

# Dissecting historical reconstruction: comparing computational approaches and the comparative method for Oceanic grammar

November 18, 2022

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Background</b>	<b>7</b>
2.1	The methods of traditional historical linguistics: the Comparative Method . . . . .	7
2.1.1	Disagreements in historical linguistics . . . . .	10
2.2	Characteristics of language structure as data . . . . .	12
<b>3</b>	<b>Material and methods</b>	<b>14</b>
3.1	Computational phylogenetic methods . . . . .	14
3.2	Calculation of similarity between predictions from the Comparative Method and computational approaches . . . . .	16
3.3	Data . . . . .	18
3.3.1	The Grambank-dataset . . . . .	18
3.3.2	Data coverage . . . . .	19
3.3.3	The trees . . . . .	22
3.3.4	Data from historical linguistics on Oceanic proto-language grammar . . . . .	26
<b>4</b>	<b>Results</b>	<b>28</b>
4.1	Concordance between traditional historical linguistics and computational methods . . . . .	28
4.2	New predictions . . . . .	31
4.3	Where the conflicts are: Ergativity . . . . .	31
<b>5</b>	<b>Conclusions</b>	<b>33</b>
<b>A</b>	<b>Appendices</b>	<b>39</b>
A	Data availability . . . . .	39
B	Technical details of methods . . . . .	39
C	Mathematics of the F1-score including half-results . . . . .	39
C.1	Standard definitions . . . . .	40
C.2	Half-result definitions of precision and recall . . . . .	40
C.3	The question . . . . .	40
C.4	The proof . . . . .	40

D	Binarisation of the Grambank features . . . . .	41
E	Grambank features . . . . .	42
F	Further details on the tree phylogeny . . . . .	53
G	Further details on the Grambank coding of proto-languages . . . . .	53
H	Table of historical linguistics sources surveyed . . . . .	54
I	Table of new predictions . . . . .	55

**Abstract**

Reconstruction is an essential part of historical linguistics. The traditional Comparative Method (CM) identifies regular sound correspondences and cognates and then infers states in proto-languages by three core principles: assume the fewest changes on the tree, assume plausible changes, and assume plausible combinations of features in proto-languages. However, there is no clear consensus on how to weigh the principles against each other, and many studies are not fully transparent regarding how precisely the principles are applied. This allows for subjective differences between studies which are hard to discern and replicate. This study aims to better understand the comparative method in general and Oceanic grammatical history in particular by comparing the reconstruction of structural features by classical historical linguists using CM to computational reconstructions with explicit and transparent mechanisms: Maximum Parsimony (MP) and Marginal Maximum Likelihood (ML). MP is similar to CM in that it infers the fewest amounts of changes along the tree. ML on the other hand infers rates of change based on the distribution of values and takes into account branch lengths. In addition, we explore an even simpler method of reconstruction, which ignores tree structure and takes into account only the most common feature in the daughter languages. The results indicate that all methods applied achieve a high level of concordance with predictions from historical linguistics, which lets us understand the mechanisms of CM better. The more methodology is made explicit, the better we are off as a field.

## 1 Introduction

Linguistic<sup>1</sup> history 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 migration paths of people and cultures. By reconstructing the words and grammars of ancient languages, we can learn about communities long gone. Historical linguistics is devoted to this endeavour and has made great strides in our understanding of human history since its inception. The field has established methods which have enabled us to classify languages into language families and reconstruct words and grammars of proto-languages (unobserved ancestors of observed languages). Conclusions from historical linguistics are also impactful outside of linguistics, for example in archaeology and history studies.

Linguists who work on historical reconstructions produce valuable and greatly inspiring work and they possess a wealth of knowledge not only of the languages themselves, but also the cultures, societies and history of the region. However, all of this knowledge is not always transmitted in each publication, which can make for outcomes that are difficult to replicate. One particular problem with the field is that analysis often relies on subjective evaluations of the weight of the evidence at hand and differing assumptions about what is plausible. This makes historical linguistics difficult to replicate. Historical linguists possess a great wealth of knowledge and understanding of a region and its cultures. All of this background knowledge underlying analytic decisions is not always made explicit and formally defined (possibly due to lack of space and necessity). This makes the analysis difficult to evaluate for someone not as steeped in the literature and region.

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 relationship between species, ancestral states and the tempo and mode of evolution (Atkinson & Gray 2005). Interestingly, the use of trees in linguistics and biology first occurred in publications just one year apart with Schlegel (1808) publishing a tree of languages and Lamarck (1809) a tree of species.<sup>2</sup> Biology and linguistics may have inspired each other, but methodologically the fields progressed separately for a long time (Greenhill 2015:370). 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 slowest? etc. The two fields have developed different methodologies, with biologists leaning more towards quantitative computational methods compared to linguists.

Historical linguistics is founded on the Comparative Method, which is a set of principles that can be

<sup>1</sup>I am fortunate to have colleagues in academia who have been generous and helped me with technical matters and working through methodology conceptually, among these are: Stephen Mann, Angela Chira, Xia Hua, Simon Greenhill, Cara Evans, Benedict King and Hannah Haynie. I am also grateful to the two anonymous reviewers at Diachronica for their valuable feedback. Any mistakes and misconceptions that remain are my own.

<sup>2</sup>However, as Greenhill (2015:370) notes, it was not until Darwin's publication of *The Origin of Species* in 1859 that the concept of species trees in biology truly took off.

applied to a set of languages and data to derive sub-groupings and ancestral states (see section 2.1). This approach has yielded great results, such as the inference of many language families and proto-languages, and has robust control for avoiding trivial similarities. However, one of the drawbacks of this approach is that it typically involves manual work and potentially subjective judgements. In contrast, computational phylogenetic methods are a set of tools that can be applied objectively in a systematic fashion. These approaches have the advantage of being able to be applied in exactly the same way across a large set of data and their mechanisms are very explicit. Computational approaches are not intended to replace traditional historical linguistics, but rather function as a complement effectivizing parts of the process. In this paper, we examine how often computational methods of reconstruction arrive at the same conclusions as traditional historical linguists using the Comparative Method. We will also investigate what the computational methods say when historical linguists disagree, and make new predictions about the grammar of proto-languages.

Applying computational methods of ancestral state reconstruction to linguistic data is becoming more common. [Jäger & List \(2018\)](#) apply three different methods (Maximum Parsimony, Maximum Likelihood and Minimal Lateral Networks) to cognate class reconstruction in three different language families. The aim of that study was primarily to evaluate how often the methods reconstructed the same state as what the authors label “the Gold Standard” (reconstructions by traditional historical linguists using the Comparative Method). The data that serves as input to the computational machinery was annotated by “hand” for cognacy by historical linguists, meaning that the identification of cognate classes is still a human affair — it was the reconstruction itself given that information that was being evaluated. Their overall result was that Maximal Likelihood performed the “best”, but that there were still several shortcomings. Most notable of these are undetected borrowings, variation within languages and paralleled independent shifts. In this paper, we address the potential for contact events by using sets of trees from a Bayesian posterior, some of which may represent an alternative contact history.

There are also two recent studies of Indo-European grammatical history: [Carling & Cathcart \(2021\)](#) and [Goldstein \(2022\)](#). [Carling & Cathcart \(2021\)](#) evaluate different theories of the history of morphosyntax of Indo-European by comparing to the product of computational Bayesian phylogenetic modelling. They find support for the “canonical” model of Indo-European syntax and illustrate clearly with case examples how their model works. 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*([Goldstein 2022:1/71](#)). This is done because scholars argue that the true tree is unknown, and that this is an appropriate method in the absence of the true tree. Goldstein argues that the appropriate action is instead to carry out reconstruction on many different trees that represent possible histories, a Bayesian posterior tree sample. He argues that this is methodologically

more sound and the results of his approach are in accord with the consensus in historical linguistics, thus strengthening their validity.

Both [Carling & Cathcart \(2021\)](#) and [Goldstein \(2022\)](#) use a Bayesian method of ancestral state known as Continuous-Time Markov Chain (CTMC)<sup>3</sup>. This approach comes with certain important assumptions, to quote from [Goldstein \(2022:77\)](#):

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

For more details on the methods, please see [Goldstein \(2022\)](#), [Pagel et al. \(2004\)](#), [Ronquist \(2004\)](#) and [Liggett \(2010\)](#)<sup>4</sup>.

Computational approaches to reconstruction not only allow us to effectivize the process by inferring the prior states of hundreds of traits in a short span of time, but it also allows us to apply exactly the same principles in exactly the same way to all pieces of data. This is much harder to do manually, since different scholars may use slightly different assumptions and judgement when applying the traditional Comparative Method.

The particular study object of this paper is the Oceanic language subgroup of the Austronesian family and the grammatical features of four of its proto-languages. We use information about the extant daughter languages from the Grambank dataset ([The Grambank Consortium 2021](#)) to infer the structure of proto-languages. The computational methods are applied using two published trees of Oceanic: Glottolog 4.0 and [Gray et al. \(2009\)](#). Findings from the historical linguistics literature have been translated into data-points in the Grambank format for four specific proto-languages: Proto-Oceanic, Proto-Central Pacific, Proto-Polynesian and Proto-Eastern Polynesian. The computational methods take as input the language-level data points in the Oceanic subgroups and then infer grammatical states of proto-languages in the tree. The state of the aforementioned four proto-languages are extracted for each tree and method and compared to conclusions from classical historical linguistics. The results are evaluated in terms

---

<sup>3</sup>The main difference between the methods of [Carling & Cathcart \(2021\)](#) and [Chang et al. \(2015\)](#) is that [Carling & Cathcart \(2021\)](#) use a tree structure informed by [Chang et al. \(2015\)](#) and comparative-historical *communis opinio* and vary the branch lengths 10,0000 times in a principled and informed manner to generate 10,000 different trees while [Goldstein \(2022\)](#) takes 100 random samples directly from the posterior of [Chang et al. \(2015\)](#).

<sup>4</sup>Another popular method of ancestral state reconstruction is Stochastic Character Mapping ([Huelsenbeck et al. 2003](#)). SCM is a procedure that can follow the same the CTMC approach employed by [Carling & Cathcart \(2021\)](#) and [Goldstein \(2022\)](#), but not necessarily. 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. Thank you to one of the anonymous reviewers for highlighting this.

of concordance between each method and the predictions from classical historical linguistics. We are evaluating how much they agree, not necessarily which one is correct. Which method is correct should be estimated based on the conceptual underpinnings and assumptions of the method and how plausible that model of change is. From the exercise in this study we can learn about which method agrees most with classical historical linguistics, and therefore dissect the comparative method into explicitly formally objectively defined mechanisms.

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

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

## 2 Background

f

### 2.1 The methods of traditional historical linguistics: the Comparative Method

In order to interpret the differences between the results of the computational approach versus the classical historical linguistics approach, it is first necessary to clarify the different methodologies and the consequences of them for the study at hand. 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 is known as the “Comparative Method”. The Comparative Method is based on finding words or morphemes in different languages that have the same (or similar enough) meaning and that display non-trivial systematic phonological similarities. By investigating these sets of words, it is possible to deduce which are inherited from a common shared ancestor, i.e. are cognates. For example, Blust (2004), Greenhill & Clark (2011) and many others have reconstructed that Māori /toru/ (meaning ‘three’) derives from the same word in an ancestral language as Hawai’ian /kolu/ (‘three’) does. These two words are “cognates” of each other and this information can be used to reconstruct a form for proto-Polynesian. Furthermore, many words that mean the same/similar thing in Māori and Hawai’ian show this pattern of t/k, e.g. Māori: /mate/, Hawai’ian: /make/ ‘to be dead’ and Māori: /whitu/, Hawai’ian: /hiku/ ‘seven’ (Greenhill et al. 2008). There is a systematic correspondence between these two sounds; regularly when there is a /t/ in Māori there is a /k/ in the corresponding position in Hawai’ian. This is known as a *systematic sound*

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

Historical linguists use cognates and systematic sound correspondences to develop hypotheses about forms in unobserved proto-languages and to propose sub-groupings based on shared innovations (c.f. how biological cladistics finds relationships between species based on shared derived characteristics from common ancestors (Maclaurin & Sterelny 2008:16-17)). The Comparative Method provides us with a) sets of words which derive from the same word in an ancestor language (cognates), b) sequences of sound changes from proto-languages to the current observable daughter languages and c) a tree or network structure of the relationships between languages.

