueurs de branche est plus approprié conceptuellement. En tant que tel, la linguistique historique devrait s'engager dans l'amélioration des méthodes dans cette direction. Une combinaison de méthodes informatiques et de connaissances qualitatives est possible à l'avenir et serait très bénéfique.

Appendices

A Data and code availability

This study involves data from Grambank (v1.0, **??**, D-PLACE (v2.2.1 **?** and Glottolog (v4.5, **?**. The study also involves the use of some scripts associated with the release of Grambank v1.0 which are found in the repository grambank-analysed (v1.0, **?**). Inside of the grambank-analysed project lies Glottolog and Grambank. The trees from **?** are stored in the D-PLACE repository, the Glottolog tree in the glottolog-cldf repository.

All the R-scripts for data wrangling, analysis and plotting are found on GitHub and Zenodo.

Zenodo locations:

- Oceanic_computational_ASR https://doi.org/10.5281/zenodo.8056616
- Grambank-analysed (v1.0) https://doi.org/10.5281/zenodo.7740822
  - Grambank (v.1.0) https://doi.org/10.5281/zenodo.7740140
  - glottolog-cldf (v4.5) https://doi.org/10.5281/zenodo.5772642
- dplace-data (v2.2.1) https://doi.org/10.5281/zenodo.5554395

GitHub locations:

- Oceanic_computational_ASR https://github.com/HedvigS/Oceanic_computational_ASR
- Grambank-analysed (v1.0) https://github.com/grambank/grambank-analysed/tree/v1.0 — which in turn contains submodules of:
  - Grambank (v1.0) https://github.com/grambank/grambank/tree/v1.0
  - glottolog-cldf (v4.5) https://github.com/glottolog/glottolog-cldf/tree/v4.5
- dplace-data (v2.2.1) https://github.com/D-PLACE/dplace-data/tree/v2.2.1

B Grambank features

Table **??** contains Grambank features which serves as the input to the analysis. Multistate features have been binarised. For more details, see appendix **??**, **?**: Materials and methods: Data and the parameters-table in the CLDF-release on Zenodo of the grambank dataset version 1 (**?**). Documentation of the features, including procedures and examples are also found on a GitHub wiki https://github.com/grambank/grambank/wiki that is updated continuously. Release-versions are published regularly, as datasets on Zenodo.

Feature ID	Name
GB024a	Is the order of the numeral and noun Num-N?
GB024b	Is the order of the numeral and noun N-Num?
GB025a	Is the order of the adnominal demonstrative and noun Dem-N?
GB025b	Is the order of the adnominal demonstrative and noun N-Dem?
GB065a	Is the pragmatically unmarked order of adnominal possessor noun and possessed noun PSR-PSD?
GB065b	Is the pragmatically unmarked order of adnominal possessor noun and possessed noun PSD-PSR?
GB130a	Is the pragmatically unmarked order of S and V in intransitive clauses S-V?
GB130b	Is the pragmatically unmarked order of S and V in intransitive clauses V-S?
GB193a	Is the order of the adnominal property word (ANM) and noun ANM-N?
GB193b	Is the order of the adnominal property word (ANM) and noun N-ANM?
GB203a	Is the order of the adnominal collective universal quantifier (UQ) and noun UQ-N?
GB203b	Is the order of the adnominal collective universal quantifier (UQ) and noun N-QU?
GB020	Are there definite or specific articles?
GB021	Do indefinite nominals commonly have indefinite articles?
GB022	Are there prenominal articles?
GB023	Are there postnominal articles?
GB026	Can adnominal property words occur discontinuously?
GB027	Are nominal conjunction and comitative expressed by different elements?
GB028	Is there a distinction between inclusive and exclusive?
GB030	Is there a gender distinction in independent 3rd person pronouns?
GB031	Is there a dual or unit augmented form (in addition to plural or augmented) for all person categories in the pronoun system?
GB035	Are there three or more distance contrasts in demonstratives?
GB036	Do demonstratives show an elevation distinction?
GB037	Do demonstratives show a visible-nonvisible distinction?
GB038	Are there demonstrative classifiers?
GB039	Is there nonphonological allomorphy of noun number markers?
GB041	Are there several nouns (more than three) which are suppletive for number?
GB042	Is there productive overt morphological singular marking on nouns?
GB043	Is there productive morphological dual marking on nouns?
GB044	Is there productive morphological plural marking on nouns?
GB046	Is there an associative plural marker for nouns?

GB047	Is there a productive morphological pattern for deriving an action/state noun from a verb?
GB048	Is there a productive morphological pattern for deriving an agent noun from a verb?
GB049	Is there a productive morphological pattern for deriving an object noun from a verb?
GB051	Is there a gender/noun class system where sex is a factor in class assignment?
GB052	Is there a gender/noun class system where shape is a factor in class assignment?
GB053	Is there a gender/noun class system where animacy is a factor in class assignment?
GB054	Is there a gender/noun class system where plant status is a factor in class assignment?
GB057	Are there numeral classifiers?
GB058	Are there possessive classifiers?
GB059	Is the adnominal possessive construction different for alienable and inalienable nouns?
GB068	Do core adjectives (defined semantically as property concepts such as value, shape, age, dimension) act like verbs in predicative position?
GB069	Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) used attributively require the same morphological treatment as verbs?
GB070	Are there morphological cases for non-pronominal core arguments (i.e. S/A/P)?
GB071	Are there morphological cases for pronominal core arguments (i.e. S/A/P)?
GB072	Are there morphological cases for oblique non-pronominal NPs (i.e. not S/A/P)?
GB073	Are there morphological cases for independent oblique personal pronominal arguments (i.e. not S/A/P)?
GB074	Are there prepositions?
GB075	Are there postpositions?
GB079	Do verbs have prefixes/proclitics, other than those that only mark A, S or P (do include portmanteau: A & S + TAM)?
GB080	Do verbs have suffixes/enclitics, other than those that only mark A, S or P (do include portmanteau: A & S + TAM)?
GB081	Is there productive infixation in verbs?
GB082	Is there overt morphological marking of present tense on verbs?
GB083	Is there overt morphological marking on the verb dedicated to past tense?
GB084	Is there overt morphological marking on the verb dedicated to future tense?

GB086	Is a morphological distinction between perfective and imperfective aspect available on verbs?
GB089	Can the S argument be indexed by a suffix/enclitic on the verb in the simple main clause?
GB090	Can the S argument be indexed by a prefix/proclitic on the verb in the simple main clause?
GB091	Can the A argument be indexed by a suffix/enclitic on the verb in the simple main clause?
GB092	Can the A argument be indexed by a prefix/proclitic on the verb in the simple main clause?
GB093	Can the P argument be indexed by a suffix/enclitic on the verb in the simple main clause?
GB094	Can the P argument be indexed by a prefix/proclitic on the verb in the simple main clause?
GB095	Are variations in marking strategies of core participants based on TAM distinctions?
GB096	Are variations in marking strategies of core participants based on verb classes?
GB098	Are variations in marking strategies of core participants based on person distinctions?
GB099	Can verb stems alter according to the person of a core participant?
GB103	Is there a benefactive applicative marker on the verb (including indexing)?
GB104	Is there an instrumental applicative marker on the verb (including indexing)?
GB105	Can the recipient in a ditransitive construction be marked like the monotransitive patient?
GB107	Can standard negation be marked by an affix, clitic or modification of the verb?
GB108	Is there directional or locative morphological marking on verbs?
GB109	Is there verb suppletion for participant number?
GB110	Is there verb suppletion for tense or aspect?
GB111	Are there conjugation classes?
GB113	Are there verbal affixes or clitics that turn intransitive verbs into transitive ones?
GB114	Is there a phonologically bound reflexive marker on the verb?
GB115	Is there a phonologically bound reciprocal marker on the verb?
GB116	Do verbs classify the shape, size or consistency of absolutive arguments by means of incorporated nouns, verbal affixes or suppletive verb stems?
GB117	Is there a copula for predicate nominals?
GB118	Are there serial verb constructions?
GB119	Can mood be marked by an inflecting word (”auxiliary verb”)?
GB120	Can aspect be marked by an inflecting word (”auxiliary verb”)?
GB121	Can tense be marked by an inflecting word (”auxiliary verb”)?

GB122	Is verb compounding a regular process?
GB123	Are there verb-adjunct (aka light-verb) constructions?
GB124	Is incorporation of nouns into verbs a productive intransitivizing process?
GB126	Is there an existential verb?
GB127	Are different posture verbs used obligatorily depending on an inanimate locatum's shape or position (e.g. 'to lie' vs. 'to stand')?
GB129	Is there a notably small number, i.e. about 100 or less, of verb roots in the language?
GB131	Is a pragmatically unmarked constituent order verb-initial for transitive clauses?
GB132	Is a pragmatically unmarked constituent order verb-medial for transitive clauses?
GB133	Is a pragmatically unmarked constituent order verb-final for transitive clauses?
GB134	Is the order of constituents the same in main and subordinate clauses?
GB135	Do clausal objects usually occur in the same position as nominal objects?
GB136	Is the order of core argument (i.e. S/A/P) constituents fixed?
GB137	Can standard negation be marked clause-finally?
GB138	Can standard negation be marked clause-initially?
GB139	Is there a difference between imperative (prohibitive) and declarative negation constructions?
GB140	Is verbal predication marked by the same negator as all of the following types of predication: locational, existential and nominal?
GB146	Is there a morpho-syntactic distinction between predicates expressing controlled versus uncontrolled events or states?
GB147	Is there a morphological passive marked on the lexical verb?
GB148	Is there a morphological antipassive marked on the lexical verb?
GB149	Is there a morphologically marked inverse on verbs?
GB150	Is there clause chaining?
GB151	Is there an overt verb marker dedicated to signalling coreference or noncoreference between the subject of one clause and an argument of an adjacent clause ("switch reference")?
GB152	Is there a morphologically marked distinction between simultaneous and sequential clauses?
GB155	Are causatives formed by affixes or clitics on verbs?
GB156	Is there a causative construction involving an element that is unmistakably grammaticalized from a verb for 'to say'?
GB158	Are verbs reduplicated?
GB159	Are nouns reduplicated?
GB160	Are elements apart from verbs or nouns reduplicated?
GB165	Is there productive morphological trial marking on nouns?
GB166	Is there productive morphological paucal marking on nouns?

