Disentangling Ancestral State Reconstruction in historical linguistics: Comparing classic approaches and new methods using Oceanic grammar

Hedvig Skirgård[1]

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

September 22, 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 ignore branch lengths. Explicitly taking branch lengths into account 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 the future and would be of great benefit.

Keywords: Oceanic languages, classical ancestral state reconstruction, computational ancestral state reconstruction, grammar

1 Introduction

Historical linguistics (HL) 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. Insights from HL have contributed to 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 HL are also influential in other historical sciences, for example archaeology (cf. **?**: S364).

HL-researchers make use of a wealth of knowledge not only of the languages themselves but also the cultures, societies and history of the regions they research. At times, it is difficult to be explicit about all the background information and context that goes into an analytical steps in HL – which makes it hard for someone else to replicate and examine the study thoroughly.

In this study, I focus on one particular subset of the HL 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. I discuss the underlying mechanisms of the conventional "manual" approaches to reconstruction in HL and compare the conceptual framework to computational alternatives. As a practical example, we compare reconstructions of Oceanic grammar. This study makes visible the opportunities for methodological expansion presented by incorporating computational approaches into mainstream HL.

Historical linguists typically engage in three different tasks simultaneously: (a) the identification of cognates and sound correspondences in languages; (b) the inference of subgrouping (networks/trees);[1] and (c) the inference of sounds/forms/patterns in proto-languages (Ancestral State Reconstruction = ASR). In conventional approaches in HL, 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. Figure **??** illustrates these three tasks for four different kinds of material: words (sounds and 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 genomic data 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. The same is true of biological traits, where biologists take care to avoid predicting impossible ancestral states (**?**) and may therefore revise their trees if these occur. Furthermore, both linguists and biologists 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 (task c; cf. re-estimating sequence alignment in genetics while estimating trees; **?**). The three tasks are not 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 from the historical linguistics literature that linking these three tasks

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

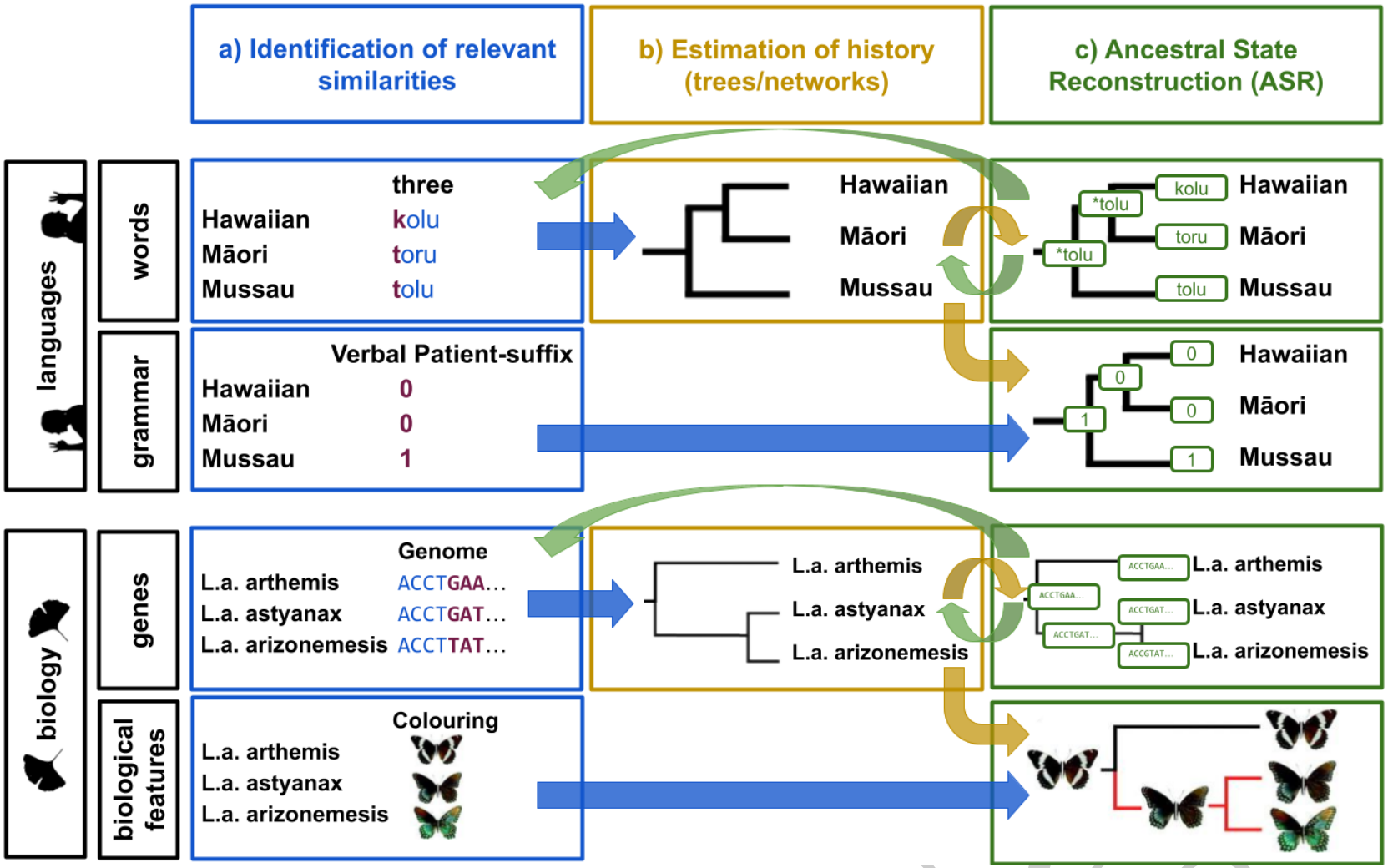


Figure 1: The three tasks involved in the historical reconstruction of linguistic matter (words & sounds), patterns (grammar) and biological traits. The tasks follow each other as indicated by the arrows. The butterfly illustrations are modified from **?**

to each other is useful – for example, 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, 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 only the third task of ASR. Specifically, this paper concerns ASR of structural data, grammatical variables of Oceanic languages from a large-scale typological database (Grambank v1.0; **?**). In addition to increasing transparency, quantitative approaches to ASR also have the benefit of speed. By interrogating the conventional methods of ASR in HL and comparing the principles and outcomes of such conventional methods to various computational approaches, it is possible to evaluate the levels of agreement between the two. In so doing, it becomes possible to improve on transparency and include into historical linguists convenient tools that can effectivize part of the labour.

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 satisfy the Double Cognacy Criterion (**?**), i.e. that the cognate sounds must occur in words which are themselves also cognates. This is relevant both for the construction of trees (task b in Figure **??**) and ASR (task c). It is difficult to apply this test to non-vocabulary data because it is not clear what what corresponds to words and phonemes are in structural data such that this criteria can be satisfied.

In cultural evolution and biology on the other hand, data are deemed appropriate for historical analysis if homoplasy can be excluded (independent convergent evolution). Excluding homoplasy means is taken to indicate that it is reasonable to assume that the tree in question estimates the history of the data and analysis can proceed (cf. **?**; **?**). One of the most common approaches to test if data are 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 see in §**??**.

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 completely 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 consistent, even over large amounts of data. Computational approaches are not intended to replace traditional historical linguistics, but rather to function as a complement, streamlining and making more transparent parts of the process. In this paper, we compare the two approaches conceptually and examine how often com-

putational methods of ASR arrive at the same conclusions as traditional historical linguistics. I will also investigate what the computational methods say when historical linguists disagree, and make new predictions about the grammar of proto-languages.

The case study used in this paper is the Oceanic subgroup of the Austronesian language family and the grammatical features of four of its proto-languages. I use information about the extant daughter languages from the Grambank dataset (**?**) to infer the structure of proto-languages given three trees: (1) Glottolog 4.5 (**?**); (2) **?** - Maximum Clade Credibility Tree (MCCT); and (3) **?** - averages over a sample of the posteriors (a random 100 trees out of 4,200, see more in §**??**).

Findings from the HL literature have been translated into datapoints 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 datapoints 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 traditional historical linguistics.

The results are evaluated in terms of concordance between each method and the predictions from traditional 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 §**??**, §**??** and §**??**). Both traditional methods of ASR in HL and the particular computational approaches in this paper have advantages and disadvantages. Much of the conceptual infrastructure is similar, and for this reason 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 systems of Proto-Polynesian and Proto-Central Pacific. This issue will be investigated and evaluated separately from the overall results of how much agreement there is between traditional HL and computational approaches.