The processes of subgrouping and reconstruction are done in tandem in historical linguistics; they are estimated simultaneously. Subgroups are proposed based on shared innovations. In order to determine what is and what is not an innovation, a certain amount of reconstruction is necessary. In order to make reconstructions, some of the tree structure needs to be approximated. Pawley (personal correspondence) notes that most of the subgrouping done in historical linguistics tends to be at the lower level. In this paper, we are only focussing on the reconstruction — not on the identification of regular sound correspondences, cognates or subgroups.

The Comparative Method in historical linguistics relies on knowledge of probability of changes — phonological and semantic. For example, many historical linguists posit that /s/ is more likely to become /h/ than it is to become a /k/<sup>5</sup>. In the above 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, Pawley (2005) reconstructs the proto-form \*panua as meaning ‘land’ or ‘inhabited territory’. In daughter languages, this has changed to ‘place’, ‘community’, ‘village’, ‘house’, ‘people’, ‘world’ and ‘weather’. These are related meanings, but not identical. Historical linguists aim to have plausible semantic connections between words that are proposed to stem from the same proto-form, they cannot be too dissimilar.

The Comparative Method is typically applied to sound changes and the lexicon, but it can also be applied to structure. Clark (1976:17-22) wrote about Proto-Polynesian syntax for example. He outlines three important principles for reconstruction that are relevant to lexion, sounds and grammar:

- (i) the number of changes posited
- (ii) the plausibility of the changes posited

---

<sup>5</sup>Historical linguists do concede that there are instances of irregular sound change (Blust 1996; Campbell 1996) and that while they can often be explained by contact, analogy or avoidance of homophony, they sometimes remain unexplained.

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

The first of these principles is the same as what is known in phylogenetics as “Maximum Parsimony”. The idea is to reconstruct states in proto-languages such that there are as few changes as possible between nodes in the entire tree. Clark (1976:17-22) explains how this works by positing an example of seven languages where there is a majority of one feature, X, and fewer of another, Y. Fig. 1 illustrates this example. If we only examine which feature is the most common, we should reconstruct X at the root of this tree. However, this would result in 2 changes (one each between the root and tips A and B). 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.

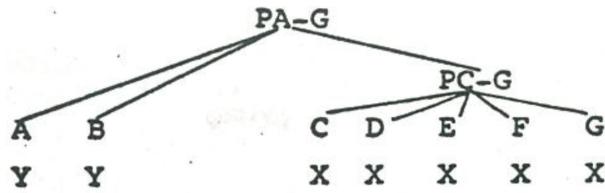


Figure 1: Tree from Clark (1976:19) illustrating Maximum Parsimony.

It is important to note that Maximum Parsimony does not take into account the length of branches, only the changes between each node of the tree (regardless of how far apart they are). Furthermore, the solution that Maximum Parsimony selects is the Parsimony solution with the fewest changes (also known as “lowest Parsimony cost”) — which can be interpreted as the solution with the slowest rate. We cannot compare Parsimony cost directly to rate of change estimated by other models, but out of all the possible solutions Maximum Likelihood generates it does not necessarily select the one with the slowest rate of change (more on this in section 3.1).

The Comparative Method procedure that has been outlined in this section can and has been applied successfully to sounds and words — be they lexical or grammatical. The application of this approach to abstract structural features is more controversial. One of the concerns is whether the underlying data represents genuine correspondences (for example by satisfying the Double Cognacy Condition suggested by Walkden (2013)). As a consequence of that, it can be said that not all reconstruction in historical linguistics is based on the Comparative Method in a strict sense. . Some of the work on reconstruction in historical linguistics cited in this paper can be said to represent *traditional approaches*, but perhaps not necessarily the *Comparative Method* in a strict sense.

This section and section 2.2 are being re-written to more clearly tease out and discuss details of reconstruction of abstract structural features.

### 2.1.1 Disagreements in historical linguistics

Reconstruction in historical linguistics includes judgements of plausibility. This requires some assumptions about what are plausible features to co-occur in language, and which pathways of language change are more plausible than others. For example, it is rare to find a language that has a gender distinction in first person, but not in third (though not impossible; c.f. [Siewierska \(2013\)](#)). If the most parsimonious reconstruction results in a proto-language with many rare features or unusual combinations of features, it may require reconsideration. If something is rare in the languages that exist today, we would expect it to be rare also in past languages. Similarly, changes from certain states to others are assumed to be less plausible. For example, a language going from having no marked dual number on nouns to having trial number would be taken as unusual by most linguists (c.f. [Kikusawa 2006:8](#)).

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

[Clark \(1976\)](#) disagrees with [Hale \(1968\)](#), [Hohepa \(1969\)](#), and [Chung \(1978\)](#) on the state of Proto-Polynesian syntax on these grounds. Chung, Hale and Hohepa argue for a theory that is technically less parsimonious, but which they say is more plausible. They posit that Proto-Polynesian had a nominative-accusative case marking system<sup>6</sup>. If this was the case, that would mean positing more changes along the tree than if we assumed, as [Clark \(1976\)](#) 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 compared to the rest of the group which most often lacks ergative-absolutive marking. Fig. 2 shows the Polynesian tree with Grambank feature GB409 values marked out<sup>7</sup>.

<sup>6</sup>Hale, Hohepa and Chung actually suggest three different specific theories for this reconstruction. For a summary of the differences between the proposals, see [Chung \(1978:247-249\)](#).

<sup>7</sup>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.

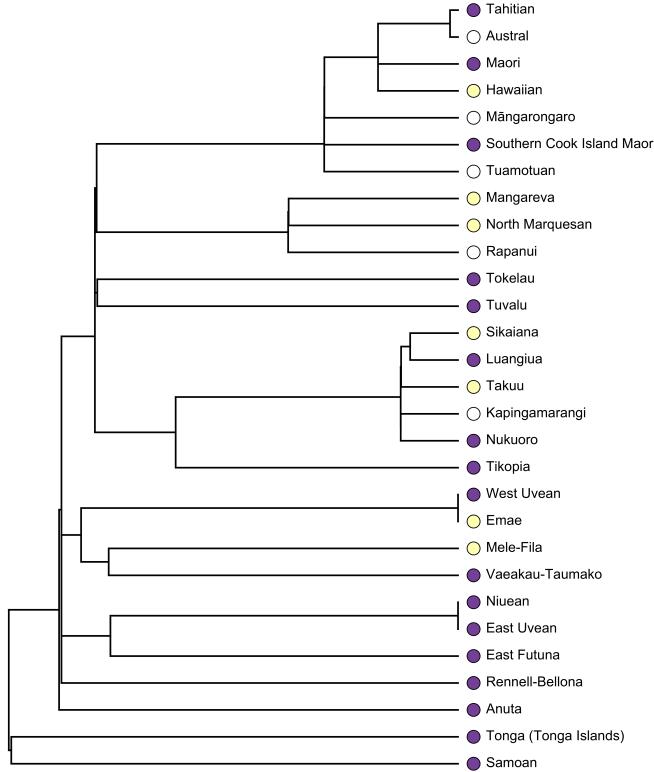


Figure 2: The Polynesian languages in the Gray et al. (2009) Maximum Clade Credibility Tree-tree, with the coding of Grambank feature GB409 “Is there any ergative alignment of flagging?” marked out. Purple = Yes, Yellow = No and white = Not enough information/not clear.

Chung’s critique of Clark’s proposal is three-fold:

- (a) the tree used is not an accurate representation of the language history (there was more interaction between Sāmoan and Tongan after splitting, i.e. horizontal transfer)
- (b) it is possible that the Proto-language contained variation and was undergoing change that was only fully realised in some of the daughters
- (c) the morpho-syntactical historical process itself is less plausible

In a review of Clark (1976), Chung writes:

*Such an approach [as Clark’s] relies on the assumption that the subgroups have developed quite independently once they split off from Proto-Polynesian, so that features shared by both must be attributed to the Proto-language. But in fact, both parts of this assumption are too strong. It is well known that the two primary subgroups of Polynesian did not develop totally separately; there was long-standing contact in pre-European times between speakers of Tongic and some Samoic-Outlier languages, as Clark himself notes (p. 27). Further, and more generally, it is*

*simply not true that every feature shared by related languages must have existed in the Proto-language uniting them. Languages are constantly undergoing change; and 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 which existed only in embryonic form, or not at all, in their common ancestor.*

Chung (1977:539)

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

In this study, any instances of conflicting data from historical linguists concerning proto-languages are evaluated separately from the overall results and will be reported in a separate section. There are three instances of this: two features related to the alignment of Proto-Polynesian (GB408 Is there any accusative alignment of flagging? and GB409: Is there any ergative alignment of flagging?) and one feature for Proto-Central Pacific, where Kikusawa (2002) and Ball (2007) disagree on the alignment as well.

## 2.2 Characteristics of language structure as data

This section is being re-written to more clearly tease out and discuss details of reconstruction of abstract structural features. c.f. Walkden (2013).

The Comparative Method is most often applied to lexical words and sounds, but it can also be applied to grammatical morphemes and features. Crowley (1985) for example traces the history of a common noun phrase marker *\*na/\*a* in Oceanic languages using the Comparative Method. (c.f. also the case studies of Goldstein (2022)).

The data in this paper does not track specific forms, as is common when reconstructing proto-languages in historical linguistics (c.f Pawley (1973); Crowley (1985); Evans (2001)). Instead we use binary features of a typological questionnaire which tracks a large part of “core” grammar — the Grambank dataset. (c.f. Carling & Cathcart (2021)) This section outlines some crucial differences between structural data and the kind of data that is typically used in historical linguistics in relation to the present study.

The kind of data used in grammatical reconstruction in historical linguistics differs from what we find in linguistic typological questionnaires such as Grambank. Crowley (1985), Clark (1976), and other scholars whose work we will compare to our results in this paper, typically apply the comparative method to specific formal expressions of structural features (the *na* article, *-Cia* suffix, *faka-* prefix etc). They take into account fossilised forms (the common noun marker *-a* fusing to roots in Paamese (Crowley 1985:141)) and related meanings (the hypothesis of *-Cia* changing from a transitivising suffix to a marker of passive voice (Hale (1968); Hohepa (1967, 1969); Chung (1978) and Jonsson (1998)). 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 participate in the function productively.

Surveys of this kind do not track forms, but abstract features such as “Does the language mark passive voice?”. This means that two languages can be coded identically for entirely different reasons and without being related. For example, Koasati [koas1236] of Louisiana, USA, and Mokilese [moki1238] on Mwoakilloa in the Federated States of Micronesia are both coded as having a construction for predicative possession of the type “Topic” by Stassen (2011). However, they belong to entirely different language families and different parts of the world. Their similarity does not necessarily imply shared inheritance. This is unlike cognacy data, where the fact that two languages have cognates in common is direct evidence of relatedness.

As an example of this difference, let’s consider definite markers in Oceanic languages. Crowley (1985) investigates “common noun phrase markers”<sup>8</sup> in Oceanic and finds that in many languages there is a reflex of proto-Oceanic *\*na/\*a*, but that in some languages there is another marker with a different origin (Māori *te* for example). In Crowley’s study, languages where there is no common noun phrase marking whatsoever and those with a marker which is not cognate with *\*na/\*a* are both included in type 1 (see Fig. 3). These languages are contrasted with those that have retained some kind of reflex of *\*na/\*a* (type 2-4 in Fig. 3). 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.

In contrast, the corresponding feature in Grambank is ‘GB022: *Are there prenominal articles?*’ (see Fig. 4). 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. We can now distinguish those that have a pre-nominal article from those that do not, but we cannot tell apart those which have retained the proto-form from those which have not. Since many reconstructions of grammar in historical linguistics rely on particular forms,

---

<sup>8</sup>This term is more or less identical to a pre-nominal definite/specific article.