GB167	Is there a logophoric pronoun?
GB170	Can an adnominal property word agree with the noun in gender/noun class?
GB171	Can an adnominal demonstrative agree with the noun in gender/noun class?
GB172	Can an article agree with the noun in gender/noun class?
GB177	Can the verb carry a marker of animacy of argument, unrelated to any gender/noun class of the argument visible in the NP domain?
GB184	Can an adnominal property word agree with the noun in number?
GB185	Can an adnominal demonstrative agree with the noun in number?
GB186	Can an article agree with the noun in number?
GB187	Is there any productive diminutive marking on the noun (exclude marking by system of nominal classification only)?
GB188	Is there any productive augmentative marking on the noun (exclude marking by system of nominal classification only)?
GB192	Is there a gender system where a noun's phonological properties are a factor in class assignment?
GB196	Is there a male/female distinction in 2nd person independent pronouns?
GB197	Is there a male/female distinction in 1st person independent pronouns?
GB198	Can an adnominal numeral agree with the noun in gender/noun class?
GB204	Do collective ('all') and distributive ('every') universal quantifiers differ in their forms or their syntactic positions?
GB250	Can predicative possession be expressed with a transitive 'habeo' verb?
GB252	Can predicative possession be expressed with an S-like possessum and a locative-coded possessor?
GB253	Can predicative possession be expressed with an S-like possessum and a dative-coded possessor?
GB254	Can predicative possession be expressed with an S-like possessum and a possessor that is coded like an adnominal possessor?
GB256	Can predicative possession be expressed with an S-like possessor and a possessum that is coded like a comitative argument?
GB257	Can polar interrogation be marked by intonation only?
GB260	Can polar interrogation be indicated by a special word order?
GB262	Is there a clause-initial polar interrogative particle?
GB263	Is there a clause-final polar interrogative particle?
GB264	Is there a polar interrogative particle that most commonly occurs neither clause-initially nor clause-finally?
GB265	Is there a comparative construction that includes a form that elsewhere means 'surpass, exceed'?
GB266	Is there a comparative construction that employs a marker of the standard which elsewhere has a locational meaning?

GB270	Can comparatives be expressed using two conjoined clauses?
GB273	Is there a comparative construction with a standard marker that elsewhere has neither a locational meaning nor a 'surpass/exceed' meaning?
GB275	Is there a bound comparative degree marker on the property word in a comparative construction?
GB276	Is there a non-bound comparative degree marker modifying the property word in a comparative construction?
GB285	Can polar interrogation be marked by a question particle and verbal morphology?
GB286	Can polar interrogation be indicated by overt verbal morphology only?
GB291	Can polar interrogation be marked by tone?
GB296	Is there a phonologically or morphosyntactically definable class of ideophones that includes ideophones depicting imagery beyond sound?
GB297	Can polar interrogation be indicated by a V-not-V construction?
GB298	Can standard negation be marked by an inflecting word ("auxiliary verb")?
GB299	Can standard negation be marked by a non-inflecting word ("auxiliary particle")?
GB300	Does the verb for 'give' have suppletive verb forms?
GB301	Is there an inclusory construction?
GB302	Is there a phonologically free passive marker ("particle" or "auxiliary")?
GB303	Is there a phonologically free antipassive marker ("particle" or "auxiliary")?
GB304	Can the agent be expressed overtly in a passive clause?
GB305	Is there a phonologically independent reflexive pronoun?
GB306	Is there a phonologically independent non-bipartite reciprocal pronoun?
GB309	Are there multiple past or multiple future tenses, distinguishing distance from Time of Reference?
GB312	Is there overt morphological marking on the verb dedicated to mood?
GB313	Are there special adnominal possessive pronouns that are not formed by an otherwise regular process?
GB314	Can augmentative meaning be expressed productively by a shift of gender/noun class?
GB315	Can diminutive meaning be expressed productively by a shift of gender/noun class?
GB316	Is singular number regularly marked in the noun phrase by a dedicated phonologically free element?
GB317	Is dual number regularly marked in the noun phrase by a dedicated phonologically free element?

GB318	Is plural number regularly marked in the noun phrase by a dedicated phonologically free element?
GB319	Is trial number regularly marked in the noun phrase by a dedicated phonologically free element?
GB320	Is paucal number regularly marked in the noun phrase by a dedicated phonologically free element?
GB321	Is there a large class of nouns whose gender/noun class is not phonologically or semantically predictable?
GB322	Is there grammatical marking of direct evidence (perceived with the senses)?
GB323	Is there grammatical marking of indirect evidence (hearsay, inference, etc.)?
GB324	Is there an interrogative verb for content interrogatives (who?, what?, etc.)?
GB325	Is there a count/mass distinction in interrogative quantifiers?
GB326	Do (nominal) content interrogatives normally or frequently occur in situ?
GB327	Can the relative clause follow the noun?
GB328	Can the relative clause precede the noun?
GB329	Are there internally-headed relative clauses?
GB330	Are there correlative relative clauses?
GB331	Are there non-adjacent relative clauses?
GB333	Is there a decimal numeral system?
GB334	Is there synchronic evidence for any element of a quinary numeral system?
GB335	Is there synchronic evidence for any element of a vigesimal numeral system?
GB336	Is there a body-part tallying system?
GB400	Are all person categories neutralized in some voice, tense, aspect, mood and/or negation?
GB401	Is there a class of patient-labile verbs?
GB402	Does the verb for 'see' have suppletive verb forms?
GB403	Does the verb for 'come' have suppletive verb forms?
GB408	Is there any accusative alignment of flagging?
GB409	Is there any ergative alignment of flagging?
GB410	Is there any neutral alignment of flagging?
GB415	Is there a politeness distinction in 2nd person forms?
GB421	Is there a preposed complementizer in complements of verbs of thinking and/or knowing?
GB422	Is there a postposed complementizer in complements of verbs of thinking and/or knowing?
GB430	Can adnominal possession be marked by a prefix on the possessor?
GB431	Can adnominal possession be marked by a prefix on the possessed noun?
GB432	Can adnominal possession be marked by a suffix on the possessor?

GB433	Can adnominal possession be marked by a suffix on the possessed noun?
GB519	Can mood be marked by a non-inflecting word (”auxiliary particle”)?
GB520	Can aspect be marked by a non-inflecting word (”auxiliary particle”)?
GB521	Can tense be marked by a non-inflecting word (”auxiliary particle”)?
GB522	Can the S or A argument be omitted from a pragmatically unmarked clause when the referent is inferrable from context (”pro-drop” or ”null anaphora”)?

Table 10: Table of Grambank fetures

C Binarisation of the Grambank features

Most of the feature questions are binary (GB027: Are nominal conjunction and comitative expressed by different elements?) but a few are multi-state (GB024 What is the order of numeral and noun in the NP? 1) Num-N, 2) N-Num, 3) both). For the analysis in this study, the multi-state features have been binarised. This is because the values of the multi-state features are not independent of each other; they all contain the value “Both”. The value “Num-N” (numeral before noun) of GB024 is more similar to “Both” than it is to the other alternative “N-Num”. The relationship between the three values are not equal or independent. The table in **??** contains a list of all the features used in this study, including the binarised features. The binarization results in a total of 201 features.

D Table of historical linguistics sources surveyed

Table 11: Table of historical linguistics publications used in this dissertation for Proto-Oceanic grammar

Citation	**Title**	**Proto-Languages**	**Domains**
?	Grammatical reconstruction and change on Polynesia and Fiji	Proto-Central Pacific	Verbal markers and aspect particles
?	Some problems in Proto-Oceanic	Proto-Oceanic and Proto-Polynesian	Possession, noun phrase marking, negation, verbal markers, clusivity, word order

Citation	Title	Proto-Languages	Domains
?	Aspects of Proto-Polynesian syntax	Proto-Oceanic and Proto-Polynesian	Alignment, negation, word order, possession, noun phrase marking, voice
?	Case marking and grammatical relations in Polynesian languages	Proto-Polynesian	Alignment, word order, voice, noun phrase marking
?	Common noun phrase marking in Proto-Oceanic	Proto-Oceanic	noun phrase marking, clusivity
?	Det polynesiska verbmorfemet *-Cia*; om dess funktion i Samoanska	Proto-Polynesian	Verbal marker
?	Polynesian languages (in Facts About the World's Languages: An encyclopaedia of the world's major languages, past and present)	Proto-Central Pacific and Proto-Polynesian	Word order, verbal markers, possession, clusivity
?	A study of valency-changing devices in Proto Oceanic	Proto-Oceanic	Verbal markers
?	On ergativity and accusativity in Proto-Polynesian and proto-Central Pacific	Proto-Polynesian	Alignment, voice
?	Rotuman and Fijian case-marking strategies and their historical development	Proto-Oceanic	Possession, pronominal number
?	Proto Central Pacific ergativity: Its reconstruction and development in the Fijian, Rotuman and Polynesian languages	Proto-Central Pacific	Alignment, word order

Citation	Title	Proto-Languages	Domains
?	The Oceanic Languages, paper 4: Proto-Oceanic	Proto-Oceanic, Proto-Central Pacific and Proto-Polynesian	Negation, word order, verbal markers, clusivity, possession, pronominal number, polar interrogation, nominalisations and more
?[30]	The morphosyntactic typology of Oceanic languages	Proto-Oceanic and Proto-Polynesian	alignment, word order, verbal markers, possession, noun phrase marking

E R packages used

All the analysis for this research project was done in the free and open source programming language R, using a multitude of packages. All code and data for this project are available in supplementary material and the locations listed in Supplementary material section **??**. The scripts have been written so that any user of R can execute them. Please see the bibliography for information on package versions. Below are citations for all used packages.

?, ? and ? .

F Technical details of ASR by Maximum Parsimony and Likelihood

For Maximum Parsimony, we are using the function `asr_max_parsimony()` from the R-package `castor` (**?**) (which is an instantiation of the method described in **?**) for calculating ancestral states and stability of features. This function produces ancestral states for all nodes and reports the number of changes that was minimally required for each feature.

Ancestral state reconstruction using Maximum Likelihood Estimation involves computing each ancestral state from the tips up to the root taking into account branch lengths and the joint likelihood of states given all nodes in the tree (**????**). The Maximum Likelihood Estimation function takes a set of observations and computes the parameter distribution that maximises the likelihood given the observed data[31]. This means that for every split in the tree — every ancestral node — the Maximum Likelihood Estimation function computes what is the most likely distribution at that point

[30] This paper makes statements about "canonical" Oceanic languages, which is technically different from *reconstruction* of Proto-Oceanic. However, the author does state that the "canonic type is probably also a reflection of the morphosyntax of Proto Oceanic" (**?**: 492) and has given personal approval for the paper to be included in this study in this manner.

[31] For a gentle introduction to the concept of Maximum Likelihood Estimation, see **?**.

given the nature of the entire tree. ML can be modified so that it allows for different rates of change. An Equal Rates (ER) model assumes that the chance of transition from state A to state B and from B to A are equal. However, we as linguists are aware that certain features are more likely to be lost than gained so this is not a reasonable assumption. Therefore, we allow the model to estimate different transition rates for going from A to B and from B to A given the data. This is known as "All Rates are Different" (ARD).

When estimating ancestral states with ML, it is possible to either a) find the state at each node that maximises the likelihood (integrating over all other states at all nodes, in proportion to their probability) at that particular node (marginal reconstruction), or b) find the set of character states at all nodes that (jointly) maximize the likelihood of the entire tree (joint reconstruction). We are using marginal reconstruction in this study since it is the recommended way to deal with uncertainty in reconstruction (**?**). These two methods often yield similar results, but can differ, see **?**: 259-260, **?**: 121-126 and **?**: 5 for more details. For our data, a trial run of joint reconstruction did not generate drastically different outcomes.

For this study, the function `R-corHMM` from the `R`-package `corHMM` (**?**) is used for marginal reconstruction of ancestral states and rates of change per feature.

Languages with missing data were pruned away in all analysis, no hidden state reconstruction of values at tips was performed. The match between Glottolog 4.5 and Grambank is 271, the match between **?** and Grambank is 132. For both MP and ML, languages with missing data were dropped from the trees in the analysis for that feature. If after this pruning less than half of the tips remained, that analysis was not carried out.