Finally, this study also yields predictions about grammatical features of the four proto-languages that were not addressed by the HL 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 Ancestral State Reconstruction 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 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 cognates. For example, **?**, **?** and many other scholars have analysed Māori [maor1246] /toru/ (meaning 'three') as deriving from the same word as Hawaiian [hawa1245] /kolu/ ('three'). These two words are thus cognates, and this information can then be used to reconstruct a form for Proto-Polynesian. Furthermore, many words that mean the same/similar thing in Māori and Hawaiian show this pattern of /t ∼ k/ correspondence, e.g., Māori /mate/ ∼ Hawaiian /make/ 'to be dead' and Māori /whitu/ ∼ Hawaiian /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 Hawaiian.[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 (i.e., the sound) and the context it occurs in (i.e., the word) need to be cognate in order to form valid data for ASR and subgrouping in conventional historical linguistics. In the above example, the sounds /t/ and /k/ are sound correspondences (sound-cognates) of each other, as are the words /toru/ and /kolu/. There is cognacy at two levels, both of the part of the word (the sound) and the word itself.

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 (cf. **?**) 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 HL that we will be comparing the computational results to (§**??**). In this paper, we will compare methods of ASR in HL 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 (task b in Figure **??**). The estimation of historical relationships between languages in traditional HL is mainly/only focused on the structure of the tree (what is subgrouped with what) and not the length of

[3]Further research into more Austronesian languages shows that Hawaiian /k/ is more likely to be an innovation and Māori /t/ a retention from an older proto-language (cf. 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.

branches between nodes (time). As one anonymous reviewer of this paper noted, branch length estimation is not a goal of the Comparative Method at all. This is important because, as we will see, trees without branch lengths are implausible models of history and are suboptimal for ASR-approaches that take branch lengths into account (e.g. Maximum Liklihood, Stochastic Character Mapping etc). There is some work in traditional historical linguistics to establish branch lengths through relative chronology of changes and archaeological anchor points (e.g., dating of clay tablets, texts, etc.; cf. **?**), but this work is mainly restricted to the Indo-European family. Estimating branch lengths need not be the same as precise chronological dating. For newer approaches to 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 with newer methods 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). All this is to say that the future of ASR in HL will most likely need to include some estimation of branch lengths as this allows for more sophisticated methods, and possibly also better incorporation of plausibility from conventional HL-ASR.

The processes of suggesting subgroupings and ASR are done in tandem in HL; they are estimated simultaneously (cf. **?**: 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 the proto-language's words and sounds (ASR) is necessary. In order to do ASR, some of the tree structure needs to be approximated. Thankfully, this feedback loop between ASR and subgrouping 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.

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 (cf. **?**; **?**).

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 outlines three general principles for ASR in traditional HL 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)

[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.

(ii) the plausibility of the changes posited

(iii) the plausibility of the reconstructed language as a human language (i.e., the degree to which the reconstructed traits work 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" (MP; **?**). 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. Figure **??** 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.

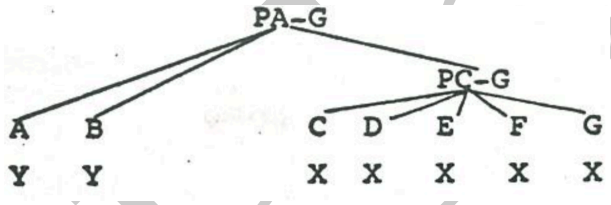


Figure 2: Tree from **?**: 19 illustrating Maximum Parsimony

It is important to note that Maximum Parsimony (MP) 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. More on this in §**??**.

The next principle of ASR in HL concerns the plausibility of changes: phonological, semantic or grammatical. For example in phonology, many historical linguists posit that /s/ is more likely to become /h/ than it is to become /k/,[6] and this information is taken into account when doing ASR. In semantics, in the earlier example from Māori and Hawai'ian, the words /toru/ and /kolu/ both mean 'three', but it is possible for cognates to have less similar meanings. For example, **?** reconstructs the form *panua as meaning 'land' or 'inhabited territory' for Proto-Oceanic. 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. The sound

[6]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.

correspondences can be guiding here. If two forms have somewhat different meanings but convincing sound correspondences, then they may still be cognates. This process can be difficult, as **?**: 229 explains: "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 (cf. **?**: 8) – it seems like the language has skipped over a necessary step, jumping from 'many' directly to 'three' without first encoding a 'two' category. Grammaticalisation theories have given rise to a number of these plausible historical changes (**?**: 594–595, 598).

Lastly, ASR in historical linguistics deals with the plausibility of the whole of the reconstructed proto-language as a system (**?**: 1). 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; cf. **?**). Likewise, 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 structures where we have more worked-out theories of plausible combinations than in the lexicon. This principle has parallels in biology as well, where researchers avoid impossible ancestral states (cf. **?**).

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 stricter applications 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 §**??**).[7]

2.1.1 Disagreements in historical linguistics

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

Plausibility is important for ASR, both in linguistics, studies of cultural change, and biology. However, this principle is sensitive to differing assumptions and theories. Besides debates over precise subgroupings, 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. **?**, **?** and **?** argue for a reconstruction that is technically

[7] It is notable that most of the discussion regarding whether structural ASR is possible has been in relation to Indo-European languages (cf. **?**). This paper offers a view from the Pacific Ocean, where the waves of the debate are more peaceful.

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. Figure **??** shows the Polynesian tree with Grambank feature GB409 values marked out.[9]

I have summarised **?**'s critique of **?**'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 explain 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 changes themselves are less plausible

In a review of **?**, **?**: 539 writes:

> Such an approach [as **?**'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 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.

This debate contains more twists and turns. In our analysis, we will be using trees that represent the history of the languages in a similar way to **?**, which means the results are sensitive to the same critique by **?** (i.e., not taking into account contact

[8] **?**, **?** and **?** 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*.

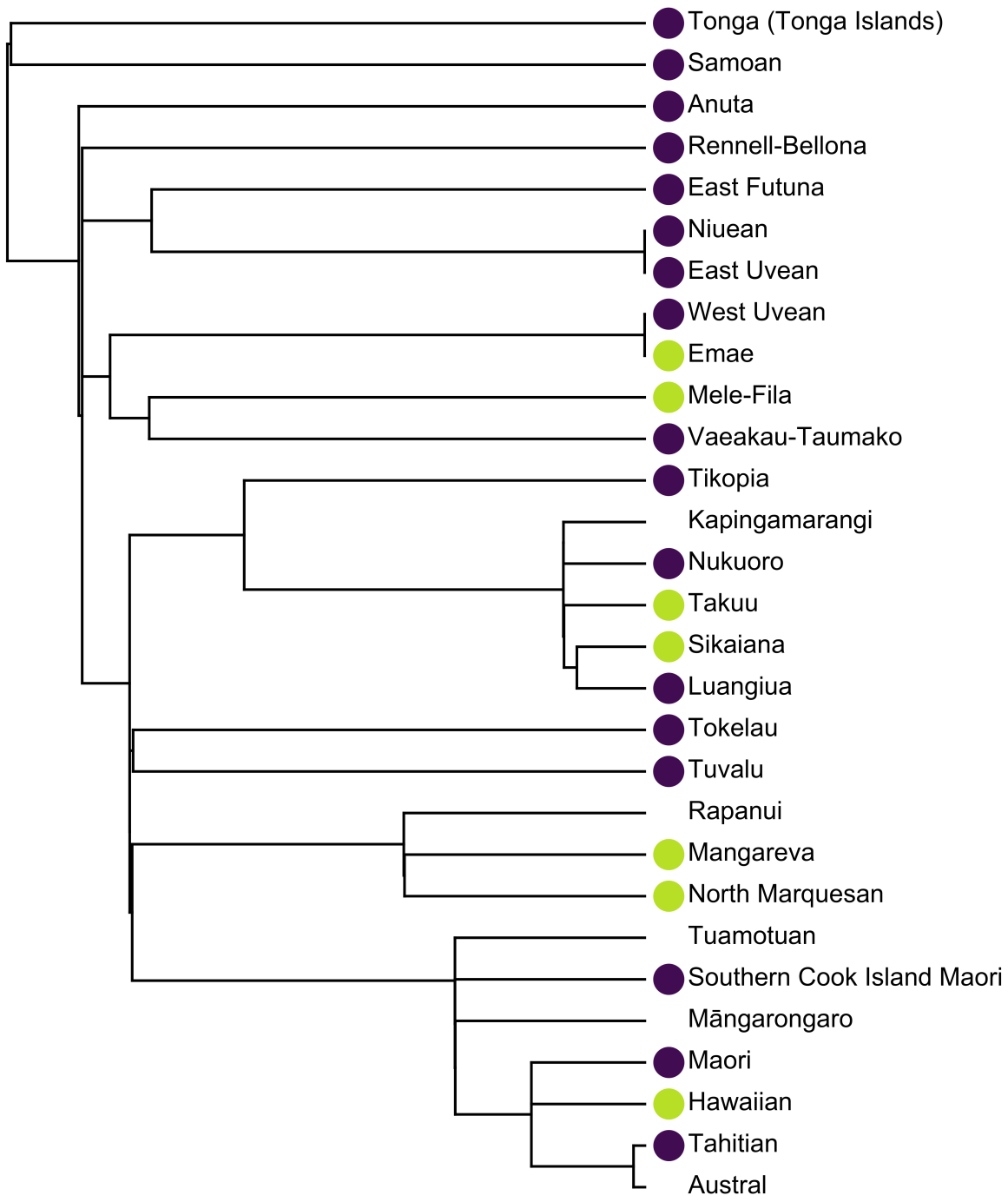


Figure 3: The Polynesian languages in the **?** Maximum Clade Credibility Tree, with the coding of Grambank feature GB409 "Is there any ergative alignment of flagging?" marked out. Purple = Yes; Green = No; absence of dot = Not enough information/not clear

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.