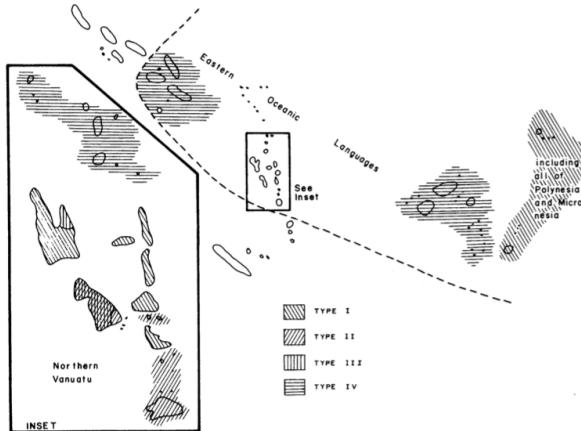


Figure 3: Map of four different types of common noun phrase markers in Oceanic from Crowley (1985:162). Type 1: absence of common noun phrase marker or marker is not a reflex of *\*na* /*\*a*, type 2: non-productive system involving a reflex of *\*na* /*\*a*, type 3: productive marking involving *\*na* /*\*a* as a prefix that is regularly separable from the noun and type 4: productive marking involving *\*na* /*\*a* generally existing as a free-standing marker.

this is an important difference. This does not matter for features such as word order.

The data in this study is composed of abstract features such as “is a grammatical distinction made between X and Y?”. This makes it different from most studies of grammatical features in historical linguistics. As was noted earlier, it is possible for two languages to be coded alike but not share ancestry. It is also possible that such abstract features track something beyond the particular forms. Ross (2004:503) 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<sup>9</sup>. This may be true of more features, and in such cases languages can share a grammatical feature but not have the same particular expression, and this could be due to inheritance. In addition, in many instances — particularly within families — abstract features such as ‘GB022: *Are there prenominal articles?*’ are in fact correlated with specific forms even if the forms are not explicitly being tracked.

### 3 Material and methods

#### 3.1 Computational phylogenetic methods

In this study, we will be reconstructing the presence or absence of structural features in proto-languages of the Oceanic subgroup using Maximum Parsimony and Maximum Likelihood. This section gives a brief overview of the methods. Further technical details concerning the precise application can be found

<sup>9</sup>He also notes that Goddard has observed similar patterns in Algonquian languages (Goddard 1993).

## GB022 ARTPre

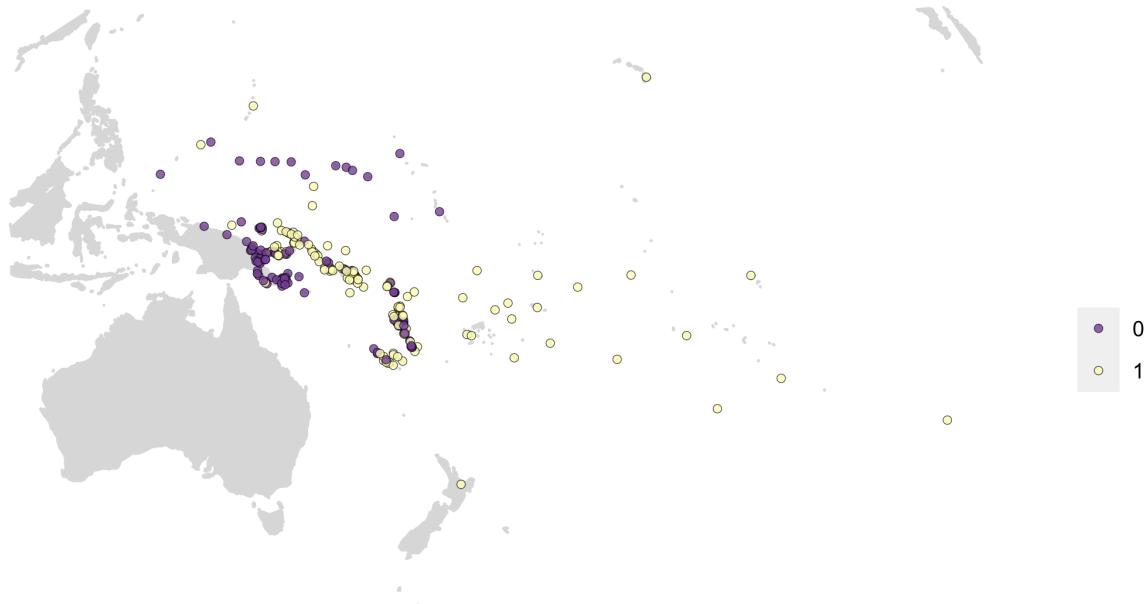


Figure 4: Map of Austronesian languages for GB022 *Are there prenominal articles?* Yellow = “yes”, purple = “no”.

in the supplementary material B. For an extensive comparison of different methods of ancestral state reconstruction and their advantages, see Joy et al. (2016).

As discussed earlier in section 2.1, **Maximum Parsimony** finds the set of ancestral states that result in the fewest number of changes between nodes. Maximum Parsimony is intuitively simple. We saw in the previous section an example of how it can play out in a small tree. While the principle of “Maximum Parsimony” is practised in traditional historical linguistics as a part of the Comparative Method, it should be noted that they rarely use the term *per se*, but rather the description of “fewest number of changes along the tree”.

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

Ancestral state reconstruction using **Maximum Likelihood** posits the most likely ancestral state distributions based on the overall probabilities given all the nodes in the tree and all branches. This approach does not assume that the slowest rate of change is the most probable one. If, for example, the distribution of values at the tips is very scattered, with sibling pairs frequently having different values, Maximum Likelihood will infer that the feature has a high rate of change and use that information when positing

ancestral states as well. The Maximum Likelihood algorithm assigns probabilities of state changes and distributions based on branch lengths. A mutation along a shorter branch is given more weight in the likelihood calculations than if it occurred in a longer branch. Furthermore, reconstruction using Maximum Likelihood allows us to use a model of change where we do not assume that the rates for losses ( $1 \rightarrow 0$ ) are equal to the rate of gains ( $0 \rightarrow 1$ ). In this study, we use an “All Rates are Different” (ARD) model, which allows for the rate of loss and gain to be different<sup>10</sup>.

It is not possible for the computational implementation of Maximum Parsimony to take into account branch lengths, nor can it assume anything but the slowest rate of change or posit different rates for losses and gains. It is however possible that the Comparative Method in historical linguistics estimates something similar when they take into account the “plausibility of the changes posited” and that this is picked up by the reconstruction by Maximum Likelihood. It is possible that scholars of historical linguistics take length of time into account, for example, or assume that a loss is more likely than a gain for a given feature. In this study, we compare Maximum Parsimony and Maximum Likelihood reconstructions to the Comparative Method. If the results from the Comparative Method are more similar to that of Maximum Likelihood, a potential explanation would be that the “plausibility of changes posited” is indeed operating along similar lines as Maximum Likelihood by taking branch length into account and assuming varying rates of change.

Finally, we will also compare the predictions of historical linguists with computational predictions based solely on which value is the most common in the daughter languages of a given proto-language — entirely disregarding the tree structure<sup>11</sup>. In the toy example in Fig. 1, this approach would reconstruct that the root had feature value “X”. Whether we prefer Maximum Parsimony, Maximal Likelihood or another approach to reconstruction, it should be the case that actually taking the tree structure into account is the sounder methodology.

### 3.2 Calculation of similarity between predictions from the Comparative Method and computational approaches

We calculate the similarity of the predictions of historical linguists and computational means with two measures: concordance and F1-scores. Concordance is known as *accuracy* in machine learning, but we wish to avoid the connotation that what is being measured is the accuracy of the reconstruction as regards real languages. Rather, concordance measures how closely the computational reconstruction matches historical linguists’ reconstruction. It is measured as the number of agreements about grammatical features (i.e. Grambank binary questions) of predicted protolanguages, divided by the total number of grammatical features predicted.

---

<sup>10</sup>Similarly to the studies by Carling & Cathcart (2021) and Goldstein (2022), rates cannot vary within the tree.

<sup>11</sup>This is similar to the frequency heuristic described in Goldstein (2022).

F1-scores are the harmonic mean of the precision and recall<sup>12</sup> (Van Rijsbergen 1979:133). It is important to note that F1-scores disregard the number of True Negatives entirely, which is relevant in our case since some of the features in proto-languages are predicted to be absent. For both measures, 0 is the worst possible score and 1 the best in terms of similarity to predictions by historical linguists.

In a similar study of ancestral states of cognate classes, Jäger & List (2018) compared three different methods of ancestral state reconstruction for lexical data (cognate classes): Maximum Parsimony, Maximum Likelihood and Minimal Lateral Networks. They found that reconstructions using Maximum Likelihood performed the most like the predictions by historical linguists. However, Jäger & List (2018) describe the general performance of all the computational reconstruction methods they used as “poor”. Jäger & List (2018) evaluated the methods using the F1-score. The highest F1-score was 0.79 (Austronesian language sample, Maximum Likelihood), and the worst was 0.44 (Indo-European, Minimal Lateral Networks).

For each feature, the methods predict a distribution of the two states (presence and absence) for every ancestral node. If the distribution is majority presence (more than 60% of the ancestral state is “1”) it is registered in the results as “Presence”; if less than 40% presence it is registered as “Absence”. If the ancestral state is between 40-60% of either state, the prediction is registered as “Half/Half”. This was done to highlight the amount of uncertainty the results sometimes contain, while at the same time making it a fair comparison between Maximum Parsimony and Maximum Likelihood. Comparing the raw distributions themselves is not a fair comparison because Maximum Parsimony is always more likely to suggest 0, 0.5 or 1 results whereas Maximum Likelihood rarely produces exactly 0 or 1.

If the reconstruction of a feature by experts for that ancestral node was “Presence” and the algorithm did predict presence with over 60%, it is counted as a “True Positive”, and so on<sup>13</sup>. Table 1 illustrates how the results are summarised.

Table 1: Table illustrating how the results of ancestral node predictions are calculated.

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

For each method, a plain concordance score (Eq (1)) and F1-score (Eq. (2)) is then calculated.

<sup>12</sup>Precision is True Positives divided by True Positives + False Positives, recall is True Positives divided by False Negatives + True Positives. F1-score =  $2 * ((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$  (Van Rijsbergen 1979).

<sup>13</sup>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. It should not be interpreted as a measure of “Truth” necessarily.

$$\frac{\text{True Negative} + \text{True Positive}}{\text{True Negative} + \text{True Positive} + \text{False Negative} + \text{False Positive}} \quad (1)$$

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

It is also important to take into account the Half-results. This count represents instances where the method was not able to say with a 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 one way, 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 type of cases into account the cut-off of 40%-60% was set and summarised as "Half" results. We can apply the concordance score and F1-scores to this summary statistic as well, as shown in Eqs. (3) and (4). See supplementary material C for more details<sup>14</sup>.

$$\frac{\text{True Negative} + \text{True Positive} + \frac{\text{Half}}{2}}{\text{True Negative} + \text{True Positive} + \text{False Negative} + \text{False Positive} + \text{Half}} \quad (3)$$

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

All four 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 True Negative (which F1-scores ignore) and it takes into account the possible uncertainty of the half-scores.

### 3.3 Data

#### 3.3.1 The Grambank-dataset

The data for the study is taken from the Grambank-project ([The Grambank Consortium 2021](#)). The Grambank dataset consists of 195 structural features which have been coded by a large group of research assistants for over 2,000 languages. This dataset includes 280 Oceanic languages.

The questionnaire's 195 questions cover what are often called the “core domains” of traditional grammatical description: word order, possession, negation, tense, aspect, mood, deixis, interrogation, comparatives and more. Features are included in the questionnaire if it is likely that it is possible to answer them for the majority of the world’s languages which are at least described in a grammar sketch (approx 4,000

---

<sup>14</sup>I am very grateful for mathematical assistance from Stephen Mann.

languages). This means that rarer features are not included. The full questionnaire is found in appendix E.

The Grambank dataset is coded by student and research assistants 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 languages. For more details on the coding workflow of Grambank, see Slingerland et al. (2020).

### 3.3.2 Data coverage

This study is focussed on the Oceanic subgroup of the Austronesian language family. The Oceanic subgroup covers almost all languages in Remote Oceania (with the exceptions of Chamorro and Palauan) and large parts of Near Oceania. Fig. 5 from Ross et al. (2016:2) shows the extent of the major subgroups of the Austronesian language family, with Oceanic covering the largest surface area. Following the language classification of Glottolog 4.0 (Hammarström et al. 2019), there are 522 languages in total in the Oceanic subgroup.