For both Maximum Parsimony and Maximum Likelihood it is possible for a structural feature to appear and disappear several times along a lineage. This is different from cognate data where a cognate class cannot re-appear.

G Supplementary Figure: distance Scatterplot Matrix

Fig. **??** shows the pairwise distances between the same tips in each of the different trees (in the case of the 100 random posterior trees it's the mean distances) and in addition Gower-distances between the same languages given all Grambank features.

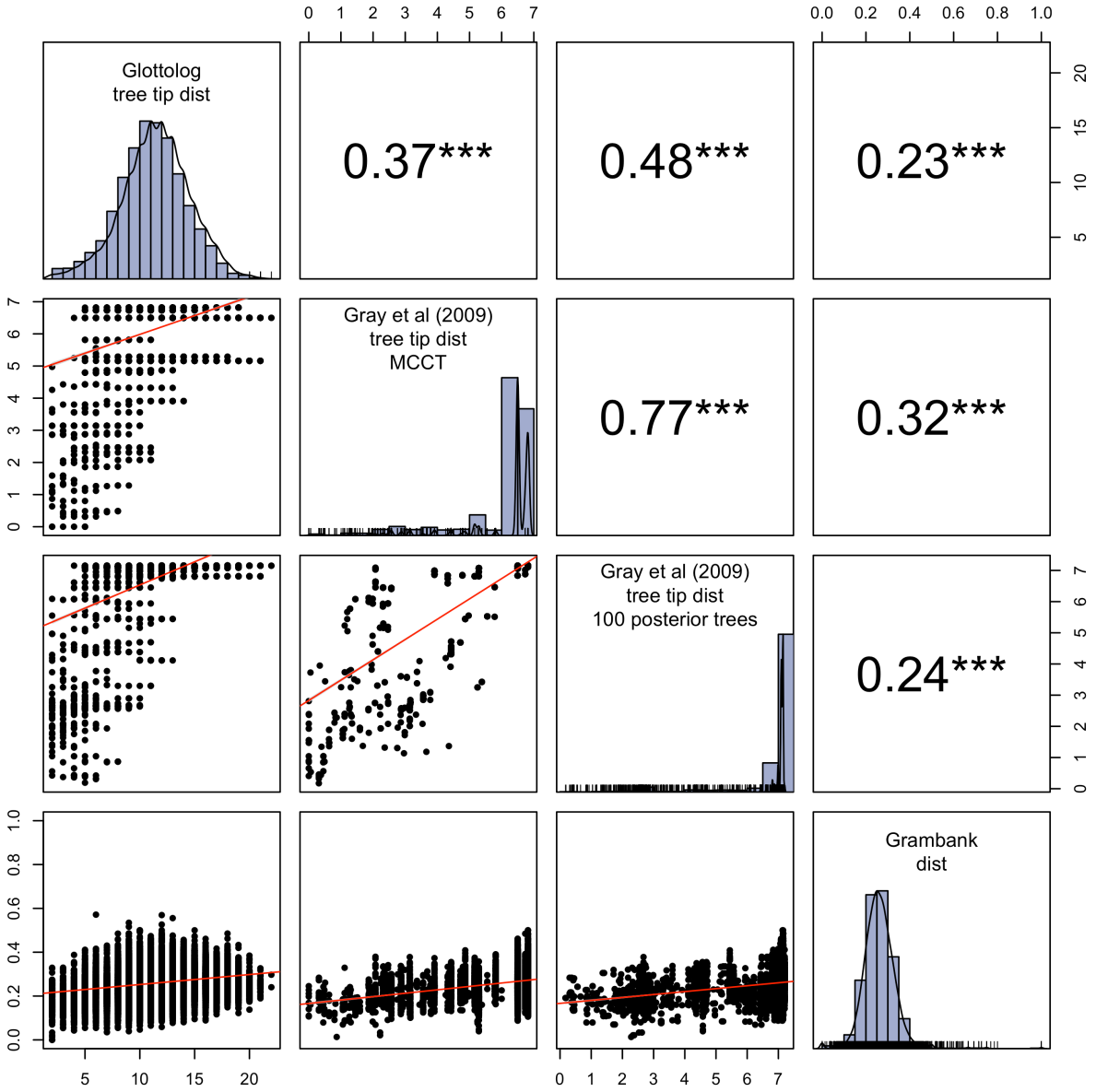


Figure 13: Comparison of distances between tips of the different trees and Grambank. Correlations are Pearson coefficients, the stars indicate the conventional p-value cut-off at 0.05.

H Supplementary Figure: tree heatmap of Gray et al (2009)-MCCT and Grambank variables

Fig. **??** shows the MCC-tree from **?** and a data-matrix of all 201 binarised Grambank variables. This data is the input for the ASR-analysis for this particular tree and the D-estimate calculation. Missing data is ignored in both sets of analysis.

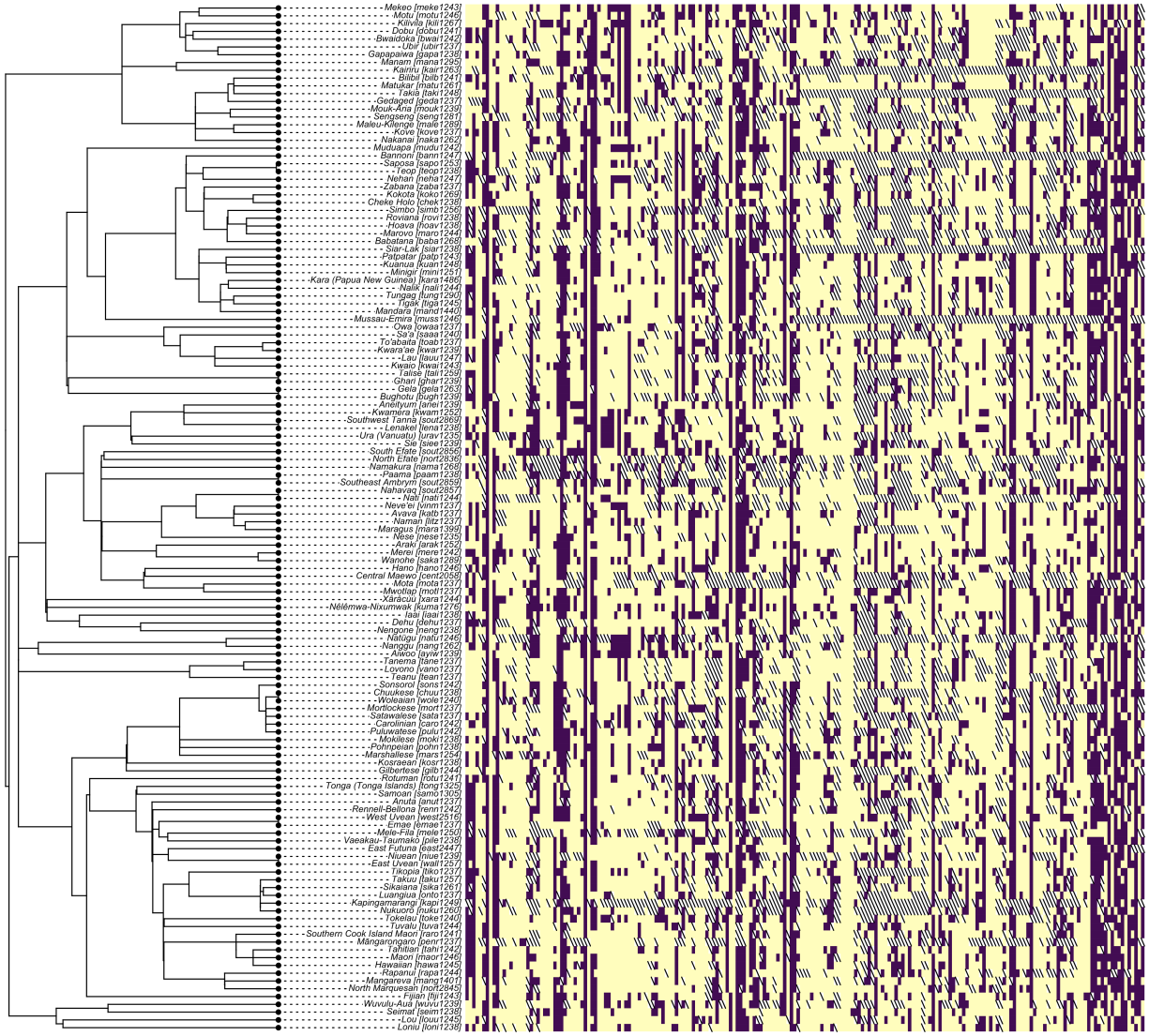


Figure 14: MCC-tree from **?** with Grambank data matrix. Purple = present, yellow = absent and striped = missing.

I Technical details on D-estimation

D-estimates are a tool for measuring phylogenetic signal in a set of binary data. Phylogenetic signal can be broadly described as the degree to which the data is generated by a given tree, or whether it was generated by some other process such as randomness. This particular method was proposed by **?** and is implemented in the R-package caper by Fritz and Orome (**?**).

The method outputs three primary values per dataset and tree: i) a D-estimate, ii) a p-value that represents how similar the data is to 0 (Brownian motion) and iii) the same kind of p-value, but instead in regard to how similar the data is to a D-estimate of 1 (randomness). If the 0-p-value is large (i.e. p>0.05) that means that the D-estimate of the data is *not dissimilar* from 0, in other worse it is *similar*. If we want to find sets of data that are similar to 0, we should look for large 0-p-values (not dissimilar = similar). The same goes for the p-values relating to 1. There can be D-estimates that are similar to both 0 *and* 1 — or neither.

The method relies on generating two kinds of simulated data: a Brownian threshold process and randomness. It then measures how similar your empirical data is to the Brownian simulation in comparison to how similar the Brownian simulation is to the random simulations. A D-estimate value of 0 represents identity to the Brownian process, 1 to the random process. D-estimates can also be smaller than 0 and larger than 1, and certainly any values in between.

The results are sensitive to how many random permutations it runs for the second set of simulated data. **?** recommends 1,000 permutations, which is also what the default value is set to for the function phylo.d in the R-package caper. However, during the work for this paper we have found further considerations that should be taken into account when working with this method — specifically in regard to the number of random permutations and skewed distributions.

I.1 D-estimate: Sensitivity to skewed distributions

While it is true that D-estimates can be smaller than 0 and larger than 1, in my experience values lower than -7 (very strong signal) and larger than 7 (very over-dispersed) are rare in empirical data. Furthermore, we would expect that if we re-run the algorithm a second time using the same data, same tree and same settings we get a similar result to the first time. This is generally true, except in certain specific situations. When the data is such that only one data point has a diverging value from the rest — for example in a set of 155 tips only one of them has the value 1 for the binary trait and all others 0 — then the algorithm struggles and produces very different results on each run, and very extreme values such as -10 on one run and 10 on another. This is problematic, and was probably not discovered by **?** and **?** because their empirical data rarely exhibited this kind of distribution (1 - 154). However, for some of the linguistic features of this study this can indeed happen.