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 are valid for phylogenetic analysis: the Double Cognacy Criterion and phylogenetic signal

There is considerable debate within historical linguistics regarding whether patterns (grammar) can indeed be analysed with the Comparative Method at all. One of 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 are suitable for ASR.

We take a leaf out of the books of cultural evolution and biology in terms of evaluating if the data are appropriate for phylogenetic analysis. In these fields, the three tasks outlined earlier are separated out; appropriate data for analysis are 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 (cf. **?**; **?**). There is a veritable smörgåsbord of methods that a scientist can choose to apply to each task. For example, a plain distances-based approach to making a tree can be used (**?**) or 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 in view of other data besides what it was directly based on, we may be able to

[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*.

carry out analysis with that tree on different data. For example, **?** analyse 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, it is also possible to test the strength of the phylogenetic signal statistically and use this as a guide. If the data have a reasonable phylogenetic signal, we assume that it is likely that the data were generated by the tree and that we can proceed with further analysis. This is what **?** do for their data, and they find that it is possible to carry out the analysis. This can also be done 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 re-arranged (**?**: 905).[11] There exist several different tests of phylogenetic signal: **?**'s λ, **?**'s K, **?**'s δ and **?**'s α, among others. 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 (**?**) 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 were: (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 are similar to what they would be if they were 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 are dissimilar from 0 (Brownian) and 1 (random).

In this study, we are primarily concerned with 84 unique Grambank features[13] and three trees: (1) Glottolog 4.5 (**?**); (2) **?**'s Maximum Clade Credibility Tree (MCCT); and (3) an aggregate of 100 random trees in the **?** posterior. We carry 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 whether they are Brownian or clumped (pval0 > 0.05). The fourth column shows the number of features for which it is 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

[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 Fritz and Purvis (**?**: 1044) for illustrative figures.

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

is because there is 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" feature (i.e., one which rarely evolves), but we cannot derive a measurement of phylogenetic signal per se. Lastly, some features are excluded because there is too much missing data which causes 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 close to 0, meaning that they have phylogenetic signal. There are, however, many features that are not close 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 linguists, been interested in inferring trees of the genetic relationships 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 focussed more on rigourous tests for which linguistic data are valid for analysis (see for example the Double Cognacy Criteria in **?**). It is possible that this is due to the material under

[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.

[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 **?**'s publication of *The Origin of Species* in **?** that the concept of species trees in biology truly took off.

study; it may be more difficult to tease out non-trivial similarities in languages than in species (especially since genome sequencing has become more readily available).

Applying computational methods of ASR to linguistic data is becoming more common. Jäger and List (**?**) apply three different methods (MP, ML and Minimal Lateral Networks) to cognate class reconstruction in three different language families (Chinese, Austronesian and Indo-European) . The aim of that study is primarily to evaluate how often the methods reconstruct 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 HL. The data that serve as input to the computational machinery in Jäger and List (**?**) are annotated "by hand" for cognacy by historical linguists, meaning that the identification of cognate classes is still an entirely human affair (task a in Figure **??**). This is also true for this study – the identification of structural features in languages is a human process. The overall result of Jäger and List (**?**) was that ML performed the most similarly to traditional HL 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 Jäger and List (**?**) also do), some of which may represent an alternative contact history (see §**??**).

There are also two recent studies of Indo-European grammatical history: Carling and Cathcart (**?**) and **?**. Carling and Cathcart (**?**) 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" (**?**: 3). 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. **?** argues that the appropriate action is instead to carry out reconstruction on many different trees that represent different possible histories – a Bayesian posterior tree sample. He argues that this is methodologically more sound and because the results of his specific case-study are in accordance with the consensus in HL it strengthens their validity.

Both Carling and Cathcart (**?**) 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

[16]The main difference between the methods of Carling and Cathcart (**?**) and **?** is that Carling and Cathcart (**?**) 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 **?**.

> 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 (**?**: 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 [emphasis in original]. Finally, rates of gain and loss are assumed not to vary across the tree.

It is always important to be 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, see **?**, **?**, **?** and **?**.

Another popular method of ASR 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 Carling and Cathcart (**?**) and **?**, but not necessarily.[17]

Computational approaches to reconstruction not only allow us to streamline 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 judgements 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.

3 Materials 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 three methods: Maximum Parsimony, Maximum Likelihood and Most Common. This section gives a brief overview of the three methods. Further technical details concerning their precise application can be found in Supplementary Material **??**. For an extensive comparison of different methods of ASR and their advantages, see **?**.

Maximum Parsimony (MP) finds the set of ancestral states that results 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 MP selects

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

the candidate solution with the slowest rate out of all possible solutions it can choose from. MP is intuitively simple.[18]

MP can be critiqued on the basis that it does not take into account branch lengths in the tree (the time between splitting events). Furthermore, MP 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 MP predicts. Both of these disadvantages are addressed in the second method we will be applying: Maximum Likelihood (ML).

ASR using ML 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. ML attempts to find the most mathematically *likely* solution; MP finds the solution with the slowest rate.[19] If, for example, the distribution of values at the tips is very scattered, with sibling pairs frequently having different feature values, ML will infer that the feature has a high rate of change and use that information "backwards" when positing ancestral states as well. The ML 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 had occurred along a longer branch.

Reconstruction using ML 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.[20,21] Specifically, we are also using a marginal ML estimation – for more details see Supplementary Material **??**.

It is impossible for MP 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 for historical linguists to estimate something similar by taking into account the length of time and the "plausibility of the changes posited", including whether losing a certain feature is more likely than gaining it. In this study, we compare MP and ML reconstructions with conventional ASR in HL. If the results from conventional ASR are more similar to that of ML, a potential explanation would be that the "plausibility of changes posited" is indeed operating along similar lines as ML 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"

[18]While the principle of MP 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".

[19]It should be noted that my use here of "rate of change" in relation to MP (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".

[20]Similarly to the studies by Carling and Cathcart (**?**) and **?**, rates cannot vary within the tree in this study.

[21]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.

which is based solely on which value is the Most Common (MC) in the daughter languages of a given proto-language, entirely disregarding the tree structure.[22] In the toy example in Figure **??**, this approach would reconstruct that the root had feature value "X". Whether we prefer MP, ML or another approach to reconstruction, actually taking the tree structure into account is generally sounder methodology.

All of the R code[23] 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.[24] 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 (qualified) majority presence (i.e., more than 60% of the ancestral state is "1"), it is registered in the results as "Presence". If the distribution is 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 is done to highlight the amount of uncertainty the results sometimes contain, while at the same time making it a fair comparison between MP and ML. Comparing the raw distributions themselves is not a fair comparison because MP is always more likely to suggest 0, 0.5 or 1 results (because the majority of the splits in the tree are binary), whereas ML rarely produces exactly 0 or 1.

Rounding into these bins ("Presence", "Absence" and "Half/Half") 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 is "Presence" and the algorithm predicts presence with over 60%, it is counted as a "True Positive", and so on.[25] Table **??** illustrates how the

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

[23] All analyses have been calculated in `R` (**?**) using the packages `castor` (MP; **?**), `phangorn` (MP; **?**) and `corHMM` (ML; **?**). The packages `ape` (**?**), `adephylo` (**?**), `phytools` (**?**), `psych` (**?**), `reshape2` (**?**) and `tidyverse` (**?**) were also used for data wrangling, analysis, summarising and visualising. For a complete record of all R-packages used, see Supplementary Material **??**.

[24] 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.

[25] 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"

results are summarised.