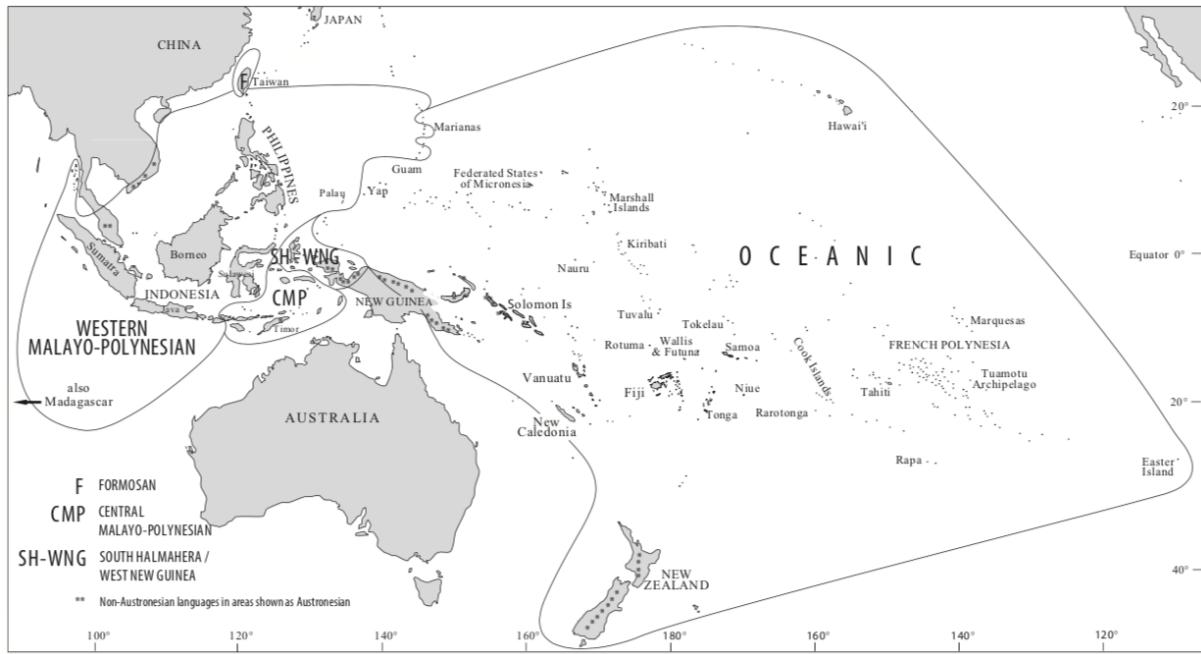


Figure 5: Map of the Austronesian language family and major subgroups, from Ross et al. (2016:2).

Not all languages of the Oceanic subgroup have a grammatical description, but out of those that have it nearly all are included in Grambank. Table 2 shows the coverage of Oceanic languages in the entire dataset. According to Glottolog, there are 289 Oceanic languages that have a grammar or grammar sketch which means that they can be included in Grambank at all. It is not always possible to fill in all the features for every language. Twenty of the Oceanic languages that are included to date are less than 50% completed for Grambank questions. This can be due to lack of access to descriptive work, or that

the content of the descriptive work does not cover the necessary domains in enough detail for the coders to answer enough questions. The map in Fig. ?? shows the same coverage information, with languages coded for their data coverage status.

Island group	More than half of the features covered in Grambank	Less than half of the features covered in Grambank	Grammar exists, but language not in Grambank (yet)	No grammar
Bismarck	42	7	0	5
Central Pacific	33	1	1	10
Central Vanuatu	48	1	0	42
Interior New Guinea	4	0	0	11
Micronesia	16	1	0	6
N Coast New Guinea	19	3	2	76
New Caledonia	14	0	2	17
Northern Vanuatu	5	0	0	9
S New Guinea	26	1	3	36
Solomons and Bougainville	30	4	1	25
Southern Vanuatu	8	0	0	1
Temotu	5	2	0	3
Total	250	20	9	241

Table 2: Table showing coverage of Oceanic languages in Grambank per island group.

The coverage of Grambank data for the Oceanic subgroup is in general better in the east than in the west. However, since we control for genealogical relatedness through the distance control approach, this is less of a problem for our methodology than if we were using traditional probability sampling (c.f. Ross (2004)).

### 3.3.3 The trees

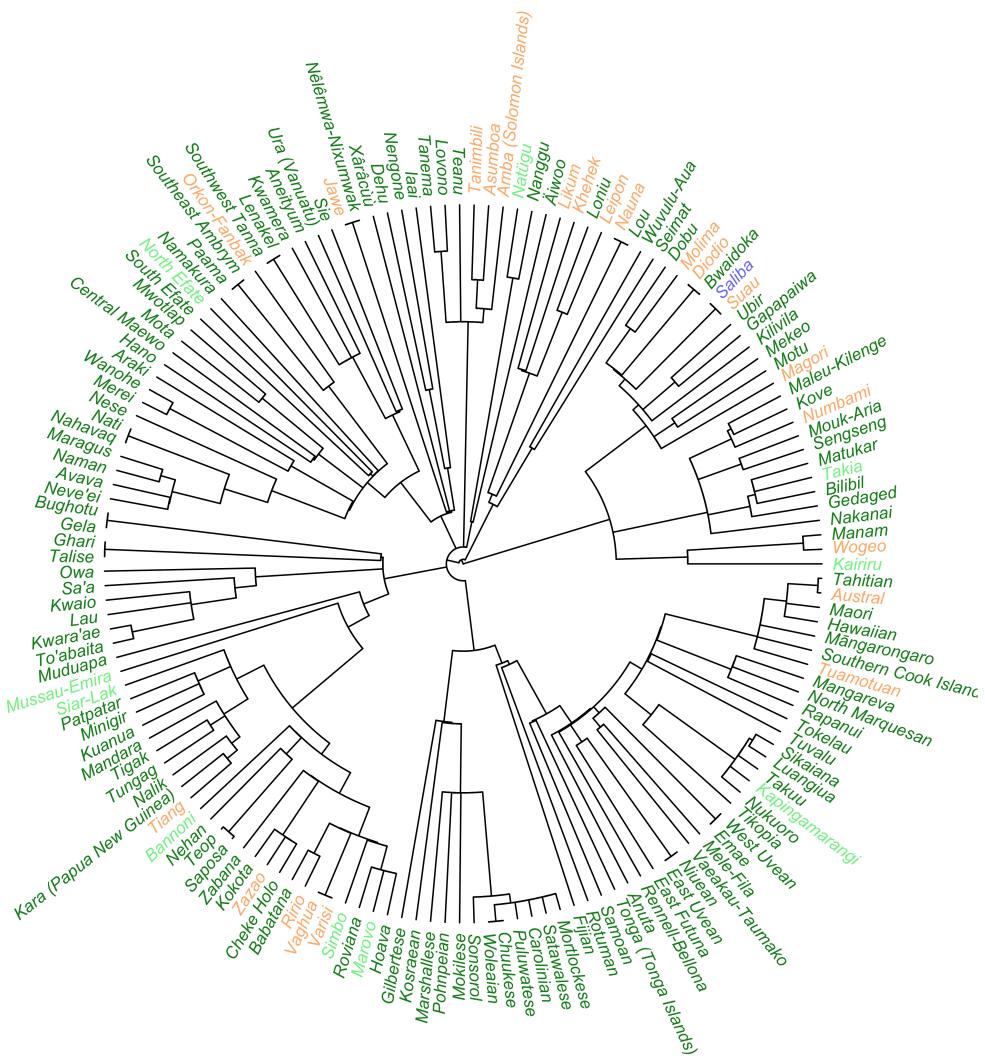
The tree phylogenies used in this study are: a) the Maximum Clade Credibility Tree (MCCT) from Gray et al. (2009); b) a random sample of 100 posterior trees from the same source; c) the tree from Glottolog 4.0<sup>15</sup>.

Figure 7 and Figure 8 show the Grambank coverage of languages over the phylogenies from the Gray et al 2009-MCC-tree and the Glottolog-tree respectively.

---

<sup>15</sup>The tree of Glottolog 4.0 (Hammarström et al. 2019) is based on work by Blust (2009, 2014) and Blust & Chen (2017).

## Coverage of the Oceanic subgroup in Grambank (Gray et al 2009 MCCT tree)



- orange: grammar doesn't exist
  - blue: grammar exists (not in GB, yet)
  - green: Less than half of features covered in GB
  - dark green: More than half of features covered in GB

Figure 7: Maximum Clade Credibility Tree of Oceanic from Gray et al. (2009), with languages coloured for coverage in Grambank.

### Coverage of the Oceanic subgroup in Grambank (Glottolog 4.0-tree)

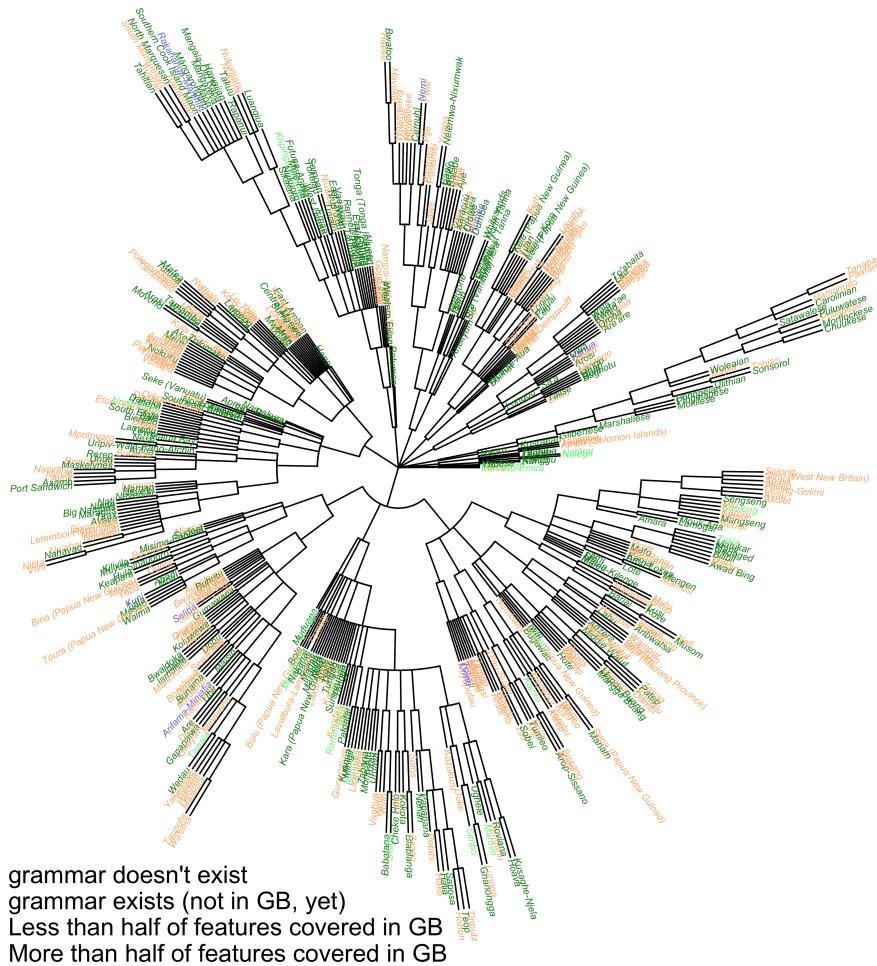


Figure 8: Tree of Oceanic from Glottolog, with languages coloured for coverage in Grambank.

One of the major differences between the trees is that the Glottolog-tree does not contain *any* information about branch lengths. All the branches in the Glottolog-tree are of the same length (1), whereas the branches in the Gray et al 2009-trees (the MCCT and the posteriors) have meaningful branch lengths

based on rates of change in the underlying data (basic vocabulary) and calibration points (archaeological dates). This can be seen in the visualisations in Figures 7 and 8 where the first has varying lengths of branches but the latter all have a uniform length (1). This has the consequence that some tips in the Glottolog-tree are much further from the root than others. This is a big disadvantage with this type of tree, since it suggests that different amounts of time has passed between the root and the languages at the tips. Further technical details of the trees can be found in Supplementary Material F.

The Glottolog-tree contains all the languages in the Oceanic subgroup. Therefore the coverage per island group that is summarised in table 2 in the previous section applies to the Glottolog-tree as well. However, the Gray et al. (2009)-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 Gray et al. (2009). Their study yielded 4,200 posterior trees. Tree topologies that are more probable occur more often, but less probable trees also appear in the posterior. By using a set of possible trees instead of just one we can include other possible histories into our analysis, which could for example estimate contact events as well as inheritance. Figure 9 shows a visualisation of the 100 trees which are used in this study.