Having identified the problem, we can also offer two solutions: a) increasing the number of random permutations and/or b) disregard data of this kind. Many thanks to (**?**) for the package documentation of caper and the paper by **?** for providing enough

methodological detail for this to be diagnosed. Stephen Mann was also invaluable to helping diagnose and address this issue mathematically.

To illustrate the problem I generated a tree with 155 tips with different distributions of binary values. The list below describes the different feature value distributions (with short names used in the plot in parenthesis) and Fig **??** shows the tree and feature value distributions

- only one tip of state 1, all other 154 tips 0 (singleton)
  - daughter with few splits from the roots (outlier)
  - in a more nested position (middle)
  - at a random position (random)
- three features with each a pair of direct sister tips of state 1, all other 153 tips 0 (sisters_a, sisters_b and sisters_c)
- two random tips with the state 1, all other 0 (two_random)
- three features with each a set of three closely related languages with the state 1, all other 152 tips 0 (triplets_a, triplets_b and triplets_c)
- three random tips with the state 1, all other 0 (three_random)
- three features with each a set of four closely related languages with the state 1, all other 151 tips 0 (quadruplets_a, quadruplets_b and quadruplets_c
- four random tips with the state 1, all other 0 (four_random)
- a cluster of 31 tips which form a clade all with 1 for the feature, all others 0 (cluster)
- 31 random tips with the same state, all others other (cluster_random)

I then proceeded to estimate the D-value for each of these 17 features, varying the number of permutations (1,000, 20,000 and 30,000). I repeated this 8 times, i.e. generating 17 * 3 * 8 D-estimates. For the entire investigation, see the script 11_phylo_d_investigation.R in the accompanying material.

The D-estimates for the singleton-features varied the most, with one iteration of the singleton outlier feature reaching a D-estimate value of 1,520 (sic). This value occurred when the number of permutations was set to 1,000. In another iteration over the same feature and the same number of iterations, the D-estimate came out as -21. While it is potentially plausible to get very small or very large values, we would expect to get *similar* values with each iteration given the same data and settings. The difference between a positive value of 1,520 and a negative of -21 is surely *unreasonably* large. When the number of permutations was increased beyond 1,000, the variance of the output with each iteration was reduced (see Fig **??**), but it was still noticeably larger in cases where the distribution was heavily skewed.

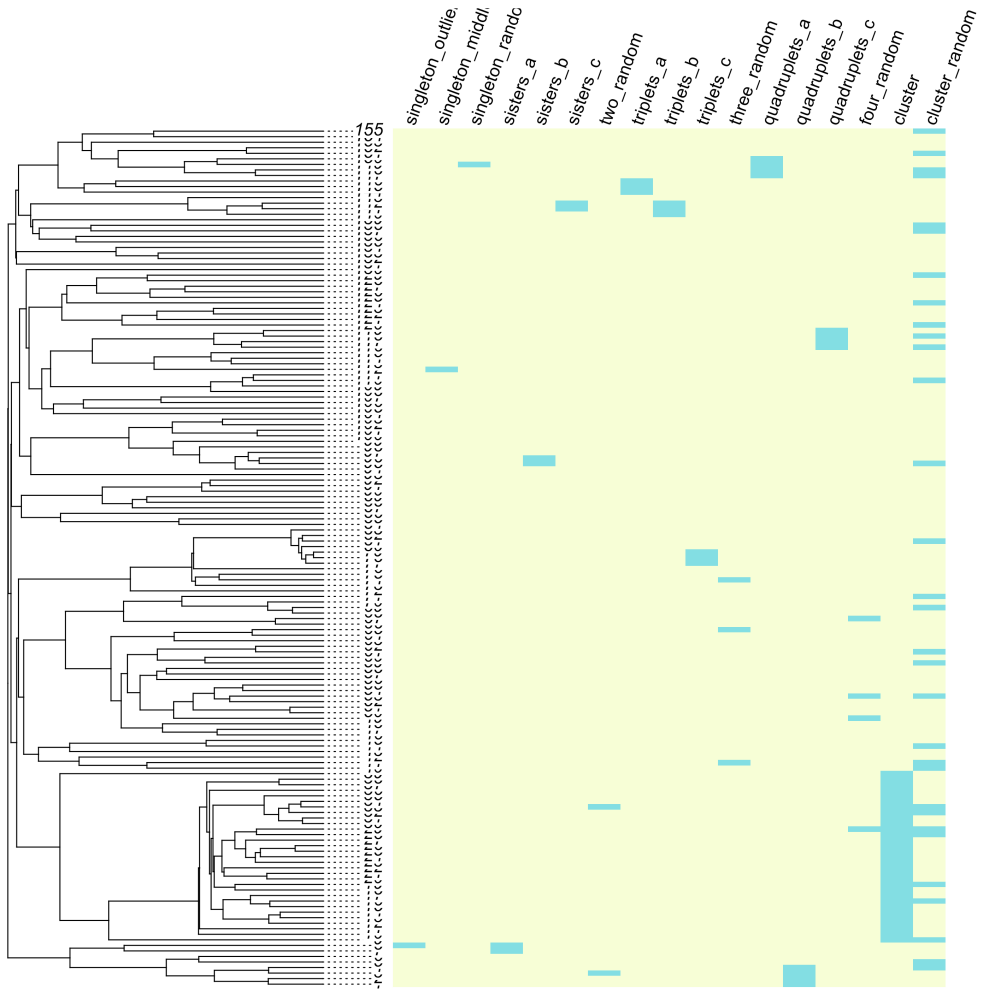


Figure 15: Tree and values heatmap for D-estimation investigation.

The cause of this issue with wildly varying D-estimates each run, especially when the feature value distribution is very skewed, has to do with the chance of generating a particular pattern of 1/154 *precisely* versus 4/151. Each time the D-estimate process is applied, a set of random and Brownian simulations are generated (the number is set by the permutations value). If the data is of the kind where 1 tip differs from all the other 154 tips (as for a few of the features in the toy example above), there is a chance that that particular position of that one value occurs in at least one of the random cases. If it does happen to occur, we would get a D-estimate that signals randomness — and conversely if it happens to be similar to the Brownian evolutionary model. If the random and Brownian simulations end up being similar the denominator (see Eq. **??**) in the formula becomes very small, which can lead to very large absolute values for the D-estimate (such as the 1,520 we saw earlier). In Eq. **??**) (**?**) r = random, b = brownian and obs = observed data.

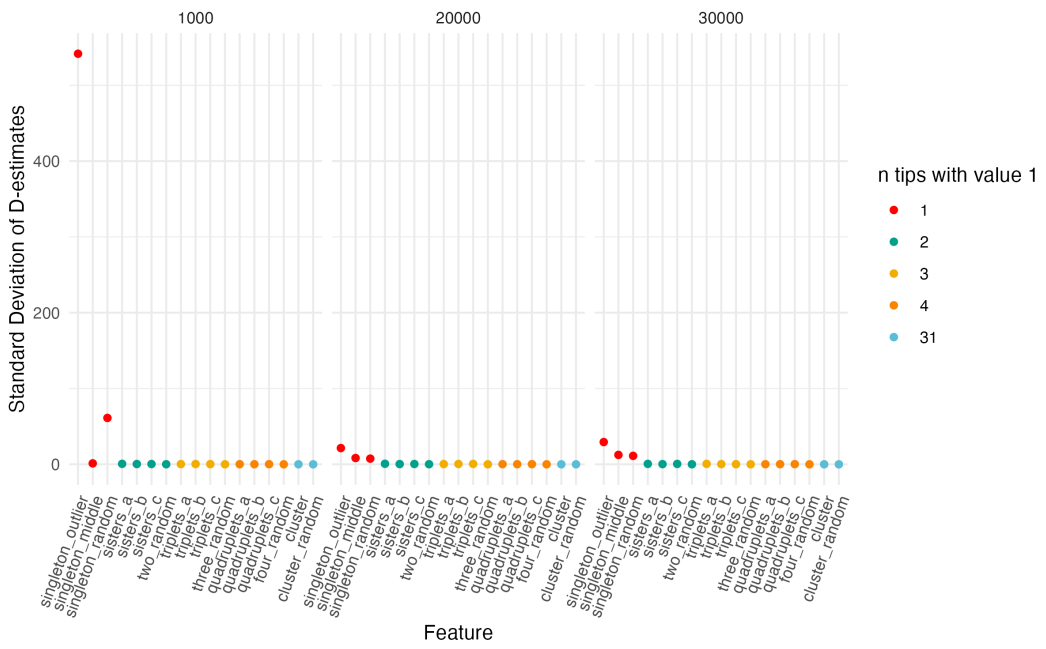


Figure 16: Scatterplot of the standard deviation of D-estimate values per feature per value of random permutations

$$\mathrm{D} = \frac{\sum d_{obs} - mean(\sum d_b)}{mean(\sum d_r) - mean(\sum d_b)} \tag{3}$$

There is less of a chance of this happening if we have more tips in each state, because those are more complicated patterns that are less likely to occur exactly in the simulated processes. Because of the possibility of this irrelevant similarity, it is necessary to increase the number of simulated permutations so that we have a larger pool of things to compare our data to. This is why the D-estimate standard deviation stabilise more in cases with skewed feature distributions if the number of permutations is increased (see Fig **??**).

Even when the number of permutations is increased to 30,000, the instances where there is a feature distribution of 1 - 154 (singeltons) are more volatile than the rest. When using this technique, it may be necessary to set aside such cases and evaluate them separately from the rest. We may want to ask ourselves: what does it mean for something that does not even form a pair to have or not have a phylogenetic signal?

If we look at the non-singleton features (the pairs, triplets, quadruplets and larger group) in the simulation example explored here in Fig. **??** we see that they behave more similarly with each iteration. Even an increase from 1 to 2 tips of the same state improves the performance of this method in terms of producing a similar value each iteration.

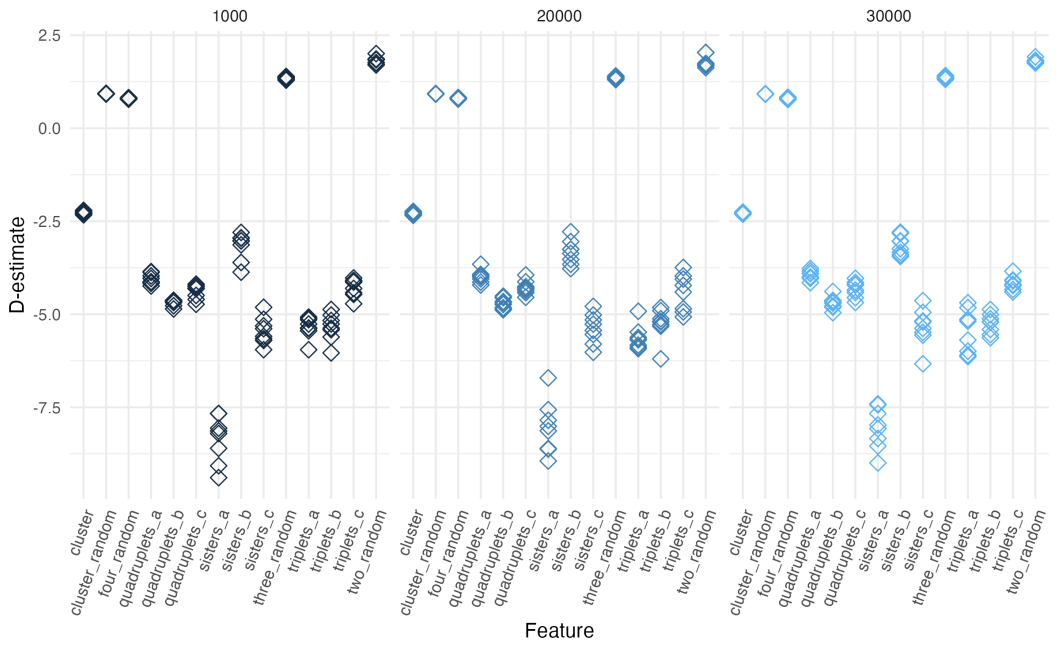


Figure 17: Scatterplot of the D-estimate values per feature per value of random permutations, for all non-singleton features. Each point represents a D-estimate value per feature, per number of permutations and per iteration.