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

Finding by Conventional Methods	**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 (Equation (**??**)) is then calculated. The score is calculated between all computational methods and the conventional historical linguists' prediction, as well as between all computational methods themselves.[26]

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

It is also important to take into account the "Half/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 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 Equation (**??**).

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

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, Jäger and List (**?**) compared three different methods of ASR for lexical data (cognate classes): MP, ML and Minimal Lateral Networks. They found that reconstructions using ML performed the most like the predictions by historical linguists. However, Jäger and List (**?**) describe the general performance of all the computational reconstruction methods they used as "poor".

necessarily.

[26] When comparing one computational method result to another, "Half/Half" – "Half/Half" count as a True pair. Otherwise the scoring is the same.

Jäger and List (**?**) evaluates the methods using F1-scores which are the harmonic means of Precision and Recall (**?**). This way of evaluating performance focusses on True Positives and ignores True Negatives altogether. It was suitable for the study by Jäger and List (**?**) 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 **??**).[27]

In addition, we also test 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 of 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 they are easily codable for the majority of the world's languages which have been described grammatically (approx 4,000 languages; 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 students, 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 §**??**.

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

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. Figure **??** from **?**: 2 shows the geographic 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.

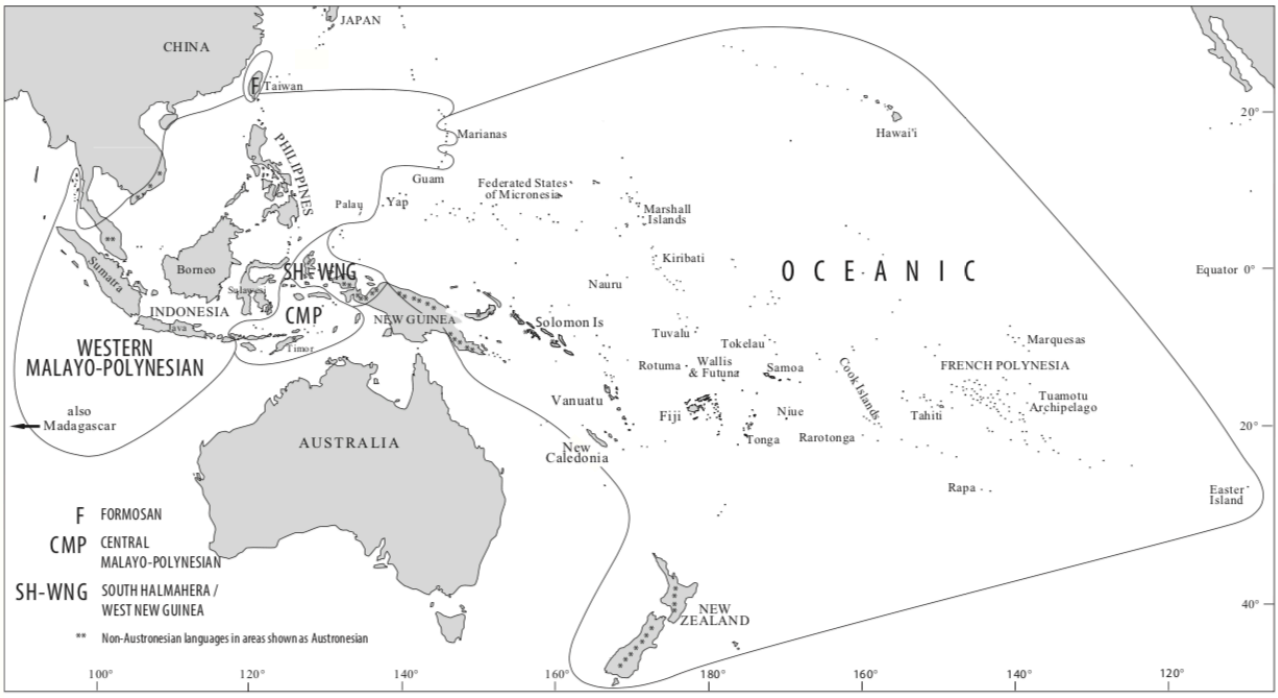


Figure 4: Map of the Austronesian language family and major subgroups from **?**: 2

Not all languages of the Oceanic subgroup have grammatical descriptions, but of those that have one, 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 Figure **??** shows the same coverage information, with languages coded for their data coverage status.

The coverage of Grambank data for the Oceanic subgroup is generally 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 (cf. **?**).

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.

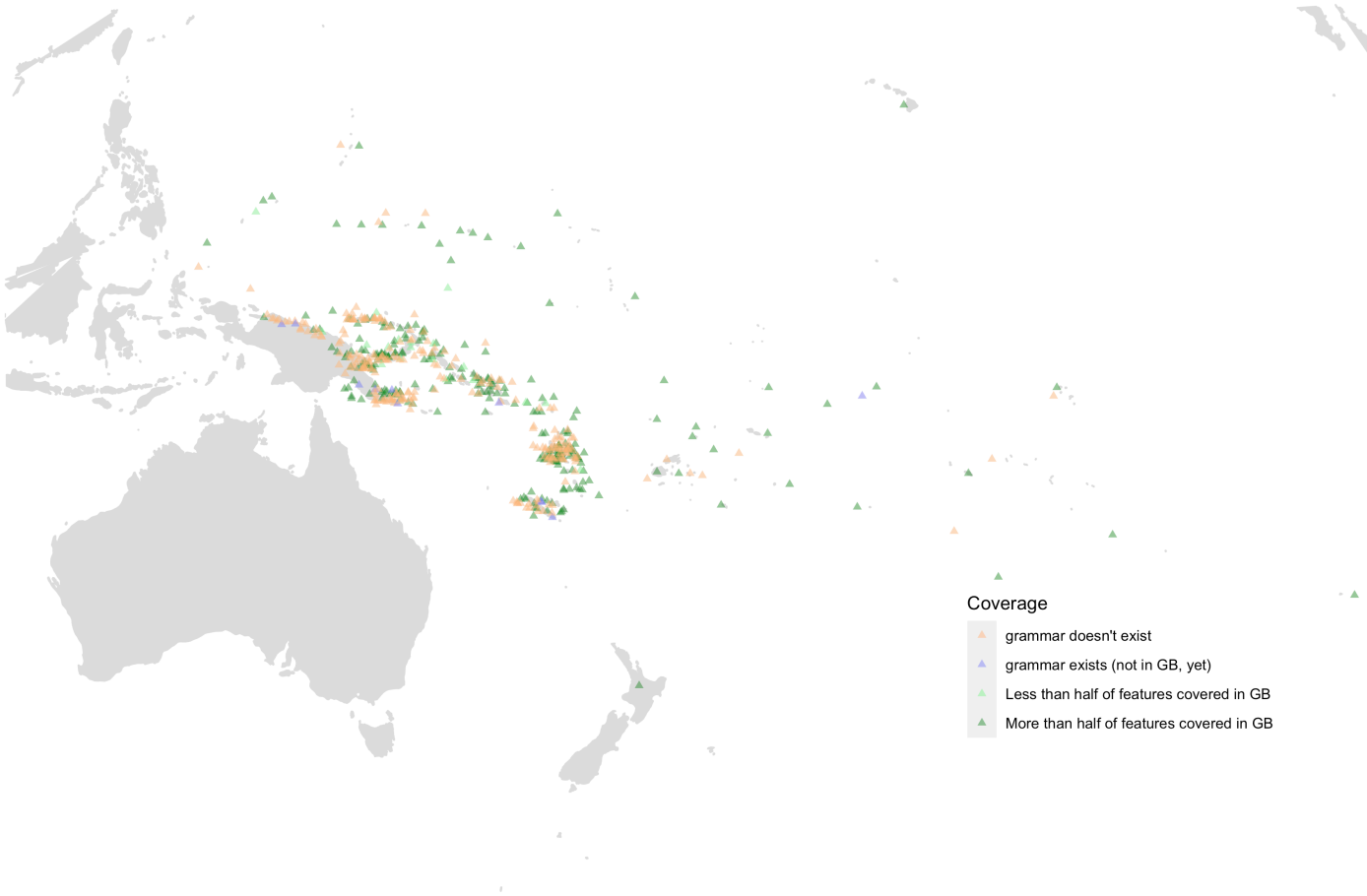


Figure 5: Map of Oceania, with Oceanic languages coloured for their coverage in Grambank

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 (**?**)[28]

Figure **??** and Figure **??** show the Grambank coverage of languages over the phylogenies from the **?** 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, whereas the branches in the **?** trees (both 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. 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 have passed between the root and the languages at the tips.

In addition, the Glottolog tree contains more non-binary splits (polytomies) than the **?** trees. Binary splits ought to be more plausible, since it is unlikely that a set of three 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, and 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 estimate contact events as well as inheritance. Figure **??** shows a DensiTree visualisation (**?**) of the 100 trees which are used in this study.