Figure 9: DensiTree (Bouckaert & Heled 2014) visualisation of the 100 random sampled trees from the Gray et al 2009-posterior. Made with the function `densiTree()` from the R-pacakge Phangorn (Schliep et al. 2019).

### 3.3.4 Data from historical linguistics on Oceanic proto-language grammar

The proto-languages of the Oceanic subbranch of the Austronesian language family are generally well researched in terms of their lexicon and phonology (see the book series on the Proto-Oceanic lexicon (Ross et al. 1998, 2007, 2008, 2011, 2016), among other publications). There is also substantial work done on the grammar of Proto-Oceanic using the comparative method. I have summarised several major works in the field and distilled their research into testable hypotheses given the Grambank data and our methods. This section gives an overview of the works included and examples of how they have been incorporated into the study. Table 7 lists the publications used for the reconstruction of proto-Oceanic by historical linguists.

For each of these publications, findings have been extracted that support a certain state of a Grambank feature at a certain node. For example, Marck (2000:4) writes that a causative prefix has been reconstructed for Proto-Polynesian (\**faka*-). In the Grambank questionnaire we have the feature GB155 ‘Are causatives formed by affixes or clitics on verbs?’. For the ancestral node that connects all the Polynesian languages, we should expect that for GB155 the state is either wholly or overwhelmingly “1” (yes/presence). For simplicity, I have only considered 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. The literature suggests that Proto-Oceanic was a language with a pre-nominal definite/specific article (Crowley 1985:136), a distinction between inclusive and exclusive first person pronouns (Pawley (1973:112), Crowley (1985:184), Ross (2004:500), Lynch et al. (2011:67, 75)), no gender distinctions in pronouns (Ross 2004:498), a dual number category in pronouns (Ross (2004:498), Lynch et al. (2011:69) and Pawley (1973:173)), a distinction between alienable and inalienable possession<sup>16</sup> (Lynch et al. 2011:69), prepositions (Pawley (1973:167), Ross (2004:498)), subject proclitics and object enclitics on the verb (Ross (2004:498-499), Lynch et al. (2011:83)), possessive suffixes on the possessed noun (Ross (2004:495), Pawley (1973:155)) and a transitivising suffix on verbs (Pawley (1970:352), Pawley (1973:171), Lynch et al. (2011:80, 92)). The reconstructions regarding ergativity will be presented separately.

Most of the time, the scholars of Proto-Oceanic are in agreement in their predictions. For example, Pawley (1973:142), Ross (2007:292), Clark (1976:xiii, 125) and Lynch et al. (2011:89) all propose that the proto-language of the Polynesian subgroup had a construction marking prohibitive that was different from declarative negatives. However, in some instances there are disagreements. As discussed earlier, one such case is the alignment system of Proto-Polynesian. Clark (1976) claims that the system was ergative while Hale (1968), Hohepa (1967, 1969) and Chung (1978) argue that Proto-Polynesian was

---

<sup>16</sup>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 into the same category.

accusative and several of the daughter languages developed ergativity later. Kikusawa (2002) and Ball (2007) also disagree on the alignment of Proto-Central Pacific. Because of this disagreement, the results for the computational ancestral reconstruction for Grambank features regarding alignment of these two proto-languages are presented separately.

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 E). 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 employ, 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. Out of the 201 (binarised) features in our questionnaire, 33% (67) were answerable for Proto-Oceanic given this material. The average completion per language in the whole of the Grambank dataset is 75%.

## 4 Results

### 4.1 Concordance between traditional historical linguistics and computational methods

We are examining results from three approaches in total: a) Maximum Parsimony, b) Maximum Likelihood and c) Most Common value in daughter languages. For (a) and (b) we are also using three different trees: i) Glottolog, ii) Gray et al. (2009) MCC-tree and ii) the mean values of reconstruction of a random selection of 100 (out of 4,200) trees in the Bayesian posterior of Gray et al. (2009). That gives  $2 * 3 + 1$  results, i.e. 7.

All results have been calculated in R (R Core Team 2019) using the packages `castor` (Parsimony, Louca & Doebeli (2017)), `phangorn` (Parsimony, Schliep et al. (2019)) and `corHMM` (Maximum Likelihood, Beaulieu et al. (2017)). The packages `ape` (Paradis et al. 2004), `aedphylo` (Jombart et al. 2017), `phytools` (Revell 2012) `reshape2` (Wickham 2020) `tidyverse` (Wickham et al. 2019) were also used for data wrangling, analysis, summarising and visualising.

Table 3 shows the number of False, Positive and Half-results for each method and tree<sup>17</sup>. One of the most striking features in Table 3 is the large amounts of Half-results in the Most common Method — the method where we simply count directly what is most common in all daughters. This means that there were many instances where this approach, which we know to be unsound, would not confidently be able to predict a presence or absence.

Given these counts, we can calculate the concordance and F1-scores (see section 3.2). These are displayed in Fig 10. A score of 1 means identical to the predictions of historical linguists and 0 means entirely dissimilar from it.

The inclusion of the half-results have the effect of evening out the differences between the performance of the different methods. The concordance scores (including half-results) and the F1-scores (including half-results) for each method are more similar to each other.

Compared to the F1-scores from the lexical reconstruction of Jäger & List (2018), all of the methods achieved higher scores. In this study, only statements about ancestral languages that could be mapped to Grambank-features were included. It is possible that the study by Jäger & List (2018) had a greater overlap between all the reconstructions made by historical linguists and the meanings that they had data for. In that case, it is possible that the features that were possible to map to Grambank data were also those that Oceanic historical linguists are the most confident about — hence the higher scores.

---

<sup>17</sup>There was one feature for the ML method and the Gray et al 2009-trees where the computation could not be carried out because all the languages had the same value. In such cases, the function used (`corHMM` from Beaulieu et al. (2017)) gives an error because it cannot compute the rates matrix.

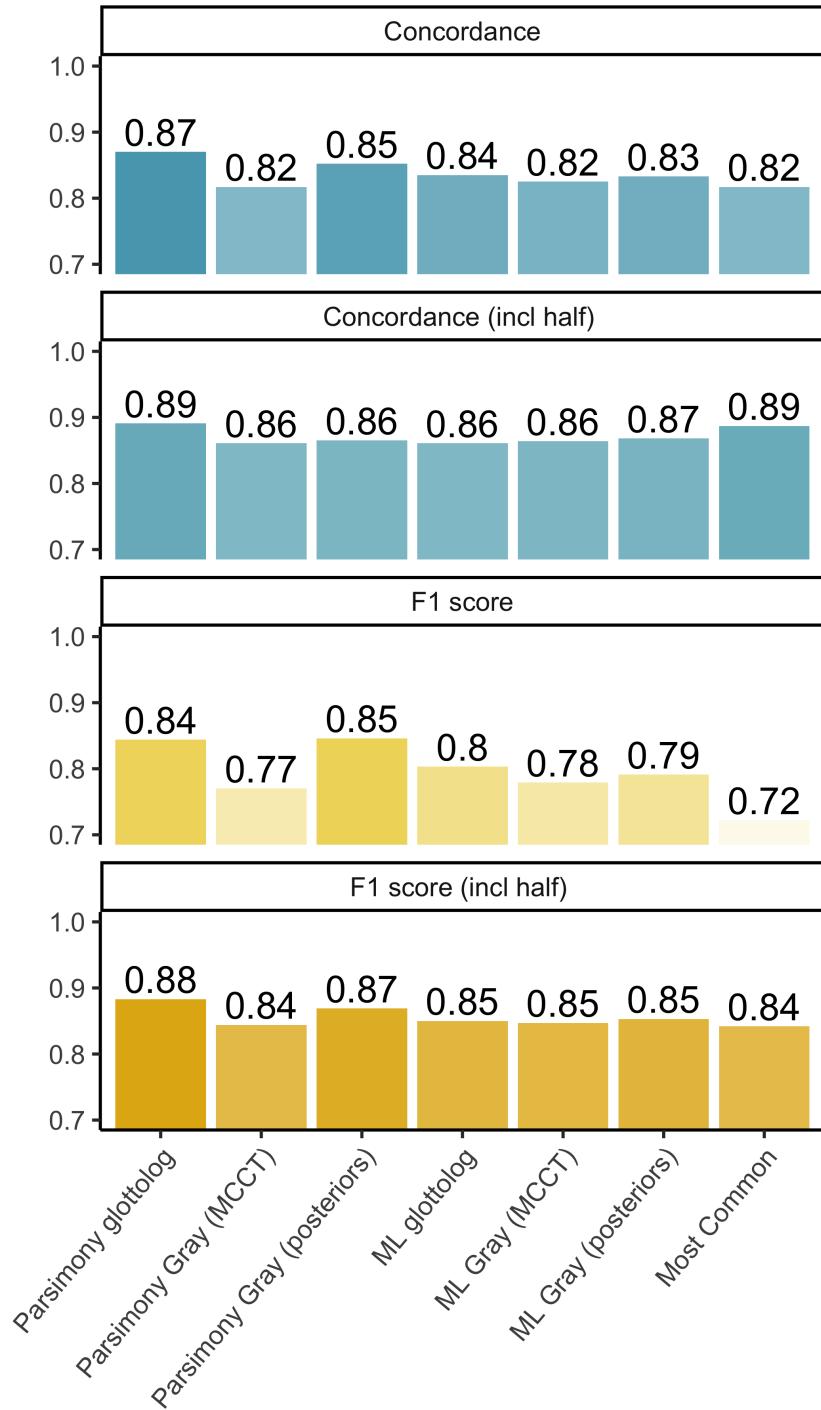


Figure 10: **Barplots of concordance and F1-scores of each method.** NB that the y-axis starts from 0.7.

Method	False Negative	False Positive	Half	True Negative	True Positive	Total
Parsimony Glottolog (4.4)	8	2	5	46	54	115
Parsimony Gray et al (2009) - MCCT	6	5	10	42	52	115
Parsimony Gray et al (2009) - posteriors	8	6	3	43	55	115
ML Glottolog (4.4)	10	3	6	45	51	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

Table 3: Table showing the amount of False Negative, False Positive, Half, True Negative and True Positive results.

The method that tends to perform most similarly to historical linguists is Parsimony + the Glottolog 4.0 tree. The Glottolog 4.0 tree has a big disadvantage; it has no branch lengths and the topology is composed of a combination and compromise from several different sources as opposed to a principled and systematic investigation of lexical data. Parts of the tree are suggested by different scholars, which means that different clades are not necessarily comparable. It does have an advantage as well, and that is the sheer number of languages it includes. The overlap between languages included in Gray et al. (2009) is lower than the overlap between the Glottolog 4.0 tree and Grambank. It is possible that it is this sheer number of tips that gives it a greater concordance with historical linguists' predictions. The results also suggests that it is likely that historical linguists, in these specific studies, do not necessarily take into account branch lengths.

Overall however, the methods preform similarly. Where they vary the most is for the plain F1-score (i.e. without including half-results). While it is appropriate to report this score as it is common practice when assessing performance in this manner, it is probably the score that is the least relevant for this particular study object since it entirely ignores True Negative-counts. Given that, if we focus instead on the concordance score there is very little that tells apart the different methods — they are giving very similar results. The results for MP and ML for the random 100 posterior samples performs slightly better than the plain MCC-tree, which may indicate that they are possibly picking up some relevant history not represented in the MCC-tree, such as horizontal transfer.

## 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. There are 110 features that are predicted as present in the four proto-languages by the two methods with each tree(s) (Glottolog, Gray et al 2009 MCCT and ditto posteriors); i.e. 6 times. For example, that Proto-Oceanic has inclusory constructions, Proto-Central Pacific uses verbs for property attribution (“adjectives”) and Proto-Polynesian has numeral classifiers. A full list of these are found in Supplementary Material I.

Many of these are probably not surprising to most historical linguists working on Oceanic languages, but they were not explicitly predicted in the sources surveyed for this study.