I.2 Categories of D-estimates that do not meet the rigours of the model

In the data in this study there were cases of inappropriate D-estimates, which were possible to diagnose both by the extremity of the D-estimates, but also by examining the p-values (dissimilarity to Brownian/clumped and random/over-dispersed).

The output is grouped into 2 groups, with 3 sub-groups each. The output in the second group is not possible to include in the analysis because the conditions do not meet the model requirements, it is either impossible to conduct the analysis (all tips one state), would generate seriously unreliable results (singleton states) or shows evidence of Brownian and random being similar which also throw suspicion on the outcome. For future work, it would be desirable if the `R-function caper::phylo.d()` also output a p-value which represents the dissimilarity between the random and Brownian simulations and in addition generated a warning when the distributions are heavily skewed (for example, only 5% tips in one state).

As with the ASR-results, I also excluded output where the number of tips that had data were fewer than half of the tips in the full Oceanic-tree, i.e. for Glottolog fewer than 135.5 and for the **?**-trees 66. See counts in table **??** in section **??**. The tables below only represent the D-estimates

The p-values that are produced by the `R-function caper::phylo.d()` represent the proportion of simulations where the observed values had a smaller sum sister-clade differences compared to the Brownian simulation, and larger than the random. Pval0 = 0 means that the observed sister-clade differences were always greater than the Brownian simulations, pval0 = 1 means they were always lower. Pval1 = 0 means that

the observed sister-clade differences were always lower than the random simulations, and pval1 = 1 means that they were always greater than the random. For more details, see the source code of `caper::phylo.d()`.

- possible to include in analysis
  - (i) observed values definitely on the Brownian/clumped end of the spectrum (pval0 > 0.05 & pval1 < 0.05)
  - (ii) observed values definitely on the random/overdispersed end of the spectrum (pval0 < 0.05 & pval1 > 0.05)
  - (iii) observed values definitely between Brownian/clumped and random/overdispersed. In all of these cases, the D-estimate is between 0 and 1. (pval0 < 0.05 & pval1 < 0.05)
- *not* possible to include in analysis
  - (i) all tips same state (D-estimate is undefined)
  - (ii) singleton (only one tip has a different state from all other tips)
  - (iii) Brownian and random simulations are not sufficiently distinct from each other to get a meaningful D-estimate, observed values appear to be similar to both (pval0 > 0.05 & pval1 > 0.05). D-estimate can be <0, in between or >1.

Tables **??** and **??** shows the number of instances of each of these categories over the trees. There are fewer instances in the problematic categories and they have been excluded from further analysis with D-estimates. Because they represent cases with skewed distributions, it is possible to interpret them as representing very rare phenomena and one interpretation of that could be a strong phylogenetic signal — but the D-estimate test is not suitable. The values for the 100 trees from the posterior are averages.

tree	**similar to 0**	**similar to 1**	**dissimilar to both**
Glottolog	36	7	33
Gray - MCCT	39	16	12
Gray - posteriors	50	9	2

Table 12: Table of types of D-estimates per tree, data-points included.

tree	**all same**	**singleton**	**similar to both**
Glottolog	0	2	6
Gray - MCCT	1	3	13
Gray - posteriors	1	3	18

Table 13: Table of types of D-estimates per tree, data-points not included.

I.3 Correlation D-estimate and HL-concurrence

Phylogenetic signal could be an indication that it is easier to reconstruct a prior state. One may for example consider that it ought to be more difficult to reconstruct a state reliably if the pattern is a random phylogenetic signal (D-estimate similar to 1), and conversely that a strong signal may make it easier to reconstruct consistently, and therefore that the agreement between conventional historical linguistics findings and the computational methods applied in this paper would be higher if the phylogenetic signal is strong (=similar to 0, Brownian). This is however not the case in this study.

Fig. **??** shows the D-estimate on the x-axis (low = strong signal, high = random) and agreement with conventional historical linguistics on the y-axis. The agreement with HL is the precise value that the method predicted for the state that HL suggests. If HL suggests that the state is present at a particular node, and the computational suggests that presence has a likelihood of 0.435, the agreement value is 0.435. This is a continuous scale, but for the parsimony results it is often 0, 0.5 or 1 because of the prevalence of binary splits in the tree and the way the method works.

The results have been grouped by method and tree. In no case, save one, does the correlation reach the conventional threshold for statistical significance for the Pearson correlation ($p > 0.05$). The exception is Parsimony - Glottolog tree, and even there the correlation is weak (0.23). Furthermore, the correlation is positive which is the opposite of what we would expect. If strong phylogenetic signal (low D-estimate) predicts high agreement, then the correlation would be negative.

Each point is mapped onto one prediction of one feature and one proto-language (Proto-Oceanic, Proto-Central Pacific, Proto-Polynesian or Proto-Eastern Polynesian), but the D-estimate is only taken for the entire Oceanic tree, not for each sub-clade. The predictions for "most common" were excluded, since there is not a tree *per se* which the D-estimate can take as input to measure the phylogenetic signal. In addition, we also excluded data-points that were ill-fitting for other reasons as discussed in the previous section.

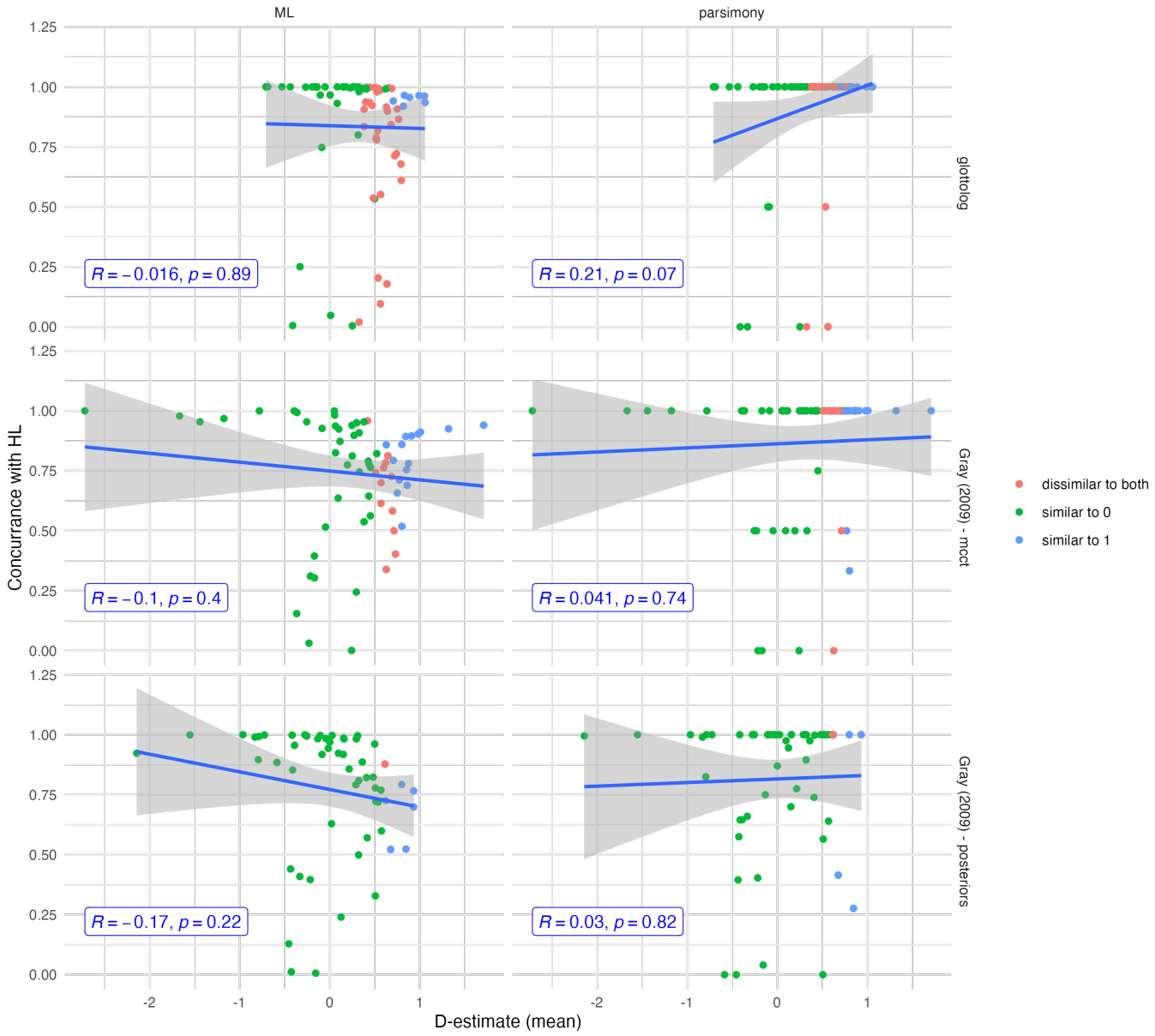


Figure 18: Scatter-plots of D-estimates (x-axis) and concurrence with conventional historical linguistics (y-axis). The points are colored based on meeting statistical thresholds of significance for being similar to 0 (Brownian) or 1 (random). The correlation statistic in blue represents a Pearson-test.

J Correlation value distributions and HL-concurrence

We can consider a much simpler approach to understanding what predicts agreement between the computational methods and historical linguists — the number of tips in each state, then we see some stronger patterns. The idea is that if very few tips are in one state and all others in the other, there is little variation that can drive disagreements between the different reconstructions. If on the other hand, the states are distributed 50%/50% then it is reasonable to assume there is a greater chance for disagreement. In fig **??**, the x-axis is the percentage of tips in the minority state — 0% indicates that all tips are of the same state (be that presence or absence) and 50% that half of the tips are in one state, half in another. 30% indicates that the state with the fewest tips had 30% of the tips. The y-axis represents concurrence with traditional historical linguistics. Each point is one structural feature in one of the four proto-languages.

All of the comparisons between HL-concurrence and the percentage of tips in minority state have a p-value lower than 0.05, which is a commonly used cut-off for significance. All correlations are negative, which is to be expected. This indicates that when tips are more evenly distributed between the two states (closer to 50% on the x-axis), there is more disagreement between each of the methods and traditional HL. Most correlations are weak (between 0.2 and 0.39), and a few are of moderate strength (between 0.40 - 0.59). There are some outliers in the lower left quadrant of the plot, these represent cases where most tips are in one state and yet there is a disagreement. One of them is discussed in greater detail in the following section.