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

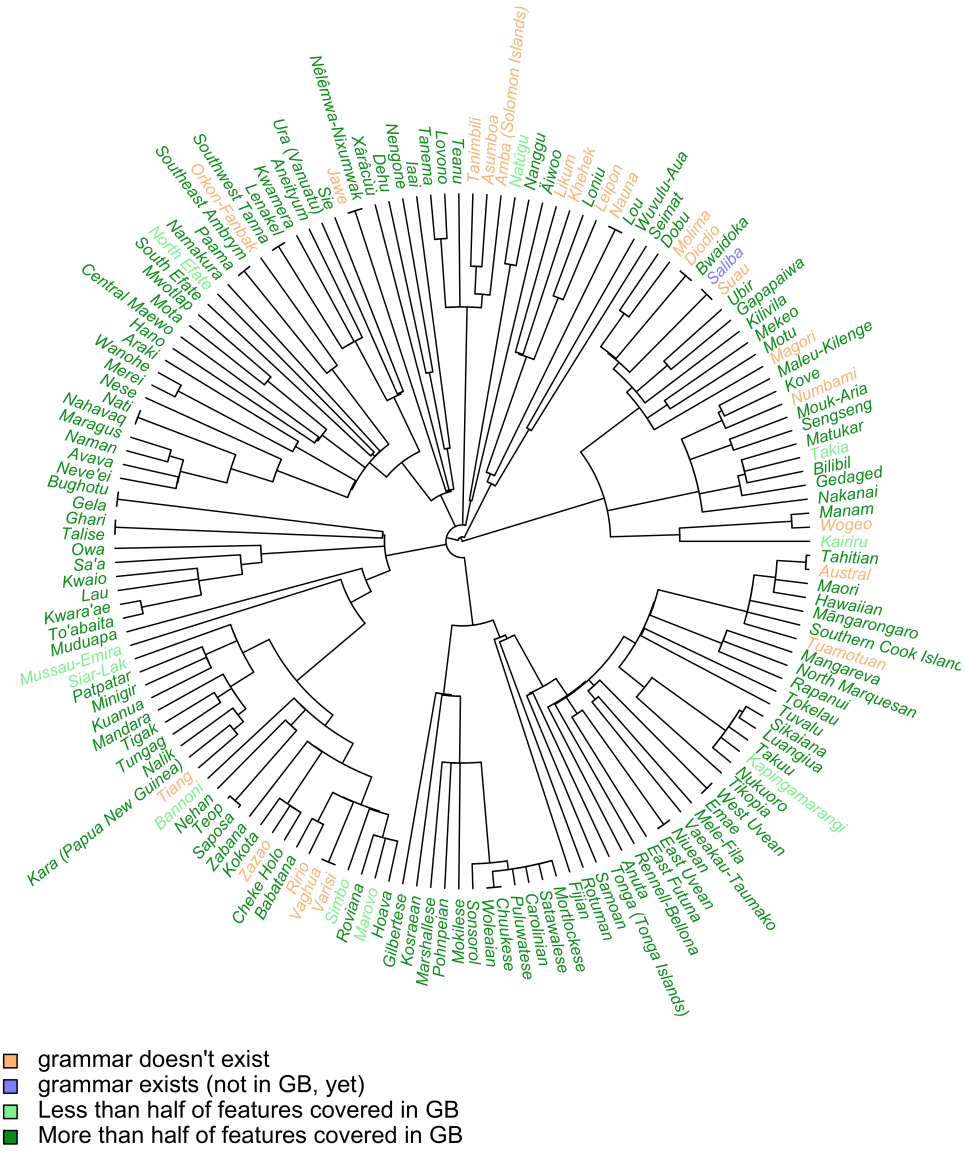


Figure 6: Maximum Clade Credibility Tree of Oceanic from **?**, with languages coloured for coverage in Grambank

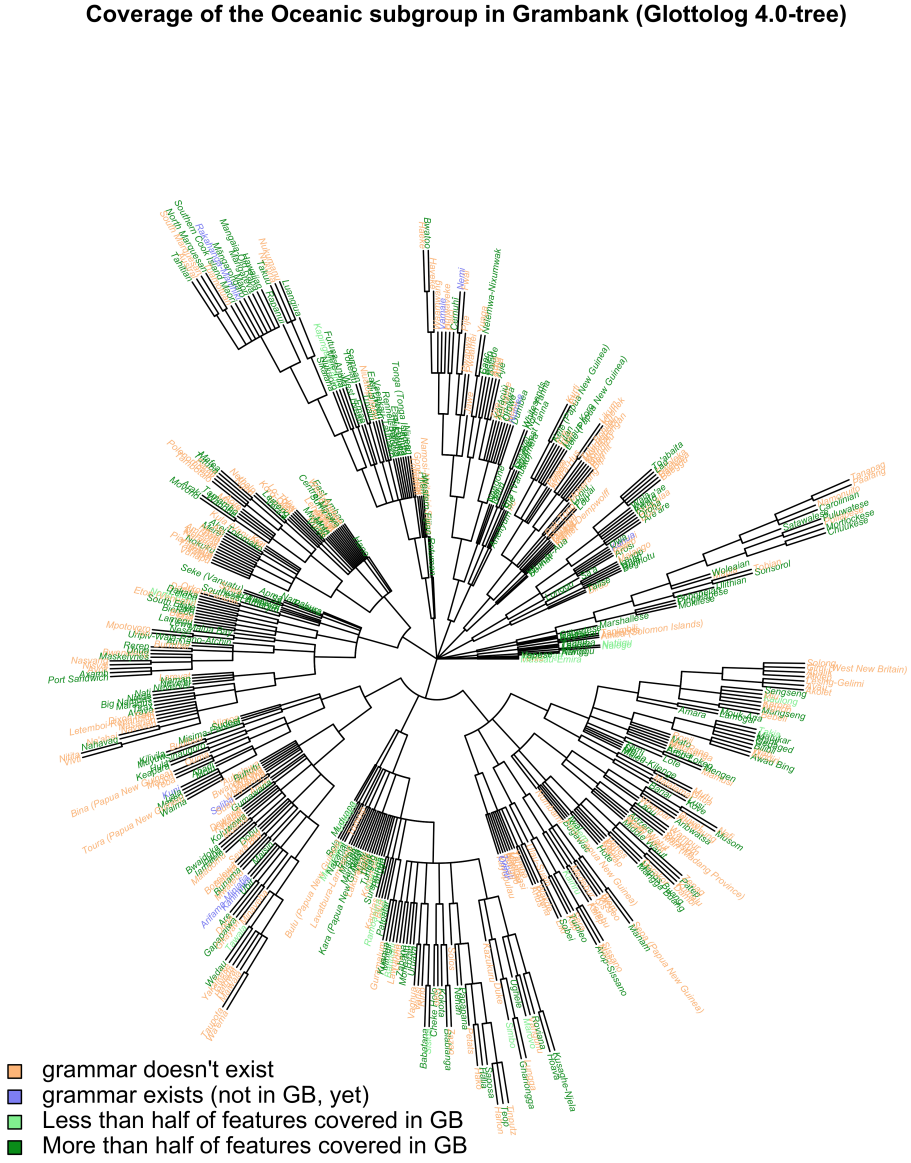


Figure 7: Tree of Oceanic from Glottolog, with languages coloured for coverage in Grambank