## 4.3 Where the conflicts are: Ergativity

The nature of the alignment system of Proto-Polynesian and Proto-Central Pacific is contested. Clark (1976) posits, primarily on the basis of Maximum Parsimony, that Proto-Polynesian was ergative whereas Hale (1968), Hohepa (1967, 1969), and Chung (1978) argue that it was accusative (while they suggest different historical pathways, they agree that Proto-Polynesian was nominative-accusative).

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 several of these, i.e., from the perspective of Grambank languages aren’t “ergative” or “accusative” — they can have both ergative and accusative flagging simultaneously. This makes it possible for us to prove both Chung and Clark “right”, the results can come out such that Proto-Polynesian had both accusative *and* ergative alignment flagging. However, the results in fact come out strongly in favour of the proposal by Clark. Table 4 shows that MP, ML and MC all reconstruct presence for ergative flagging in Proto-Polynesian. On the matter of nominative-accusative marking, there is disagreement with the MP results all suggesting absence but the ML and MC giving a half-result.

As was noted earlier, the computational reconstructions differ from those arrived at through the comparative method primarily because the data used in this study is abstract presence or absence of structural features whereas historical linguists use specific concrete forms instead (c.f. Crowley (1985)). In the case of alignment systems, the matter of concrete markers is less of an issue. However, besides the parsimony principle (as laid out by Clark (1976:19) for example), expert historical linguists also take into account the plausibility of the proposed proto-language and the chain of changes posited (Chung 1977). It is not

Method		GB408 Is there any accusative alignment of flagging? Proto-Polynesian	GB409 Is there any ergative alignment of flagging? Proto-Polynesian	GB409 Is there any ergative alignment of flagging? Proto-Central Pacific
Parsimony (v4.4)	Glottolog	Absent	Present	Absent
Parsimony Gray et al (2009) - MCCT		Absent	Present	Absent
Parsimony Gray et al (2009) - posteriors		Absent	Present	Absent
ML Glottolog (v4.4)		Absent	Present	Absent
ML Gray et al (2009) - MCCT		Half	Present	Absent
ML Gray et al (2009) - posterioris		Half	Present	Absent
Most Common		Half	Present	Present

Table 4: Table showing the results for the features where historical linguists disagree.

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

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, it is very likely that GB409 would be reconstructed as present for Proto-Polynesian. As Clark pointed out, it is the most parsimonious solution. However, it could still have been the case that GB408 (accusative) would have been reconstructed for Proto-Polynesian. The reasons for this may lie in different definitions of what counts as nominative-accusative or neutral in different descriptions, and/or in discussions of plausibility. As has been discussed earlier, it was not possible to include plausibility as a factor in this study.

The proposals of Hale (1968), Hohepa (1967, 1969), and Chung (1978) also involve reconstruction of passive voice that relate 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 records 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 Hale (1968); Hohepa (1967, 1969); Chung (1978).

Concerning the alignment of Proto-Central Pacific, all the results, save the Most Common-model, predict

an absence of ergative-marking. This is likely to be because Rotuman [rotu1241], Wester Fijian [west2519] and Fijian [fiji1243] are all coded as 0 for this feature and they split off early from the Proto-Central Pacific node. This supports the argument put forward by Ball (2007). Similarity to the Polynesian case, given the tree structure it is difficult for the computational approaches to produce another result in lieu of more information on the particulars of the development of alignment systems or possible horizontal transfer.

## 5 Conclusions

In this paper, we have investigated the history of structural features of Oceanic languages to examine how computational reconstructive methods compare to reconstructions by historical linguists, including contributing to the debate on alignment in Oceanic proto-languages.

We have found that computational reconstructions show a high degree of concordance with reconstructions from expert historical linguists. Reconstructions by both Maximum Parsimony and Maximum Likelihood agreed to a very large extent with the findings from historical linguistics. This suggests that the mechanisms at work in historical linguistic reconstruction are similar to the computational methods presented in this paper. This means that we have support for using conclusions regarding rate of change etc from the computational methods to estimate knowledge of historical linguists. This can aid us in making the implicit explicit. Whether or not the predictions from classical historical linguistics are correct or not is a different matter, which should be investigated by assessing the soundness of the principles at play and comparing with other evidence.

The results show that the methods which do not take into account branch lengths (Maximum Parsimony and Most Common) achieve a high concordance with historical linguistics predictions, and the Glottolog-tree typology performed slightly better than the Gray et al. (2009)-results for ML. This is potentially troubling since it seems a sound principle that branch lengths in trees matter. After all, an equal amount of time has past between the existence of a proto-community to today's extant languages. It is possible that some languages are more conservative than others and we should put ancient languages further back in the tree, but the differences in root to tip distances that a tree like Glottolog suggests (see Fig. 8) seem extreme. The Comparative Method is well developed in terms of identifying cognates and sound correspondences, but the estimation of branch lengths has received less attention. This need not mean a return to glottochronology, this appears to be a misguided and overly simplistic approach (Greenhill 2015:371), there are other alternatives.

The future of research on the history of languages probably lies in the combination of human and computational labour. Computational endeavours of curating lexical cognate data (List et al. 2022) and constructing trees (Gray et al. 2009) still rely on teams of expert linguists annotating wordlists for cognacy. Methods are being developed for automatic cognate detection (List et al. 2017), but they are not

yet ready to replace the vast human knowledge and experience of the experts of historical linguistics. However, once cognate classes, regular sound correspondences and structural features are identified the work then turns to subgrouping (constructing trees/networks) and reconstructing earlier states. For these tasks, there are suitable computational methods that can be applied ([Greenhill \(2015\)](#); [Greenhill & Gray \(2009\)](#)). These tasks can be greatly improved and effectivized by computational tools, which in turn can be given sensible priors and parameters to produce viable results.

In order to improve these methods, we could 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. It is no doubt difficult to convey all of this wealth of contextual information in each and every academic paper. However, in order for research to be more easily replicable by others and transparent, more of this knowledge should probably be formalised and/or put in print. Some or all of this kind of information can be incorporated to guide computational methods, for example as priors in models. They should not be given the power to entirely constrain the outcomes, but guide the conclusions the method reaches given the priors and data. In order for this to happen, even more information needs to be made explicit in historical linguistics studies.

The more information is made explicit in methodology the easier it is to assess the soundness of the study, replicate it and improve upon it. This study aims at increasing the transparency of both the principles of reconstruction in classical historical linguistics and the corresponding computational approaches. Hopefully this study (alongside [Carling & Cathcart \(2021\)](#) and [Goldstein \(2022\)](#)) can be a starting point for more joint ventures into our grammatical past.

## References

- Atkinson, Quentin D & Russell D Gray. 2005. Curious parallels and curious connections—phylogenetic thinking in biology and historical linguistics. *Systematic biology* 54(4). 513–526.
- Ball, Douglas. 2007. On ergativity and accusativity in proto-polynesian and proto-central pacific. *Oceanic Linguistics* 128–153.
- Beaulieu, Jeremy M., Jeffrey C. Oliver & Brian O'Meara. 2017. Package ‘corHMM’. URL <https://CRAN.R-project.org/package=corHMM>.
- Blust, Robert. 1996. The Neogrammarian Hypothesis and Pandemic Irregularity. In Mark Durie & Malcolm Ross (eds.), *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, 135–156. Oxford: Oxford University Press.
- Blust, Robert. 2014. Some Recent Proposals Concerning the Classification of the Austronesian Languages. *Oceanic Linguistics* 53(2). 300–391.
- Blust, Robert A. 2004. \*t to k: An Austronesian Sound Change Revisited. *Oceanic Linguistics* 43(2). 365–410.
- Blust, Robert A. 2009. *The Austronesian Languages*. Canberra: Pacific Linguistics.
- Blust, Robert A. & Victoria Chen. 2017. The Pitfalls of Negative Evidence: ‘Nuclear Austronesian’, ‘Ergative Austronesian’, and Their Progeny. *Language and Linguistics* 18(4). 577–621.
- Bouckaert, Remco R & Joseph Heled. 2014. Densitree 2: Seeing trees through the forest. *BioRxiv* 012401.
- Brooks-Bartlett, Jonny. 2018. Probability Concepts Explained: Maximum Likelihood Estimation. *Towards Data Science* URL <https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1>.
- Campbell, Lyle. 1996. On Sound Change and Challenges to Regularity. In Mark Durie & Malcom Ross (eds.), *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, 72–89. Oxford: Oxford University Press.
- Carling, Gerd & Chundra Cathcart. 2021. Reconstructing the evolution of indo-european grammar. *Language* .
- Chang, Will, David Hall, Chundra Cathcart & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the indo-european steppe hypothesis. *Language* 194–244. doi: 10.1353/lan.2015.0005.
- Chung, Sandra. 1977. Review of Clark, R: Aspects of Proto-Polynesian Syntax. *The Journal of the Polynesian Society* 86(4). 537–540.
- Chung, Sandra. 1978. *Case Marking and Grammatical Relations in Polynesian Languages*. Austin: University of Texas.
- Clark, D Ross. 1976. *Aspects of Proto-Polynesian Syntax*, vol. 6. Linguistic Society of New Zealand.
- Crowley, Terry. 1985. Common Noun Phrase Marking in Proto-Oceanic. *Oceanic Linguistics* 24(1/2). 135–193.
- Cunningham, Clifford W, Kevin E Omland & Todd H Oakley. 1998. Reconstructing Ancestral Character States: A Critical Reappraisal. *Trends in Ecology & Evolution* 13(9). 361–366.
- Darwin, Charles. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: Murray.
- Evans, Bethwyn. 2001. *A Study of Valency-Changing Devices in Proto Oceanic*. Research School of Pacific and Asian Studies, The Australian National University. PhD Thesis.
- Felsenstein, Joseph. 2004. *Inferring Phylogenies*, vol. 2. Sunderland, MA: Sinauer Associates.
- Fisher, Ronald A. 1912. On an Absolute Criterion for Fitting Frequency Curves. *Messenger of Mathematics* 41. 155–156.

- Goddard, Ives. 1993. Contamination in Morphological Change in Algonquian Languages. Talk given at Henk Aertsen & Robert J Jeffers (eds.), *Historical Linguistics 1989: Papers from the 9th International Conference on Historical Linguistics, New Brunswick, 14-18 August 1989*, 129–140. John Benjamins Publishing.
- Goldstein, David. 2022. There's no escaping phylogenetics. In Hannes A. Fellner Laura Grestenberger, Charles Reiss & Gabriel Z. Pantillon (eds.), *Ha! linguistic studies in honor of mark r. hale*, 71–91. Wiesbaden: Reichert. URL <https://davidgoldstein.netlify.app/publication/2022-no-escaping-phylogenetics/2022-no-escaping-phylogenetics.pdf>.
- Gray, R.D., A.J. Drummond & S.J. Greenhill. 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science* 323. 479–483.
- Greenhill, Simon. 2015. Evolution and Language: Phylogenetic Analyses. In JD Wright (ed.), *The International Encyclopaedia of the Social and Behavioural Sciences*, Oxford: Elsevier 2nd edn.
- Greenhill, Simon & Russell Gray. 2009. Austronesian Language Phylogenies: Myths and Misconceptions about Bayesian Computational Methods. In Alexander Adelaar & Andrew Pawley (eds.), *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, 375–397. Canberra: Pacific Linguistics.
- Greenhill, Simon J., Robert Andrew Blust & Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics. *Evolutionary Bioinformatics* 4. 271–283.
- Greenhill, Simon J & Ross Clark. 2011. POLLEX-Online: The Polynesian Lexicon Project Online. *Oceanic Linguistics* 50(2). 551–559.
- Hale, Kenneth. 1968. Review of Hohepa 1967. *Journal of the Polynesian Society* 77(1). 83–99.
- Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019. Glottolog/Glottolog: Glottolog Database 4.0. URL <https://doi.org/10.5281/zenodo.3260726>.
- Hohepa, Patrick W. 1967. *A Profile Generative Grammar of Maori*. Indiana University Publications in Anthropology and Linguistics, Memoir 20.
- Hohepa, Patrick W. 1969. The Accusative-to-Ergative Drift in Polynesian Languages. *Journal of the Polynesian Society* 78(3). 295–329.
- Huelsenbeck, John P, Rasmus Nielsen & Jonathan P Bollback. 2003. Stochastic mapping of morphological characters. *Systematic biology* 52(2). 131–158.
- Jäger, Gerhard & Johann-Mattis List. 2018. Using Ancestral State Reconstruction Methods for Onomasiological Reconstruction in Multilingual Word Lists. *Language Dynamics and Change* 8(1). 22–54.
- Jombart, Thibaut, Stéphane Dray & Anders Ellern Bilgrau. 2017. Package ‘Adephylo’. URL <https://cran.r-project.org/web/packages/adephylo/index.html>.
- Jonsson, Niklas. 1998. Det Polynesiska Verbmorfemet - Cia; Om Dess Funktion i Samoanska.
- Joy, Jeffrey B, Richard H Liang, Rosemary M McCloskey, T Nguyen & Art FY Poon. 2016. Ancestral Reconstruction. *PLoS Computational Biology* 12(7). e1004763.
- Kikusawa, Ritsuko. 2001. Rotuman and Fijian Case-Marking Strategies and Their Historical Development. *Oceanic Linguistics* 40(1). 85–111.
- Kikusawa, Ritsuko. 2002. *Proto Central Pacific Ergativity: Its Reconstruction and Development in the Fijian, Rotuman and Polynesian Languages*. Canberra: Pacific Linguistics.
- Kikusawa, Ritsuko. 2006. On the Development of Number Systems in Oceanic Pronouns. Talk given at 6th International Conference on Oceanic Linguistics (COOL6), Port Vila, Vanuatu.
- Lamarck, Jean-Baptiste. 1809. Philosophie zoologique, ou exposition des considérations relatives à l'histoire naturelle des animaux, 2 voll., dentu, paris; trad. it. parz. *Filosofia zoologica. Prima parte*.
- Liggett, Thomas Milton. 2010. *Continuous time markov processes: an introduction*, vol. 113. American Mathematical Soc.