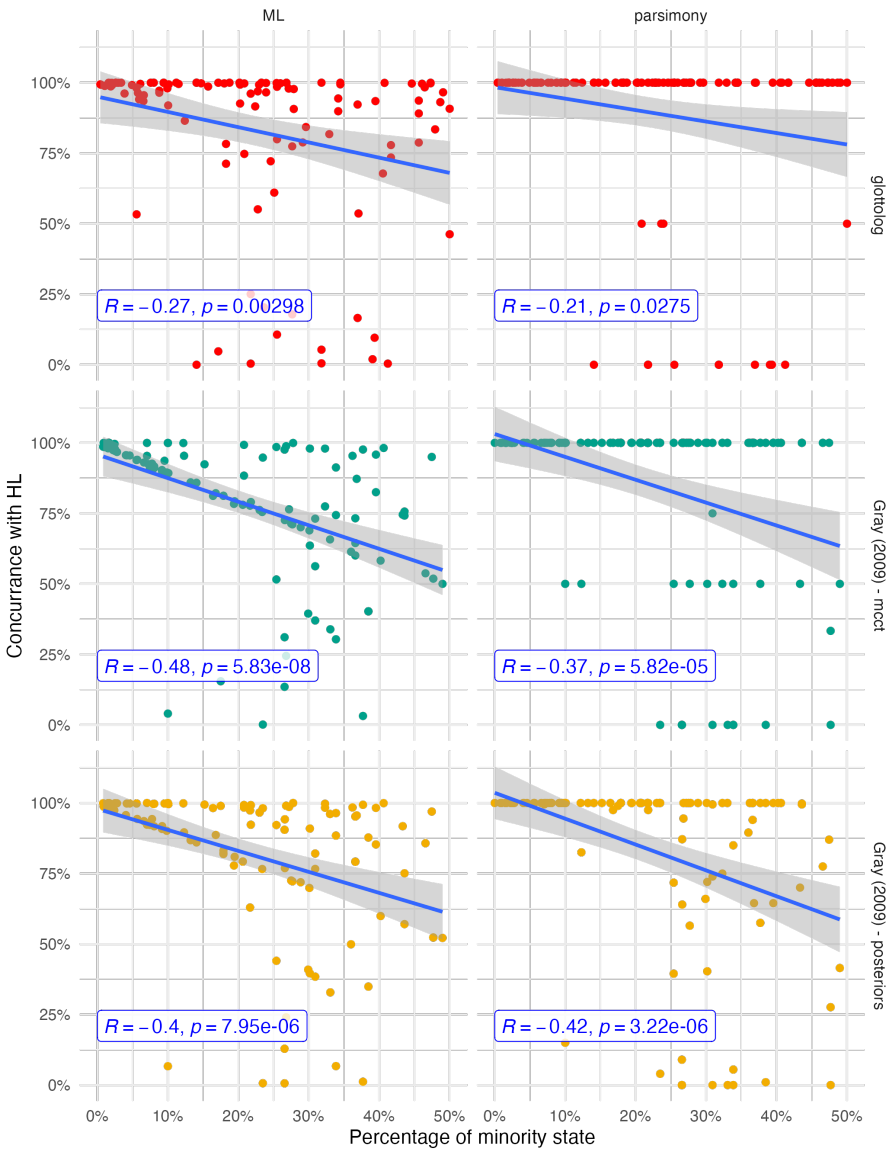


Figure 19: Scatter-plots of precentage of tips in minority state (x-axis) and concurrence with conventional historical linguistics (y-axis). The correlation statistic in blue represents a Pearson-test.

K Disagreement between methods detail: present, 50%/50% versus absent

One example of disagreement between conventional HL, Maximum Parsimony and Maximum Likelihood is GB133 'Is a pragmatically unmarked constituent order verb-final for transitive clauses¿ for Proto-Oceanic. This feature has a very low concurrence with HL for the ML method (0.04) and (**?**)) MCC-tree, despite the tip state distribution being 10%/90% which we saw in the previous section usually predicts high agreement.

Let us first consider the historical linguistics literature and the feature at hand. The coding of Proto-Oceanic as present for this feature according to conventional historical linguistics is based on the following passage from **?**:

> *Capell's suggestion that the SOV order found in many New Guinea Oceanic languages is the result of influence by Papuan (non-Austronesian) languages, almost all of which show SOV order, seems reasonable.* [..] *Still, the fact that the better-known SVO languages also tolerate certain other orders (for non- pronominal constituents) suggests that some variation occurred in POC* [Proto-Oceanic]. *In particular, occurrences of OSV and VOS order are widely distributed enough to indicate that both were possible in POC.*[32]
>
> **?**: 118

Unlike the chapter in the World Atlas of Language Structures on order in the transitive clause (**?**), the Grambank feature questionnaire does not ask about the "dominant"-type, but has 3 different binary questions about the "pragmatically unmarked" order.

- GB131 Is a pragmatically unmarked constituent order verb-initial for transitive clauses?
- GB132 Is a pragmatically unmarked constituent order verb-medial for transitive clauses?
- GB133 Is a pragmatically unmarked constituent order verb-final for transitive clauses?

It is possible for a language to be answered "yes" for more than one question if multiple orders occur (without changing the pragmatics). However, most Oceanic languages were still coded as absent for GB133. The Maximum Parsimony and Maximum Likelihood all disagree with conventional HL regarding GB133 for Proto-Oceanic - but in different ways.

Fig **??** shows the Ancestral Nodes of GB133 on the Gray et al (2009)-MCCT with the parsimony method, and Fig **??** the same tree but with the Maximum Likelihood method. These two tree figures have the same exact topology and tip states, they only vary in the reconstruction of internal nodes (proto-languages). In each of the figures, there is a set of languages at the bottom of the tree that are coded as "yes" for GB133 and these are located on the island of New Guinea or nearby. Their location in the tree is such that they form a clade that is an early offshoot from the root. For the parsimony method, that means that even though most of the tips are of another state,

[32]Pawley does note that the "basic" word order in Proto-Oceanic is likely to be SVO (Subject-Verb-Object).

this group carries a lot of weight. The parsimony method suggests that the state of the root, of Proto-Oceanic, is 50%/50%. However, the Maximum Likelihood method takes into account branch lengths and the overall tendency in the tree for the trait value "absent" to be stable (because it estimates asymmetric rates, unlike MP which assumes symmetric rates). This results in a ML-estimation of proto-Oceanic as overwhelmingly absent of verb-finality.

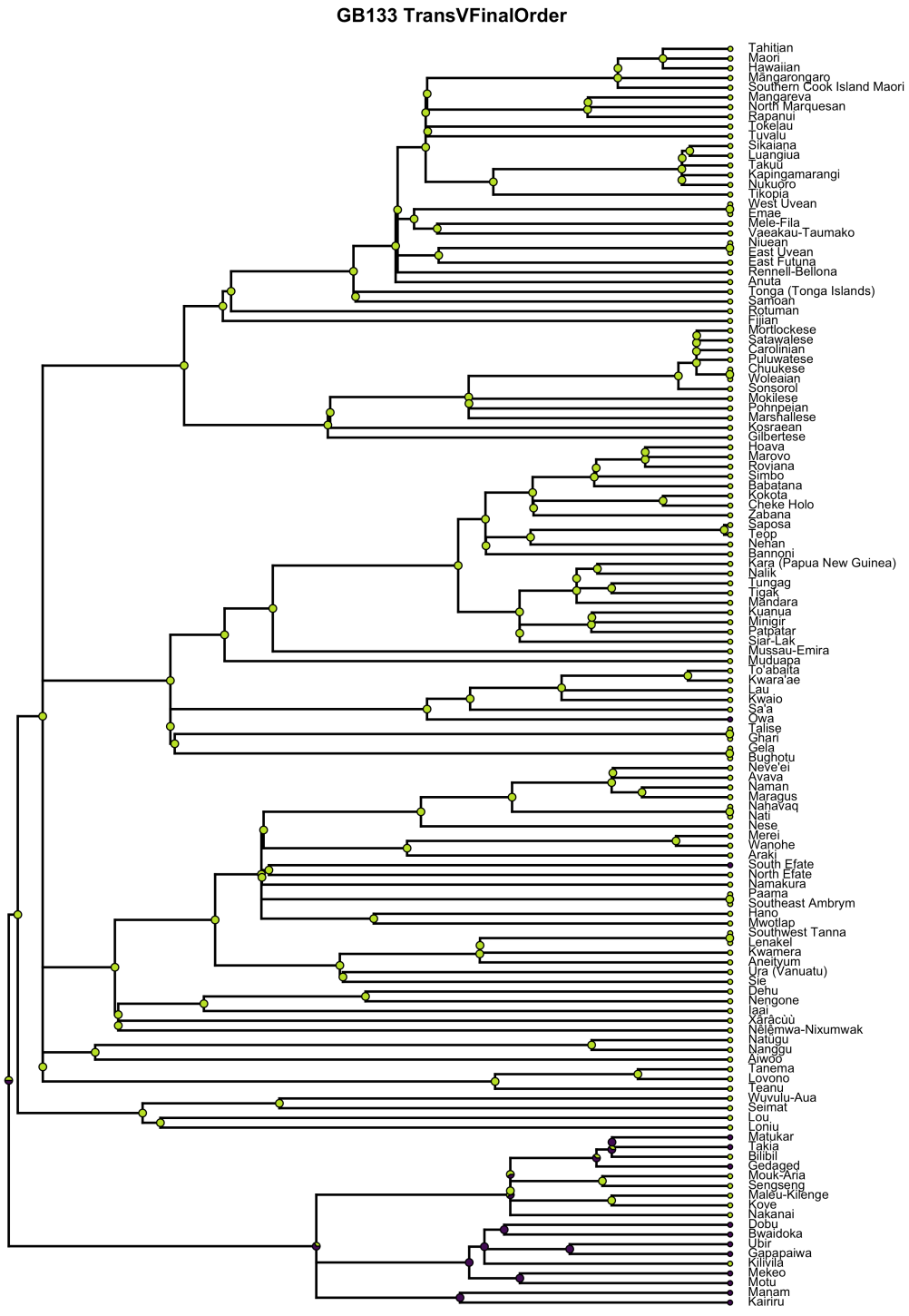


Figure 20: Gray et al 2009-tree with Maximum Parsimony method, Proto-Oceanic is reconstructed as half/half present/absent. Green = absent, purple = present. Root edge added in for visualisation purposes only.