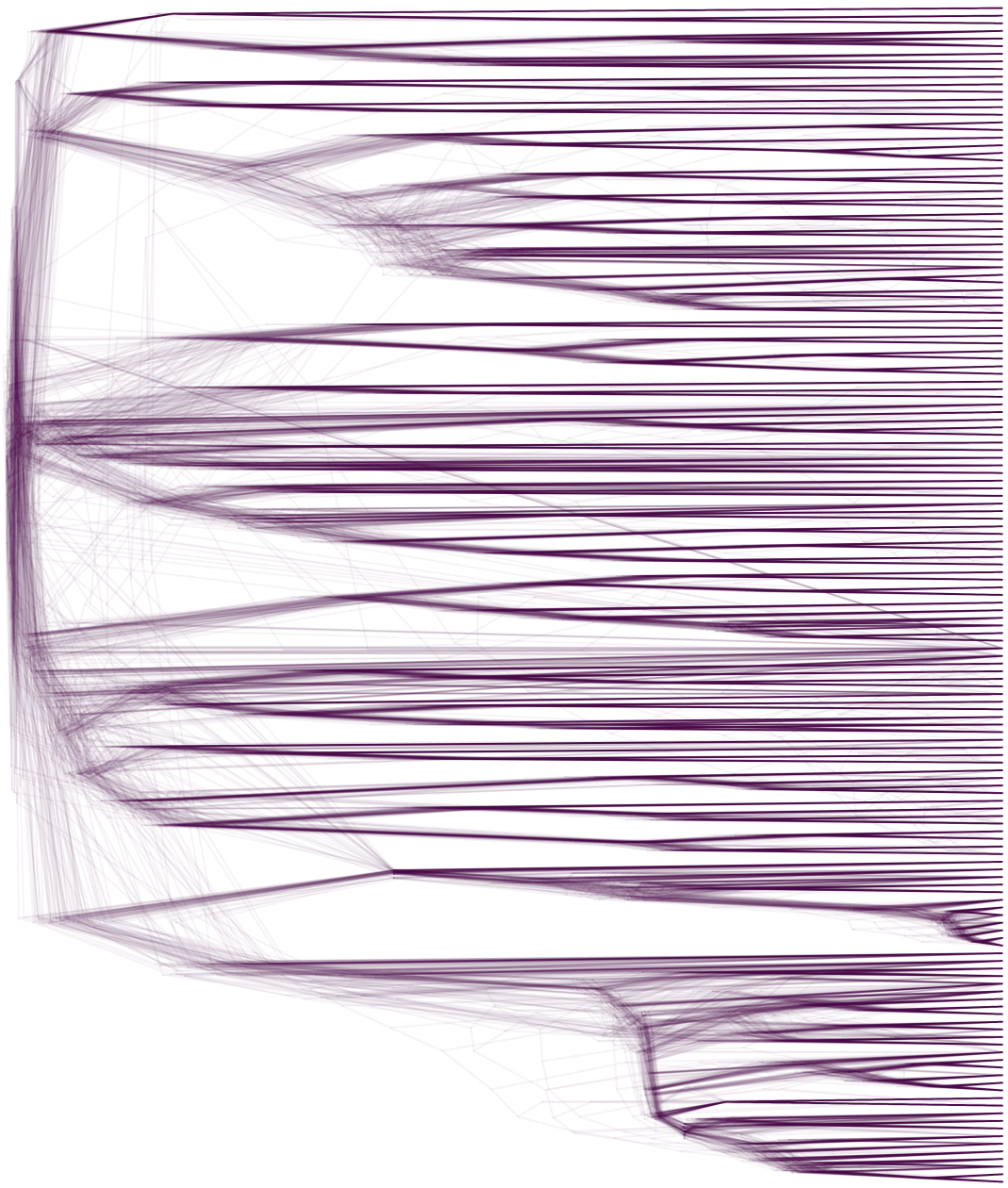


Figure 8: DensiTree (**?**) visualisation of the 100 random sampled trees from the **?** posterior. Made with the function `densiTree()` from the R package `phangorn` (**?**)

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, among other publications, the book series on the Proto-Oceanic lexicon; **?????**). There also exists substantial work done on the grammar of Proto-Oceanic using conventional methods in HL. 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.[29]

For each of these publications on the grammars of proto-languages, findings have been extracted that support a certain coding in the Grambank questionnaire for each proto-language. For example, **?**: 4 writes that a causative prefix **faka-* can be reconstructed for Proto-Polynesian. In the Grambank questionnaire, we have the feature GB155 "Are causatives formed by affixes or clitics on verbs?". So, in Proto-Polynesian for the feature 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 HL 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.), where we read grammars that describe particular forms, paradigms, etc. and then translate this information into Grambank datapoints.

When doing ASR in conventional HL, 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 (**?**); (**?**); (**?**); (**?**); (**?**). 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"[30] in Oceanic and finds that in many languages there is a reflex of what is taken to be proto-Oceanic **na/*a*, but in some languages there is another marker with a different origin (Māori *te*, for example). In **?**'s study, languages where there is no common noun phrase marking whatsoever and those with a marker that is not cognate with **na/*a*, are both included in Type 1 (see Figure **??**). These languages are contrasted with those

[29] Table **??** in Supplementary Material **??** lists all of the publications used here as representations for the reconstruction of grammar in Oceanic linguists by conventional historical linguistics means.

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

that have retained some kind of reflex of **na/*a* (Types 2–4 in Figure **??**). 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.

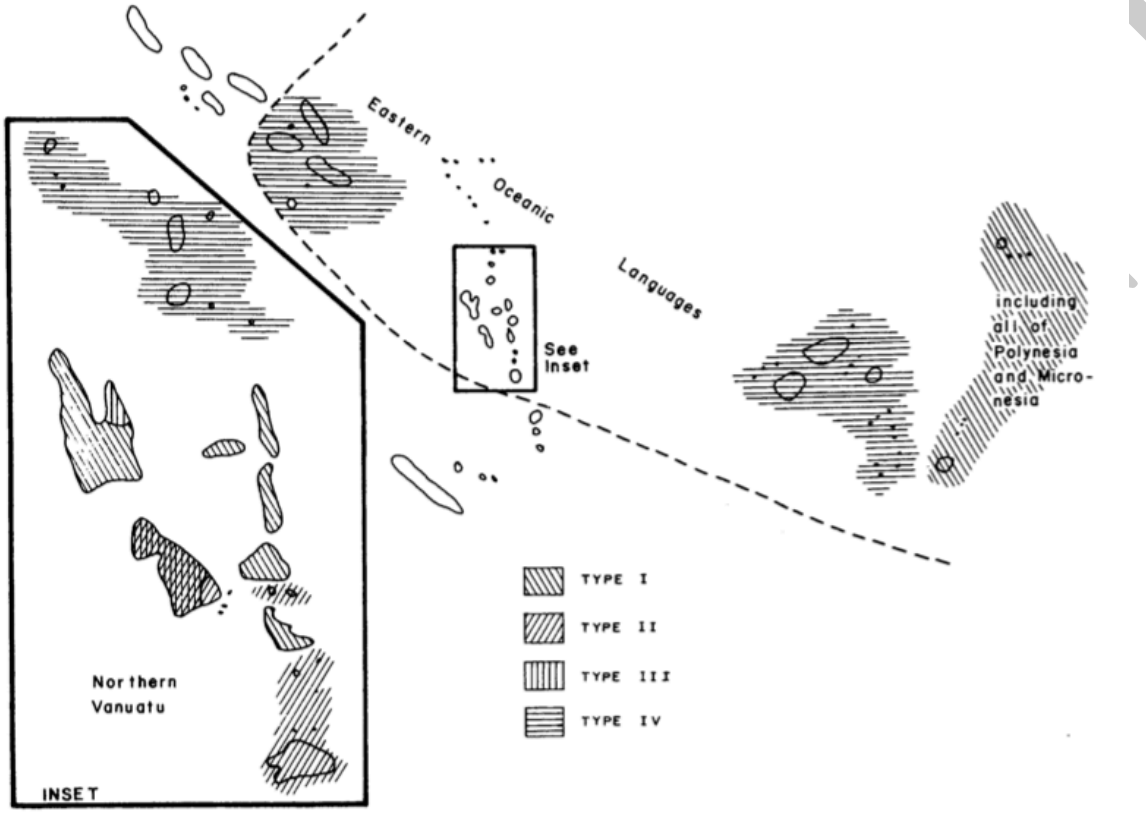


Figure 9: Map of four different types of common noun phrase markers in Eastern Oceanic from **?**: 162. Type 1: absence of common noun phrase marker or marker is not a reflex of **na/*a*; Type 2: non-productive system involving a reflex of **na/*a*; Type 3: productive marking involving **na /*a* as a prefix that is regularly separable from the noun; Type 4: productive marking involving **na/*a* generally existing as a free-standing marker. Areas with cross-hatching show a distribution of both Type 1 and Type 2 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 Figure **??**). 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 therefore distinguish those that have a prenominal article from those that do not, but we cannot tell apart those which have retained the proto-form **na/*a* and those which have not.