- List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch & Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Sci Data* .
- List, Johann-Mattis, Simon J Greenhill & Russell D Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLoS one* 12(1). e0170046.
- Louca, Stilianos & Michael Doebeli. 2017. Efficient Comparative Phylogenetics on Large Trees. *Bioinformatics* 34(6). 1053–1055.
- Lynch, John, Malcolm Ross & Terry Crowley. 2011. Proto Oceanic. In John Lynch, Malcolm Ross & Terry Crowley (eds.), *The Oceanic Languages* Curzon Language Family Series, 54–91. Richmond: Curzon 2nd edn.
- MacLaurin, James & Kim Sterelny. 2008. *What Is Biodiversity?* University of Chicago Press.
- Marck, Jeffrey C. 2000. Polynesian Languages. In J. Garry & C. Rubino (eds.), *Facts About the World's Languages: An Encyclopaedia of the World's Major Languages, Past and Present*, New York: H.W. Wilson.
- Pagel, Mark. 1994. Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255(1342). 37–45.
- Pagel, Mark, Andrew Meade & Daniel Barker. 2004. Bayesian estimation of ancestral character states on phylogenies. *Systematic biology* 53(5). 673–684.
- Paradis, Emmanuel, Julien Claude & Korbinian Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R Language. *Bioinformatics* 20(2). 289–290.
- Pawley, Andrew. 1970. Grammatical Reconstruction and Change in Polynesia and Fiji. In SA Wurm & DC Laycock (eds.), *Studies in Honour of Arthur Capell*, 301–368. Canberra: Pacific Linguistics.
- Pawley, Andrew. 1973. Some Problems in Proto-Oceanic Grammar. *Oceanic Linguistics* 12(1/2). 103–188.
- Pawley, Andrew. 2005. The Meaning(s) of Proto Oceanic\* Panua. In D.A. Counts C. Gross, H.D. Lyons (ed.), *A Polymath Anthropologist: Essays in Honour of Ann Chowning* (Research in Anthropology and Linguistics Monograph 6), 133–145. Auckland: Department of Anthropology, University of Auckland.
- R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. URL <https://www.R-project.org/>.
- Revell, Liam J. 2012. Phytools: An R Package for Phylogenetic Comparative Biology (and Other Things). *Methods in Ecology and Evolution* 3(2). 217–223.
- Revell, Liam J. 2014. Ancestral State Reconstruction. Talk given at *AnthroTree Workshop 2014*, Duke University, Durham, NC. URL <http://www.phytools.org/anthrotree/ancestral-states.pdf>.
- Ronquist, Fredrik. 2004. Bayesian inference of character evolution. *Trends in ecology & evolution* 19(9). 475–481. doi: 10.1016/j.tree.2004.07.002.
- Ross, Malcolm, Andrew Pawley & Meredith Osmond (eds.). 1998. *Material Culture*, vol. 1 The Lexicon of Proto Oceanic: The Culture and Environment of Ancestral Oceanic Society. Canberra: ANU Electronic Press.
- Ross, Malcolm, Andrew Pawley & Meredith Osmond (eds.). 2007. *The Physical Environment*, vol. 2 The Lexicon of Proto Oceanic: The Culture and Environment of Ancestral Oceanic Society. Canberra: ANU Electronic Press 2nd edn.
- Ross, Malcolm, Andrew Pawley & Meredith Osmond (eds.). 2008. *Plants*, vol. 3 The Lexicon of Proto Oceanic: The Culture and Environment of Ancestral Oceanic Society. Canberra: ANU Electronic Press.
- Ross, Malcolm, Andrew Pawley & Meredith Osmond (eds.). 2011. *Animals*, vol. 4 The Lexicon of Proto Oceanic: The Culture and Environment of Ancestral Oceanic Society. Canberra: ANU Electronic Press.

- Ross, Malcolm, Andrew Pawley & Meredith Osmond (eds.). 2016. *People: Body and Mind*, vol. 5 The Lexicon of Proto Oceanic: The Culture and Environment of Ancestral Oceanic Society. Canberra: ANU Electronic Press.
- Ross, Malcolm D. 2004. The Morphosyntactic Typology of Oceanic Languages. *Language and Linguistics* 5(2). 491–541.
- Ross, Malcolm D. 2007. Two Kinds of Locative Construction in Oceanic Languages: A Robust Distinction. In Jeff Siegel, John Lynch & Diana Eades (eds.), *Language Description, History and Development: Linguistic Indulgence in Memory of Terry Crowley*, 281–295. John Benjamins Publishing Company.
- Sankoff, David. 1975. Minimal Mutation Trees of Sequences. *SIAM Journal on Applied Mathematics* 28(1). 35–42.
- Schlegel, F. 1808. *Über die sprache und weisheit der indier: Ein beitrag zur begründung der alterthumskunde*. Heidelberg: Mohr und Zimmer.
- Schliep, Klaus, Alastair Potts, Tim W. White, Cyrill Stachniss, Michelle Kendall, Keren Halabi, Richel Bilderbeek & Kristin Winchell. 2019. Package ‘Phangorn’. URL <https://CRAN.R-project.org/package=phangorn>.
- Siewierska, Anna. 2013. Gender Distinctions in Independent Personal Pronouns. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. URL <https://wals.info/chapter/44>.
- Slingerland, Edward, Quentin D Atkinson, Carol R Ember, Oliver Sheehan, Michael Muthukrishna, Joseph Bulbulia & Russell D Gray. 2020. Coding culture: challenges and recommendations for comparative cultural databases. *Evolutionary Human Sciences* 2.
- Stassen, Leon. 2011. Predicative Possession. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library. URL <http://wals.info/chapter/117>.
- The Grambank Consortium (ed.). 2021. *Grambank*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at <http://grambank.clld.org>, Accessed on 2021-07-01.
- Van Rijsbergen, Cornelis Joost. 1979. *Information Retrieval*. Butterworths.
- Walkden, George. 2013. The correspondence problem in syntactic reconstruction. *Diachronica. International Journal for Historical Linguistics. Founded by E.F.K. Koerner, General Editor, 1984–2001* 30(1). 95–122. doi: 10.1075/dia.30.1.04wal. URL <https://doi.org/10.1075/dia.30.1.04wal>.
- Wickham, Hadley. 2020. reshape2: Flexibly reshape data: a reboot of the reshape package, version 1.4.4. See <https://cran.r-project.org/web/packages/reshape2/index.html>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo & Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43). 1686. doi: 10.21105/joss.01686.
- Wilks, Samuel S. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics* 9(1). 60–62.
- Yang, Ziheng. 2006. *Computational Molecular Evolution*. Oxford University Press.

## Appendices

### A Data availability

The Grambank-dataset (version 1.0) will be published shortly and can be provided for anonymous reviewers under the condition that it is not shared before official publication. All code relevant to this paper will be published as a Git repository and how-ever the journal might want it published. The Git repository is private and is possible for reviewers to access under certain restrictions before publication.

### B Technical details of methods

For Maximum Parsimony, we are using the function `asr_max_parsimony()` from the R-package `castor` (Louca & Doebeli 2017) (which is an instantiation of the method described in Sankoff (1975)) for calculating ancestral states and stability of features. This function produces ancestral states for all nodes and reports the number of changes that was minimally required for each feature.

Ancestral state reconstruction using Maximum Likelihood Estimation involves computing each ancestral state from the tips up to the root taking into account branch lengths and the joint likelihood of states given all nodes in the tree (Wilks 1938; Fisher 1912; Pagel 1994; Cunningham et al. 1998). The Maximum Likelihood Estimation function takes a set of observations and computes the parameter distribution that maximises the likelihood given the observed data<sup>18</sup>. This means that for every split in the tree — every ancestral node — the Maximum Likelihood Estimation function computes what is the most likely distribution at that point given the nature of the entire tree. ML can be modified so that it allows for different rates of change. An Equal Rates (ER) model assumes that the chance of transition from state A to state B and from B to A are equal. However, we as linguists are aware that certain features are more likely to be lost than gained so this is not a reasonable assumption. Therefore, we allow the model to estimate different transition rates for going from A to B and from B to A given the data. This is known as “All Rates are Different” (ARD).

When estimating ancestral states with ML, it is possible to either a) find the state at each node that maximises the likelihood (integrating over all other states at all nodes, in proportion to their probability) at that particular node (marginal reconstruction), or b) find the set of character states at all nodes that (jointly) maximize the likelihood of the entire tree (joint reconstruction). We are using marginal reconstruction in this study since it is the recommended way to deal with uncertainty in reconstruction (Revell 2014). These two methods often yield similar results, but can differ, see Felsenstein (2004:259-260), Yang (2006:121-126) and Joy et al. (2016:5) for more details. For our data, a trial run of joint reconstruction did not generate drastically different outcomes.

For this study, the function `corHMM` from the R-package `corHMM` (Beaulieu et al. 2017) is used for marginal reconstruction of ancestral states and rates of change per feature.

Languages with missing data were pruned away in all analysis, no hidden state reconstruction of values at tips was preformed. The match between Glottolog and Grambank is 226, the match between Gray et al. (2009) and Grambank is 112. For the parsimony analysis of each feature, languages with missing data were dropped from the trees in the analysis for that feature. Features which could only be assigned values for less than half of the languages in the tree were excluded from the analysis.

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

### C Mathematics of the F1-score including half-results

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

---

<sup>18</sup>For a gentle introduction to the concept of Maximum Likelihood Estimation, see Brooks-Bartlett (2018).

### C.1 Standard definitions

The F1-score is the harmonic mean of precision and recall (Van Rijsbergen 1979).