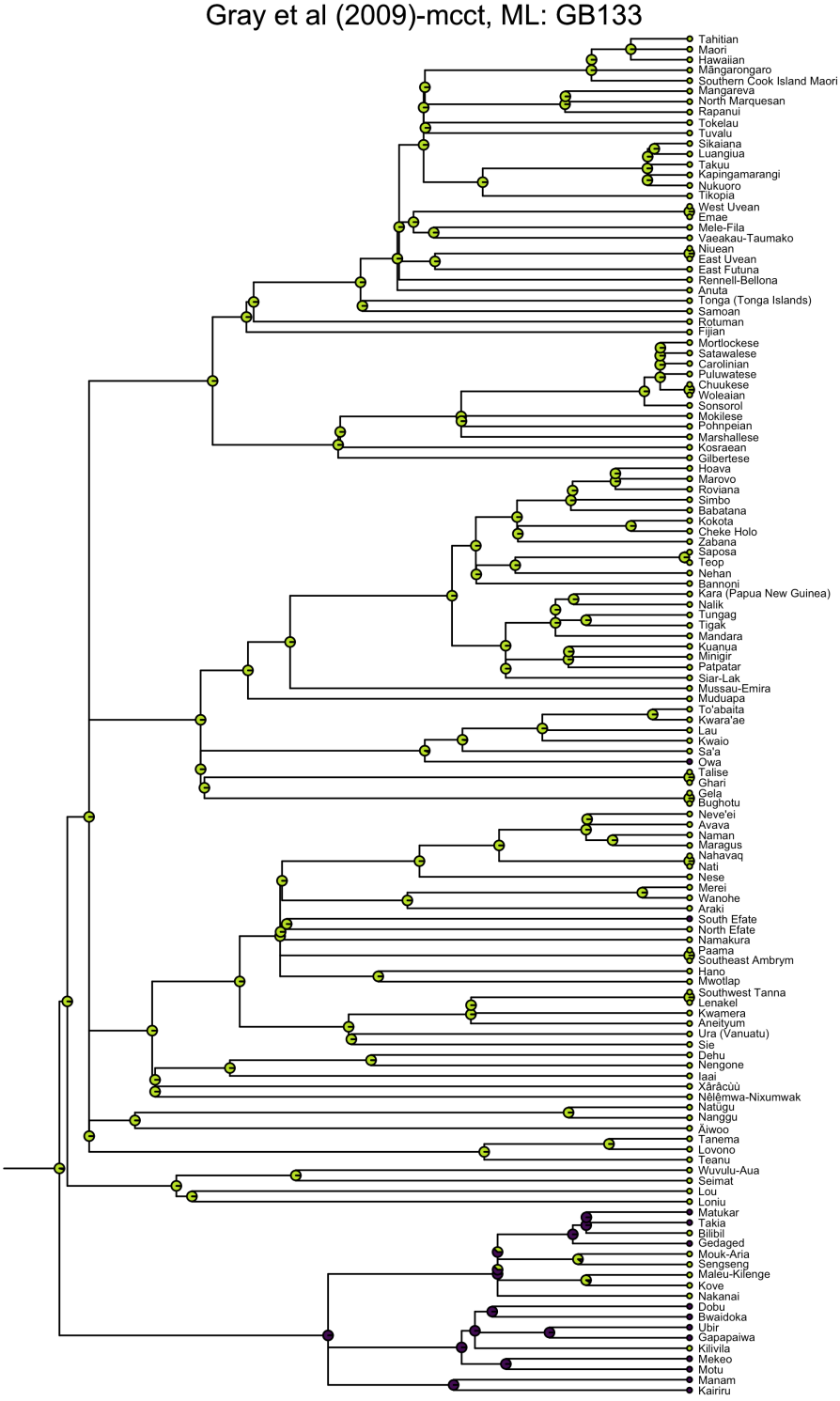


Figure 21: Gray et al 2009-tree with ML method, Proto-Oceanic is reconstructed as absent. Green = absent, purple = present. Root edge added in for visualisation purposes only.

L F1-score results

F1-scores are the harmonic mean of the precision and recall[33] (**?**: 133). It is important to note that F1-scores disregard the number of True Negatives entirely, which is relevant in our case since some of the features in proto-languages are predicted to be absent. For both measures, 0 is the worst possible score and 1 the best in terms of similarity to predictions by historical linguists.

In a similar study of ancestral states of cognate classes, **?** compared three different methods of ancestral state reconstruction for lexical data (cognate classes): Maximum Parsimony, Maximum Likelihood and Minimal Lateral Networks. They found that reconstructions using Maximum Likelihood performed the most like the predictions by historical linguists. However, **?** describe the general performance of all the computational reconstruction methods they used as "poor". **?** evaluated the methods using the F1-score. The highest F1-score was 0.79 (Austronesian language sample, Maximum Likelihood), and the worst was 0.44 (Indo-European, Minimal Lateral Networks).

The formula for F1-scores is given in Eq. **??**.

$$\frac{\text{True Positive}}{\text{True Positive} + \frac{1}{2} \times (\text{False Positive} + \text{False Negative})} \tag{4}$$

As stated in section **??**, the half-results are also interesting, the formula for F1-scores including half-results is given in Eq. **??**. For more on the caluclation of the F1-score including half results, see Supplementary Material **??**.

$$\frac{\text{True Positive} + \frac{\text{Half}}{2}}{\text{True Positive} + \frac{1}{2} \times (\text{False Positive} + \text{False Negative}) + \text{Half}} \tag{5}$$

The results of the F1-scores are shown in Fig **??**, alongside the concordance scores. The result for the plain F1-score differs from the other three, this is precisely because it ignores True Negatives. While True Negatives are not included *per se* in the calculation of F1 including half-results score, the inclusion of the half-similarity still has an impact as it makes all the methods more similar.

Compared to the F1-scores from the lexical reconstruction of **?**, all of the methods achieved higher scores. The highest ("best") F1-score in **?** was 0.79 (Austronesian language sample, Maximum Likelihood), and the worst was 0.44 (Indo-European, Minimal Lateral Networks). In this study, only statements about ancestral languages that could be mapped to Grambank-features were included. It is possible that the study by **?** had a greater overlap between all the reconstructions made by historical linguists and the meanings that they had data for. In that case, it is possible that the features that were possible to map to Grambank data were also those that Oceanic historical linguists are the most confident about — hence the higher scores of agreement (quantified as F1-scores) compared to **?**.

[33] Precision is True Positives divided by True Positives + False Positives, recall is True Positives divided by False Negatives + True Positives. F1-score = 2 * ((precision*recall) / (precision + recall)) (**?**).

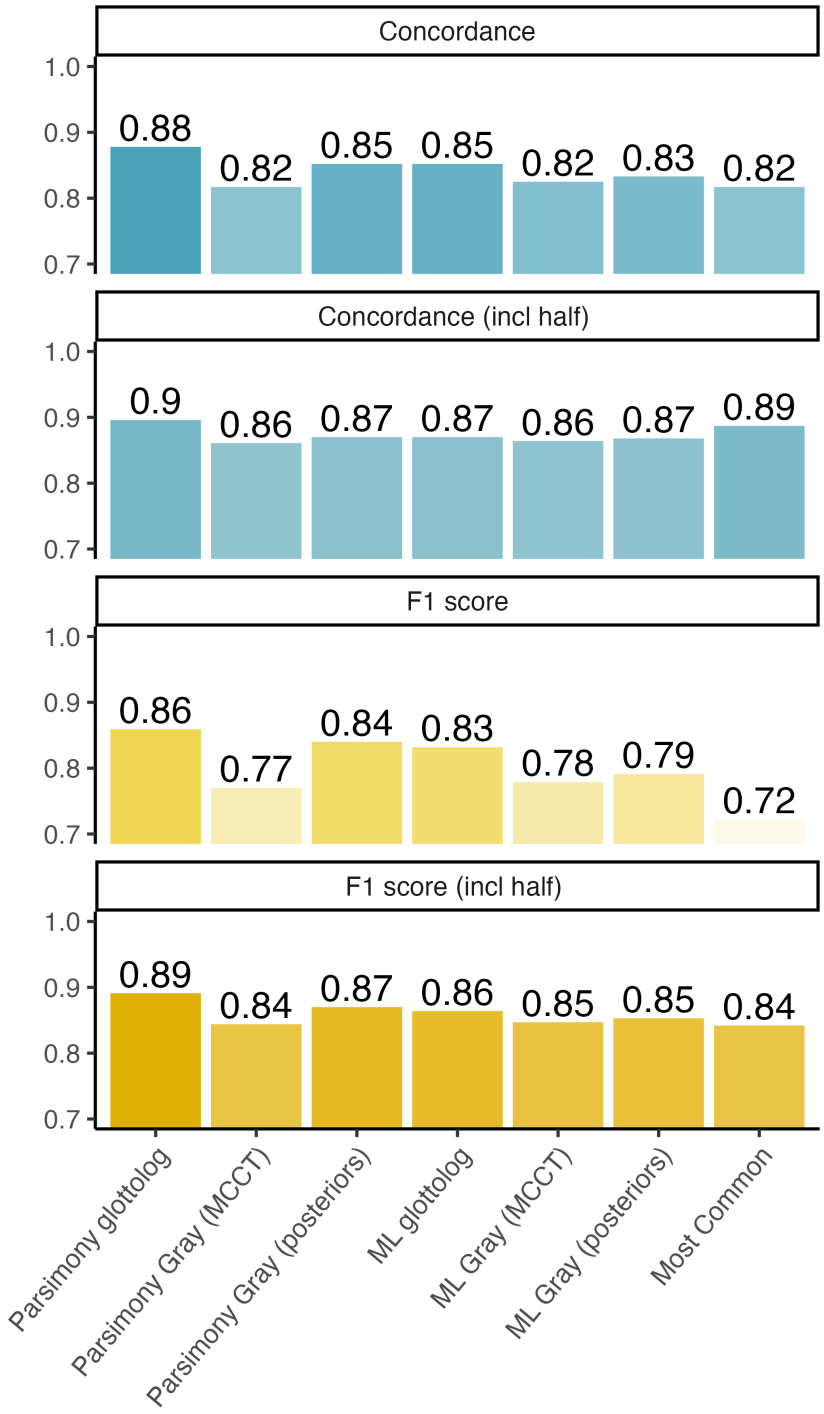


Figure 22: **Barplots of concordance and F1-scores of each method.** NB that the y-axis starts from 0.7.

M Mathematics of the F1-score including half-results

I am very grateful for the assistance of Stephen Mann in working out the mathematics of these scores as they incorporate the Half-results.