This is a difference in the kind of data that goes into the analysis, not a difference

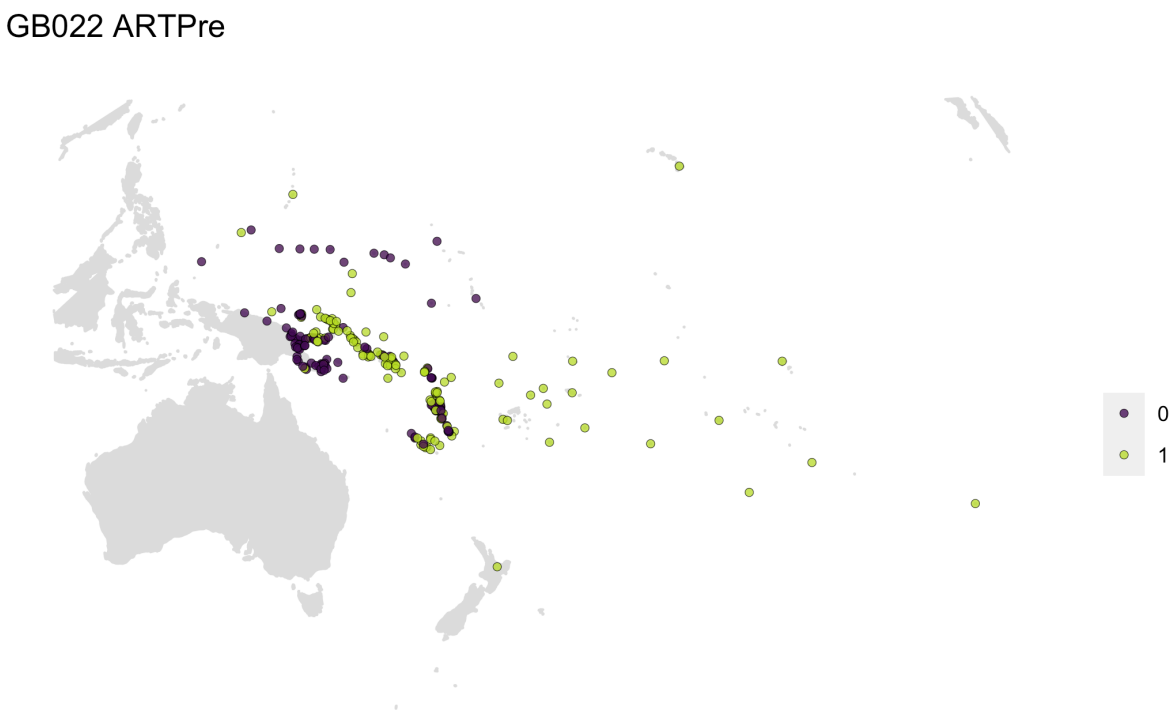


Figure 10: Map of Austronesian languages for GB022 “Are there prenominal articles?”. Yellow = “yes”; Purple = “no”

in the analytical methods themselves (compare with tasks a and c in Figure **??**, respectively). While this principal difference is important, it should also be noted that Grambank feature GB022 has strong phylogenetic signal (negative D-estimates which are statistically similar to 0; see §**??** for details). This gives us confidence that we can move ahead.

As we have seen, Grambank data are 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 HL, which tend to be more focused 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.[31] 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 HL. The average completion per language in the whole of the Grambank dataset 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; **?**: 173); a distinction between alienable and inalienable possession[32] (**?**: 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 Table **??** in Supplementary Material **??**.

[31] **?** also notes that **?** observes similar patterns in Algonquian languages.

[32] 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.

Most of the time, 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, there are some disagreements (as discussed in §**??**). In total, there are 115 datapoints where there is either just one publication supporting the statement or – if there were several – they agreed. There are three datapoints where there is disagreement (all of which concern 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 (MC). For (a) and (b), we are also using three different trees: (i) Glottolog, (ii) **?** MCC tree and (iii) the mean values of reconstruction of a random selection of 100 (out of 4,200) trees in the Bayesian posterior of **?**. That gives us 7 results, in total.

The results are divided into three subsections: §**??**, concurrence with conventional historical linguistics; §**??**, new predictions; and §**??** 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.[33] 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 for the MC 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.

Given these counts, we can calculate the concordance scores (see §**??**). These are displayed in Figure **??**. 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 include half-results for each method are more similar to each other.

The method that performs most similarly to historical linguists is MP * the Glottolog 4.5 tree. The Glottolog 4.5 tree has a significant issue; it has no branch lengths

[33] There was one feature for the ML analysis of the **?** 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 * **?** trees.

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.

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: 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 MP method (a method that disregards branches altogether).

Overall, however, the methods perform similarly. There is very little that differentiates 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, 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 HL (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 vary more, the agreement is reduced. This is expected; 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 §?? 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) 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. Figure ?? 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 HL. The reason is most likely related to not only the difference in methodology but also the underlying data. All of the computational methods use Grambank data and partially the same trees, whereas the underlying language-level data and tree structure in conventional HL ASR are different. For the purposes of this study, the HL literature has been translated into Grambank datapoints, 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 dataset. Reconstruction in conventional HL 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.