M.1 Standard definitions

The F1-score is the harmonic mean of precision and recall (**?**).

$$\begin{aligned} F_1 &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \\ &= \frac{\text{TP}}{\text{TP} + \frac{1}{2} \times (\text{FP} + \text{FN})} \end{aligned}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

M.2 Half-result definitions of precision and recall

The half-result-definitions of precision and recall add one half of the half-counts to the numerator, and all of the half-counts to the denominator:

$$\text{precision}_{\text{half}} = \frac{\text{TP} + \frac{\text{H}}{2}}{\text{TP} + \text{FP} + \text{H}}$$

$$\text{recall}_{\text{half}} = \frac{\text{TP} + \frac{\text{H}}{2}}{\text{TP} + \text{FN} + \text{H}}$$

M.3 The question

We want to define $F_{1,\text{half}}$. A natural way to do it would be to follow the rule defined above, i.e.

$$F_{1,\text{half?}} = \frac{\text{TP} + \frac{\text{H}}{2}}{\text{TP} + \frac{1}{2} \times (\text{FP} + \text{FN}) + \text{H}}$$

However, we want to ensure $F_{1,\text{half}}$ has the same relationship with $\text{precision}_{\text{half}}$ and $\text{recall}_{\text{half}}$ as F_1 has with precision and recall. So we need to determine whether the following equation is true:

$$2 \times \frac{\text{precision}_{\text{half}} \times \text{recall}_{\text{half}}}{\text{precision}_{\text{half}} + \text{recall}_{\text{half}}} \stackrel{?}{=} \frac{\text{TP} + \frac{\text{H}}{2}}{\text{TP} + \frac{1}{2} \times (\text{FP} + \text{FN}) + \text{H}} \tag{6}$$

M.4 The proof

We will expand the left-hand side of (**??**) and show it is equal to the right-hand side. Let's forget about the 2× for now (we will reintroduce it at the end). Expanding the numerator gives:

$$\frac{\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)}{\left(\mathrm{TP}+\mathrm{FP}+\mathrm{H}\right)\left(\mathrm{TP}+\mathrm{FN}+\mathrm{H}\right)}$$

Expanding the denominator gives:

$$\begin{aligned}
&\frac{\mathrm{TP}+\frac{\mathrm{H}}{2}}{\mathrm{TP}+\mathrm{FP}+\mathrm{H}}+\frac{\mathrm{TP}+\frac{\mathrm{H}}{2}}{\mathrm{TP}+\mathrm{FN}+\mathrm{H}} \\
&=\frac{\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)\left(\mathrm{TP}+\mathrm{FN}+\mathrm{H}\right)}{\left(\mathrm{TP}+\mathrm{FP}+\mathrm{H}\right)\left(\mathrm{TP}+\mathrm{FN}+\mathrm{H}\right)}+\frac{\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)\left(\mathrm{TP}+\mathrm{FP}+\mathrm{H}\right)}{\left(\mathrm{TP}+\mathrm{FN}+\mathrm{H}\right)\left(\mathrm{TP}+\mathrm{FP}+\mathrm{H}\right)} \\
&=\frac{\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)\left(2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+2\times\mathrm{H}\right)}{\left(\mathrm{TP}+\mathrm{FP}+\mathrm{H}\right)\left(\mathrm{TP}+\mathrm{FN}+\mathrm{H}\right)}
\end{aligned}$$

When we put the numerator back on top of the denominator, both of their respective denominators cancel out, because they are both (TP+FP+H)(TP+FN+H). So we end up with *the numerator of the numerator* on top of *the numerator of the denominator*, like so:

$$\begin{aligned}
&\frac{\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)}{\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)\left(2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+2\times\mathrm{H}\right)} \\
&=\frac{\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)}{2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+2\times\mathrm{H}}
\end{aligned}$$

Finally, we bring back the 2× from the beginning:

$$\begin{aligned}
&2\times\frac{\left(\mathrm{TP}+\frac{\mathrm{H}}{2}\right)}{2\times\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+2\times\mathrm{H}} \\
&=\frac{\mathrm{TP}+\frac{\mathrm{H}}{2}}{\mathrm{TP}+\frac{1}{2}\times\left(\mathrm{FP}+\mathrm{FN}\right)+\mathrm{H}}
\end{aligned}$$

And we have our suggested definition of $F_{1,\mathrm{half}}$ as required.

N Further details on the tree phylogeny

The tree from **?** contains duplicates in terms of glottocodes (see for example Nakanai). This is because it is a tree of word-lists for languages (doculects) rather than languages themselves. There are also some instances where multiple dialects of one language are

included. For the analysis, only one tip per language was retained, based on which had best coverage in the underlying data for the tree (i.e. the Austronesian Basic Vocabulary Database, ABVD (**?**)). This means that duplicate glottocodes were reduced to one, be it due to multiple word-lists or dialects. The specific analytical choices are found in the following three R-scripts:

- Oceanic_computational_ASR/code/01_requirements.R
- Oceanic_computational_ASR/code/analysis_scripts_gray_mcct/03_get_gray_tree_mcct.R
- Oceanic_computational_ASR/code/analysis_scripts_gray_all_posterior/03_process_gray_tree_posteriors.R

For both Maximum Parsimony and Maximum Likelihood the tree were first pruned down to only languages where there is data in Grambank for each given feature, i.e. the ASR-analysis never contains tips with missing or ambiguous data. Missing data varies with features, so each analysis per tree and method differs in number of tips.

Regarding branch lengths, most of the trees in the analysis are not ultrametric, i.e. the distances between the tips and the roots are not all the same. If we use trees to represent history and time, then an ultrametric, or near-Ultrametric, tree is a more reasonable representation of said histories when we assume that the languages at the tips existed at the same time. Fig **??** illustrates different configurations of branch lengths using only Nuclear Polynesian languages. It is reasonable to assume that the data gathered on these languages represent similar time-slices to each other, i.e. the representation of Rapa Nui is not considerably "younger" as a language than Tongan. If the tips included ancient languages, such as Sanskrit or Akkadian, it may be possible for such tips to have a shorter distance to the root than the others. However, if the languages are of a similar "age", the tree ought to be ultrametric or near-ultrametric if we understand tree length as representing time.

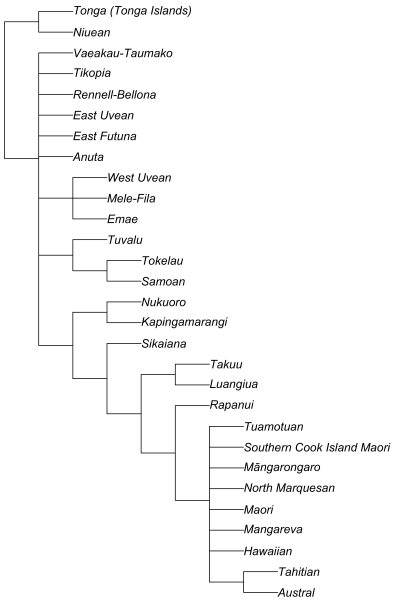
The Glottolog genealogical classification follows common principles in historical linguistics by focusing on the validity of sub-grouping, not branch lengths. The Glottolog 4.5 tree does not include any information about branch lengths. This is interpreted as the same as if all branches are of the same length (they are explicitly all set to 1 in the analysis). In order to illustrate what this entails, consider the difference between Fig. **??** and Fig. **??**. The first is the Glottolog tree of Nuclear Polynesian as found originally, i.e. with all branches of the same length (1). The second is a transformation of the first, it has been made ultrametric by Grafen's transform (**?**), which is one of several approaches to making a tree ultrametric in lieu of branch lengths directly from the data. The second tree is *not* used in the analysis, it is included here only to illustrate how the same sub-grouping of languages can be expressed when the branch lengths are changed. Transformations of branch lengths should be carried out with great care and with good reason. It is not clear what transformation of branch lengths in the Glottolog 4.5-tree is appropriate, which is why none has been carried out in the analysis of this paper. Keeping the Glottolog 4.5-tree lacking branch lengths may also be more true to historical linguistics methodology.

The branch lengths in the trees from **?** are derived from the dynamics of the data — the word-lists — and certain priors regarding island settlement. The MCC-tree of **?** is not ultrametric, but very close to it (see Fig **??**. After pruning to the subset that overlaps with Oceanic languages in Grambank (as described above), the difference between the tip with the largest distance to the root and the smallest is tiny (3.408377 - 3.408362). In the random sample of 100 from the 4,200 posteriors trees, the case is the same as with the MCCT — they are not perfectly ultrametric but very close (see Fig **??**. The trees from **?** may not be perfectly ultrametric, but they are much more closer to ultrametric than the Glottolog tree, as can be seen by comparing the visualisations in Fig **??**.

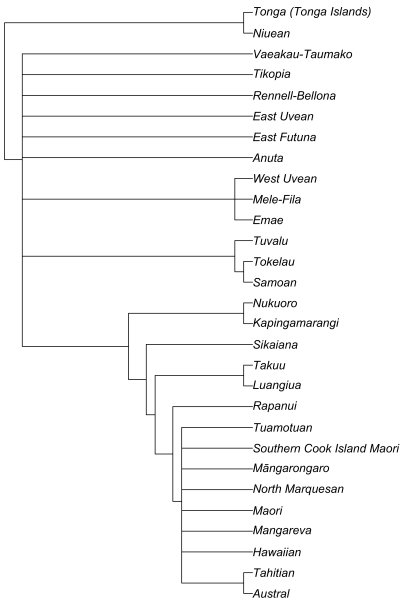
Concerning binary splits, there are non-binary splits in the Glottolog 4.5 tree, the **?** MCCT and in 64 of the 100 posterior trees. For this analysis, we have chosen to not resolve these polytomies into binary splits in order to stay as true as possible to the original phylogeny. There are branches of length 0 in the MCCT and posterior trees. It is not possible to collapse these into polytomies as this may in cases introduce basal polytomies. Instead, 0.00011 length was added to all branches. Doing this removes branches of length 0 while maintaining the relative lengths of all branches in the tree.

In some cases, pruning a given posterior tree to the relevant tips resulted in the tree becoming unrooted. In such cases, the tree was re-rooted using midpoint rooting `castor::root_at_midpoint()` **?**. There were 4 such cases in the random sample of posteriors trees (random seed = 147).

All of the wrangling of the trees is found in the data analysis R-scripts that accompany this paper.



(a) Glottolog tree, all branches have the same length.



(b) Glottolog tree, made ultrametric with Grafen's method (**?**).

O Further details on the Grambank coding of proto-languages

Another example of how information in the publications was turned into Grambank feature coding relates to verbal markers encoding subjects and objects, as proposed by **?** among others. In their book, there is a paper on reconstructions of grammar for Proto-Oceanic and in the section on the basic verb phrase we find the statement below:

> *Attached to the verb root were a subject proclitic and, if the verb had a non-generic object, an object enclitic.*
>
> **?**: 83

This statement, together with a verb schema provided in the section, support the notion that Proto-Oceanic had subject proclitics and object enclitics. We can also infer from this publication as a whole that the authors believe Proto-Oceanic in fact did *not* have subject *en*clitics and object *pro*clitics. This second prediction relies on absence of evidence and is less strong than the first, but given that the whole paper is void of any description of object proclitics or subject enclitics being a possibility (including the verb schema) and argument structure is well-discussed, we may dare to make this leap. This information can be translated into the Grambank questionnaire by positing absence and presence for the six relevant features that concern argument marking on the verb (where S stands for subject of intransitive, A for subject of transitive and O for object; see table **??**).

Table 14: Example of predictions from historical linguistics as rendered in Grambank features.

Grambank ID	**Question**	**Proto-language**	**Expert prediction**	**Reference**
GB089	Can the S argument be indexed by a suffix/enclitic on the verb in the simple main clause?	Proto-Oceanic	Absent	**?**: 498-499, **?**: 83
GB090	Can the S argument be indexed by a prefix/proclitic on the verb in the simple main clause?	Proto-Oceanic	Present	**?**: 498-499, **?**: 83
GB091	Can the A argument be indexed by a suffix/enclitic on the verb in the simple main clause?	Proto-Oceanic	Absent	**?**: 498-499, **?**: 83
GB092	Can the A argument be indexed by a prefix/proclitic on the verb in the simple main clause?	Proto-Oceanic	Present	**?**: 498-499, **?**: 83
GB093	Can the P argument be indexed by a suffix/enclitic on the verb in the simple main clause?	Proto-Oceanic	Present	**?**: 498-499, **?**: 83
GB094	Can the P argument be indexed by a prefix/proclitic on the verb in the simple main clause?	Proto-Oceanic	Absent	**?**: 498-499, **?**: 83