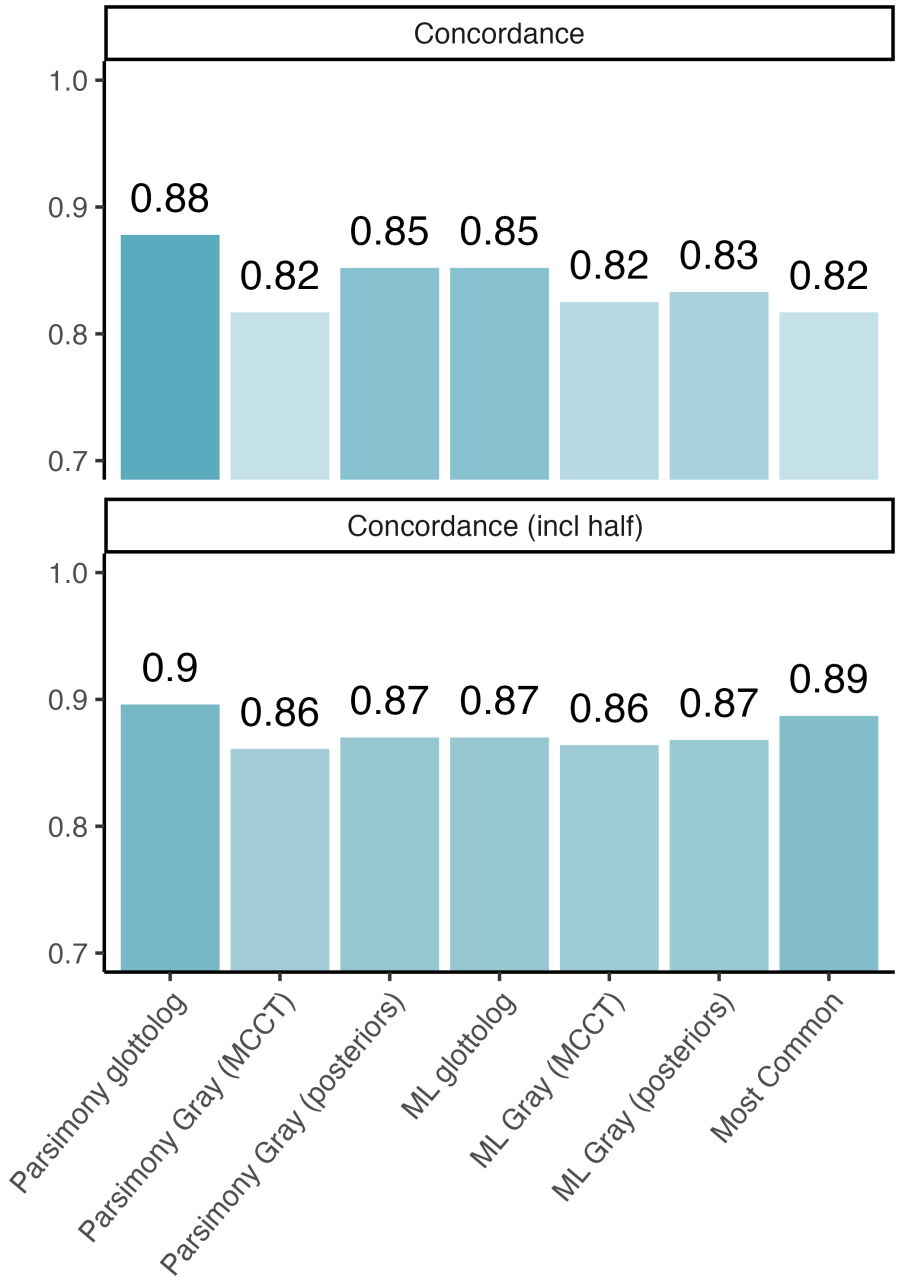


Figure 11: Barplots of concordance scores of each method

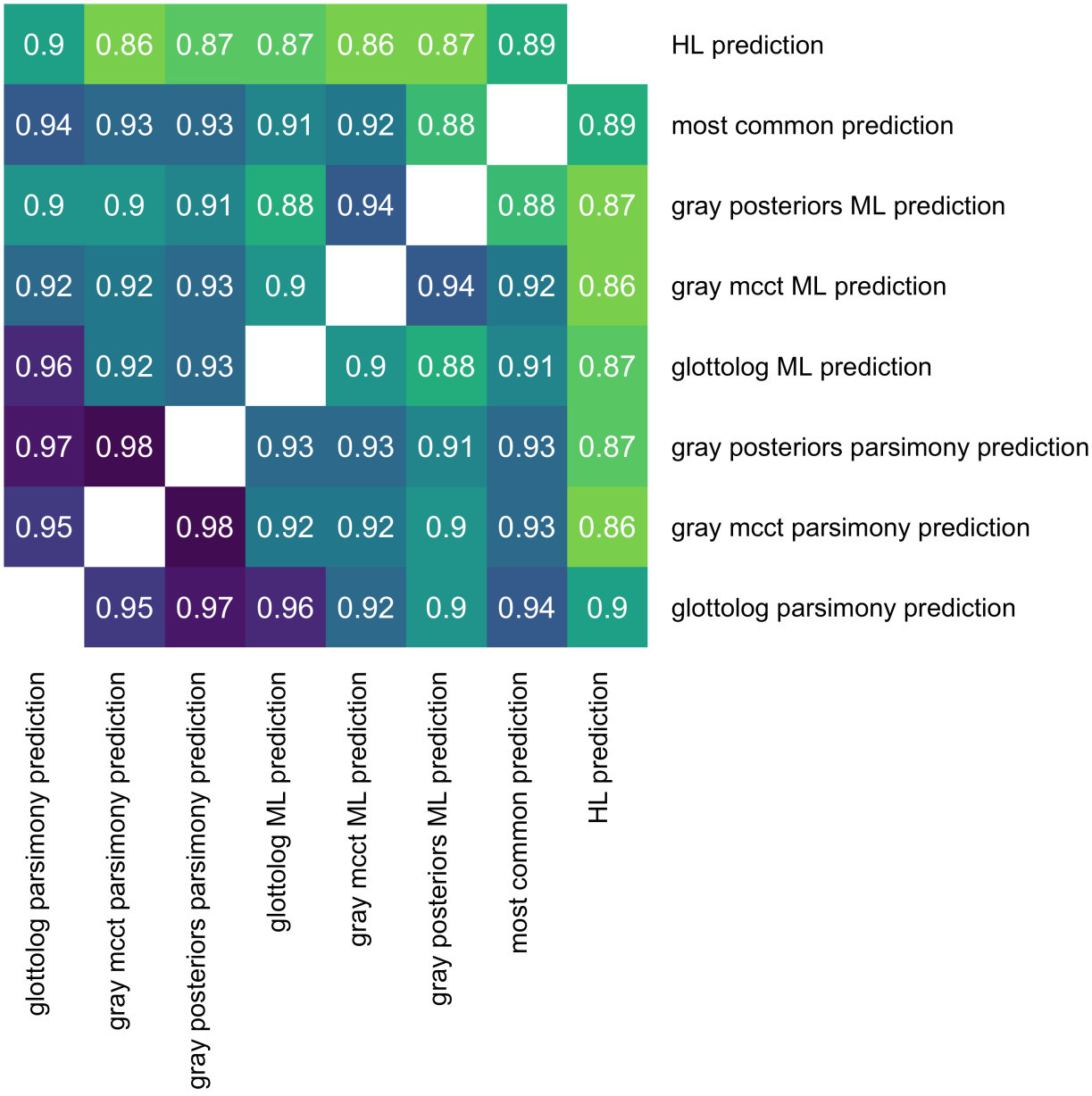


Figure 12: Heatmap of accuracy score (including half) between reconstruction, per tree and method. Dark blue = high concurrence; Light green = low concurrence

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, **?** – MCCT and **?** – posteriors); i.e., six 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 little surprise that all the methods agree. However, there are four 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 §**??**). 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 **?** and **?** "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 **?**. Table **??** shows that MP, ML and MC all reconstruct presence for ergative flagging in Proto-Polynesian. There is disagreement on the matter of nominative-accusative marking, 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 §**??**, the computational reconstructions differ from those arrived at through the conventional ASR in HL primarily because the data used in this study are the more abstract presence or absence of structural features, whereas historical linguists use specific concrete forms instead (cf. **?**). Besides the parsimony principle (as laid out by **?**: 19), 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 **?**'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 (cf. Figure **??**), it is very likely that GB409 would be reconstructed as present for Proto-Polynesian. As **?** points out, this is the most parsimonious solution. However, GB408 (accusative) could still be 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 rely on changes in the specifics of the passive voice construction that we, unfortunately, do not track. Given our data, which simply record 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 **???** and **?**.

Concerning the alignment of Proto-Central Pacific, all the results (save the MC model) predict an absence of ergative-marking. This is likely to be because Rotuman [rotu1241], West 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 without 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 §**??**, §**??** 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 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

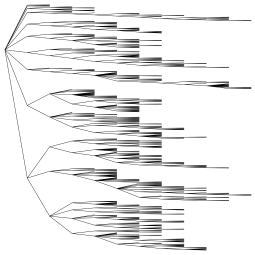
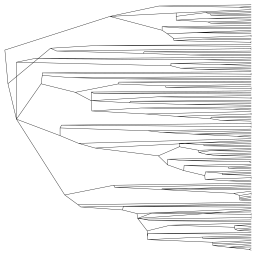
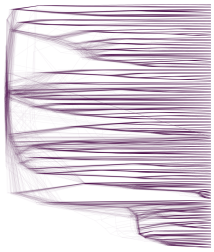
Table 8: Summary of conceptual pros and cons of the ASR-methods

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

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, ML 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 Figure **??** 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 uncertainty in a better way and involves fewer polytomies.

Conceptually, the most reliable results *a priori* are those derived from ML * 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 MP and ML agreed to a very large extent with the findings from HL. This suggests that the mechanisms at work in HL 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, ML * random sample over **?** posterior, did not have the highest concordance (including half-results) score with HL, at 0.87. The variation between the results was not large, however. The method that was the most similar to conventional HL, MP * Glottolog, achieved a concordance of 0.9. The Glottolog tree contained more matches to Grambank datapoints, which is probably why it outperformed the **?** trees in concordance with conventional HL.

The methods which do not take into account branch lengths (MP and MC) achieve a somewhat higher concordance with HL 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 HL as they first may appear.

The classical HL toolkit is well-developed in terms of subgrouping, 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 subgrouping done in HL 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 difficult to estimate rates of change, historical linguists do have knowledge that may rein in analysis, avoiding fantastically slow or fast rates.

Computational methods need not be in conflict with conventional approaches – the two can be complementary. There is certainly room for improvements in computational approaches based on knowledge from classical HL. 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 the future. For example, it is possible to supply ML 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 (cf. **?**), but they are not yet ready to replace the vast human knowledge and experience of the experts in HL. However, once cognate classes, regular sound correspondences and structural features are identified, the work then turns to reconstructing history (subgrouping or constructing trees/networks) and ASR (cf. Figure **??**). For these tasks, there are suitable computational methods that can be applied, such as those in this paper and others (cf. **?**; **?**; **?**). Research into linguistic history can be greatly improved and streamlined by computational tools, which in turn can be given sensible priors and parameters to produce more reliable results in future joint ventures 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 HL studies.

It is no doubt difficult to convey this wealth of contextual information in each and every academic paper. The task becomes more 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.

Furthermore, it is also desirable that computational methodologies and phylogenetics be made more accessible to the wider linguistics community and incorporated into HL education. It is my perception that there is at times a disconnect between newer and classical approaches in this space, which is unnecessary and even 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 HL more generally.

The more methodology and analytical choices are made explicit, the easier it is to assess the soundness of a 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 HL and the corresponding computational approaches. Hopefully, this study (alongside **?** and **?**) can be a starting point for more joint ventures into our cultural past.

Funding information

The research leading to this paper has received funding from the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology, the School of Culture, History and Language at the Australian National University and from the Australian Research Council through The Wellsprings of Linguistic Diversity Laureate Project (awarded to Professor Nicholas Evans 2014-2018) and the Centre of Excellence for the Dynamics of Language (CoEDL).

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, editors and proof-readers 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

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

Address of author

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

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 des Baumes, Plausibilität von Änderungen und Plausibilität der resultierenden Protosprachen. Dieser Ansatz weist einige Probleme auf, insbesondere die Definition von plausibel 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 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 Richtung befassen. Eine Kombination aus computergestützten Methoden und qualitativem Wissen ist künftig möglich und wäre von großem Nutzen.

Résumé La reconstruction de l'état ancestral (ASR) est une partie essentielle de la linguistique historique (HL). L'ASR conventionnel en 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), sur les plans conceptuel et pratique. 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 la HL classique à s'accorder davantage avec les approches qui ignorent les longueurs de branche. La prise en compte explicite des longueurs de branche est plus appropriée du point de vue conceptuel. 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.