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

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

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

### C.2 Half-result definitions of precision and recall

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

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

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

### C.3 The question

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

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

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

$$2 \times \frac{\text{precision}_{\text{half}} \times \text{recall}_{\text{half}}}{\text{precision}_{\text{half}} + \text{recall}_{\text{half}}} \stackrel{?}{=} \frac{\text{TP} + \frac{H}{2}}{\text{TP} + \frac{1}{2} \times (\text{FP} + \text{FN}) + H} \quad (5)$$

### C.4 The proof

We will expand the left-hand side of (5) and show it is equal to the right-hand side. Let's forget about the  $2 \times$  for now (we will reintroduce it at the end). Expanding the numerator gives:

$$\frac{(\text{TP} + \frac{H}{2})(\text{TP} + \frac{H}{2})}{(\text{TP} + \text{FP} + H)(\text{TP} + \text{FN} + H)}$$

Expanding the denominator gives:

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

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

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

Finally, we bring back the  $2 \times$  from the beginning:

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

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

## D Binarisation of the Grambank features

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

## E Grambank features

Feature_ID	Name
GB024a	Is the order of the numeral and noun Num-N?
GB024b	Is the order of the numeral and noun N-Num?
GB025a	Is the order of the adnominal demonstrative and noun Dem-N?
GB025b	Is the order of the adnominal demonstrative and noun N-Dem?
GB065a	Is the pragmatically unmarked order of adnominal possessor noun and possessed noun PSR-PSD?
GB065b	Is the pragmatically unmarked order of adnominal possessor noun and possessed noun PSD-PSR?
GB130a	Is the pragmatically unmarked order of S and V in intransitive clauses S-V?
GB130b	Is the pragmatically unmarked order of S and V in intransitive clauses V-S?
GB193a	Is the order of the adnominal property word (ANM) and noun ANM-N?
GB193b	Is the order of the adnominal property word (ANM) and noun N-ANM?
GB203a	Is the order of the adnominal collective universal quantifier (UQ) and noun UQ-N?
GB203b	Is the order of the adnominal collective universal quantifier (UQ) and noun N-QU?
GB020	Are there definite or specific articles?
GB021	Do indefinite nominals commonly have indefinite articles?
GB022	Are there prenominal articles?
GB023	Are there postnominal articles?
GB026	Can adnominal property words occur discontinuously?
GB027	Are nominal conjunction and comitative expressed by different elements?
GB028	Is there a distinction between inclusive and exclusive?
GB030	Is there a gender distinction in independent 3rd person pronouns?
GB031	Is there a dual or unit augmented form (in addition to plural or augmented) for all person categories in the pronoun system?
GB035	Are there three or more distance contrasts in demonstratives?
GB036	Do demonstratives show an elevation distinction?

- GB037      Do demonstratives show a visible-nonvisible distinction?
- GB038      Are there demonstrative classifiers?
- GB039      Is there nonphonological allomorphy of noun number markers?
- GB041      Are there several nouns (more than three) which are suppletive for number?
- GB042      Is there productive overt morphological singular marking on nouns?
- GB043      Is there productive morphological dual marking on nouns?
- GB044      Is there productive morphological plural marking on nouns?
- GB046      Is there an associative plural marker for nouns?
- GB047      Is there a productive morphological pattern for deriving an action/state noun from a verb?
- GB048      Is there a productive morphological pattern for deriving an agent noun from a verb?
- GB049      Is there a productive morphological pattern for deriving an object noun from a verb?
- GB051      Is there a gender/noun class system where sex is a factor in class assignment?
- GB052      Is there a gender/noun class system where shape is a factor in class assignment?
- GB053      Is there a gender/noun class system where animacy is a factor in class assignment?
- GB054      Is there a gender/noun class system where plant status is a factor in class assignment?
- GB057      Are there numeral classifiers?
- GB058      Are there possessive classifiers?
- GB059      Is the adnominal possessive construction different for alienable and inalienable nouns?
- GB068      Do core adjectives (defined semantically as property concepts such as value, shape, age, dimension) act like verbs in predicative position?
- GB069      Do core adjectives (defined semantically as property concepts; value, shape, age, dimension) used attributively require the same morphological treatment as verbs?
- GB070      Are there morphological cases for non-pronominal core arguments (i.e. S/A/P)?

- GB071 Are there morphological cases for pronominal core arguments (i.e. S/A/P)?
- GB072 Are there morphological cases for oblique non-pronominal NPs (i.e. not S/A/P)?
- GB073 Are there morphological cases for independent oblique personal pronominal arguments (i.e. not S/A/P)?
- GB074 Are there prepositions?
- GB075 Are there postpositions?
- GB079 Do verbs have prefixes/proclitics, other than those that only mark A, S or P (do include portmanteau: A & S + TAM)?
- GB080 Do verbs have suffixes/enclitics, other than those that only mark A, S or P (do include portmanteau: A & S + TAM)?
- GB081 Is there productive infixation in verbs?
- GB082 Is there overt morphological marking of present tense on verbs?
- GB083 Is there overt morphological marking on the verb dedicated to past tense?
- GB084 Is there overt morphological marking on the verb dedicated to future tense?
- GB086 Is a morphological distinction between perfective and imperfective aspect available on verbs?
- GB089 Can the S argument be indexed by a suffix/enclitic on the verb in the simple main clause?
- GB090 Can the S argument be indexed by a prefix/proclitic on the verb in the simple main clause?
- GB091 Can the A argument be indexed by a suffix/enclitic on the verb in the simple main clause?
- GB092 Can the A argument be indexed by a prefix/proclitic on the verb in the simple main clause?
- GB093 Can the P argument be indexed by a suffix/enclitic on the verb in the simple main clause?
- GB094 Can the P argument be indexed by a prefix/proclitic on the verb in the simple main clause?
- GB095 Are variations in marking strategies of core participants based on TAM distinctions?
- GB096 Are variations in marking strategies of core participants based on verb classes?

GB098	Are variations in marking strategies of core participants based on person distinctions?
GB099	Can verb stems alter according to the person of a core participant?
GB103	Is there a benefactive applicative marker on the verb (including indexing)?
GB104	Is there an instrumental applicative marker on the verb (including indexing)?
GB105	Can the recipient in a ditransitive construction be marked like the monotransitive patient?
GB107	Can standard negation be marked by an affix, clitic or modification of the verb?
GB108	Is there directional or locative morphological marking on verbs?
GB109	Is there verb suppletion for participant number?
GB110	Is there verb suppletion for tense or aspect?
GB111	Are there conjugation classes?
GB113	Are there verbal affixes or clitics that turn intransitive verbs into transitive ones?
GB114	Is there a phonologically bound reflexive marker on the verb?
GB115	Is there a phonologically bound reciprocal marker on the verb?
GB116	Do verbs classify the shape, size or consistency of absolutive arguments by means of incorporated nouns, verbal affixes or suppletive verb stems?
GB117	Is there a copula for predicate nominals?
GB118	Are there serial verb constructions?
GB119	Can mood be marked by an inflecting word ("auxiliary verb")?
GB120	Can aspect be marked by an inflecting word ("auxiliary verb")?
GB121	Can tense be marked by an inflecting word ("auxiliary verb")?
GB122	Is verb compounding a regular process?
GB123	Are there verb-adjunct (aka light-verb) constructions?
GB124	Is incorporation of nouns into verbs a productive intransitivizing process?
GB126	Is there an existential verb?
GB127	Are different posture verbs used obligatorily depending on an inanimate locatum's shape or position (e.g. 'to lie' vs. 'to stand')?

- GB129      Is there a notably small number, i.e. about 100 or less, of verb roots in the language?
- GB131      Is a pragmatically unmarked constituent order verb-initial for transitive clauses?
- GB132      Is a pragmatically unmarked constituent order verb-medial for transitive clauses?
- GB133      Is a pragmatically unmarked constituent order verb-final for transitive clauses?
- GB134      Is the order of constituents the same in main and subordinate clauses?
- GB135      Do clausal objects usually occur in the same position as nominal objects?
- GB136      Is the order of core argument (i.e. S/A/P) constituents fixed?
- GB137      Can standard negation be marked clause-finally?
- GB138      Can standard negation be marked clause-initially?
- GB139      Is there a difference between imperative (prohibitive) and declarative negation constructions?
- GB140      Is verbal predication marked by the same negator as all of the following types of predication: locational, existential and nominal?
- GB146      Is there a morpho-syntactic distinction between predicates expressing controlled versus uncontrolled events or states?
- GB147      Is there a morphological passive marked on the lexical verb?
- GB148      Is there a morphological antipassive marked on the lexical verb?
- GB149      Is there a morphologically marked inverse on verbs?
- GB150      Is there clause chaining?
- GB151      Is there an overt verb marker dedicated to signalling coreference or noncoreference between the subject of one clause and an argument of an adjacent clause ("'''switch reference'''")?
- GB152      Is there a morphologically marked distinction between simultaneous and sequential clauses?
- GB155      Are causatives formed by affixes or clitics on verbs?
- GB156      Is there a causative construction involving an element that is unmistakably grammaticalized from a verb for 'to say'?
- GB158      Are verbs reduplicated?

GB159	Are nouns reduplicated?
GB160	Are elements apart from verbs or nouns reduplicated?
GB165	Is there productive morphological trial marking on nouns?
GB166	Is there productive morphological paucal marking on nouns?
GB167	Is there a logophoric pronoun?
GB170	Can an adnominal property word agree with the noun in gender/noun class?
GB171	Can an adnominal demonstrative agree with the noun in gender/noun class?
GB172	Can an article agree with the noun in gender/noun class?
GB177	Can the verb carry a marker of animacy of argument, unrelated to any gender/noun class of the argument visible in the NP domain?
GB184	Can an adnominal property word agree with the noun in number?
GB185	Can an adnominal demonstrative agree with the noun in number?
GB186	Can an article agree with the noun in number?
GB187	Is there any productive diminutive marking on the noun (exclude marking by system of nominal classification only)?
GB188	Is there any productive augmentative marking on the noun (exclude marking by system of nominal classification only)?
GB192	Is there a gender system where a noun's phonological properties are a factor in class assignment?
GB196	Is there a male/female distinction in 2nd person independent pronouns?
GB197	Is there a male/female distinction in 1st person independent pronouns?
GB198	Can an adnominal numeral agree with the noun in gender/noun class?
GB204	Do collective ('all') and distributive ('every') universal quantifiers differ in their forms or their syntactic positions?
GB250	Can predicative possession be expressed with a transitive 'habeo' verb?
GB252	Can predicative possession be expressed with an S-like possessum and a locative-coded possessor?
GB253	Can predicative possession be expressed with an S-like possessum and a dative-coded possessor?
GB254	Can predicative possession be expressed with an S-like possessum and a possessor that is coded like an adnominal possessor?

GB256	Can predicative possession be expressed with an S-like possessor and a possessum that is coded like a comitative argument?
GB257	Can polar interrogation be marked by intonation only?
GB260	Can polar interrogation be indicated by a special word order?
GB262	Is there a clause-initial polar interrogative particle?
GB263	Is there a clause-final polar interrogative particle?
GB264	Is there a polar interrogative particle that most commonly occurs neither clause-initially nor clause-finally?
GB265	Is there a comparative construction that includes a form that elsewhere means 'surpass, exceed'?
GB266	Is there a comparative construction that employs a marker of the standard which elsewhere has a locational meaning?
GB270	Can comparatives be expressed using two conjoined clauses?
GB273	Is there a comparative construction with a standard marker that elsewhere has neither a locational meaning nor a 'surpass / exceed' meaning?
GB275	Is there a bound comparative degree marker on the property word in a comparative construction?
GB276	Is there a non-bound comparative degree marker modifying the property word in a comparative construction?
GB285	Can polar interrogation be marked by a question particle and verbal morphology?
GB286	Can polar interrogation be indicated by overt verbal morphology only?
GB291	Can polar interrogation be marked by tone?
GB296	Is there a phonologically or morphosyntactically definable class of ideophones that includes ideophones depicting imagery beyond sound?
GB297	Can polar interrogation be indicated by a V-not-V construction?
GB298	Can standard negation be marked by an inflecting word ("'''auxiliary verb'''")?
GB299	Can standard negation be marked by a non-inflecting word ("'''auxiliary particle'''")?
GB300	Does the verb for 'give' have suppletive verb forms?
GB301	Is there an inclusory construction?

- GB302 Is there a phonologically free passive marker ("""particle"""" or """auxiliary""")?
- GB303 Is there a phonologically free antipassive marker ("""particle"""" or """auxiliary""")?
- GB304 Can the agent be expressed overtly in a passive clause?
- GB305 Is there a phonologically independent reflexive pronoun?
- GB306 Is there a phonologically independent non-bipartite reciprocal pronoun?
- GB309 Are there multiple past or multiple future tenses, distinguishing distance from Time of Reference?
- GB312 Is there overt morphological marking on the verb dedicated to mood?
- GB313 Are there special adnominal possessive pronouns that are not formed by an otherwise regular process?
- GB314 Can augmentative meaning be expressed productively by a shift of gender/noun class?
- GB315 Can diminutive meaning be expressed productively by a shift of gender/noun class?
- GB316 Is singular number regularly marked in the noun phrase by a dedicated phonologically free element?
- GB317 Is dual number regularly marked in the noun phrase by a dedicated phonologically free element?
- GB318 Is plural number regularly marked in the noun phrase by a dedicated phonologically free element?
- GB319 Is trial number regularly marked in the noun phrase by a dedicated phonologically free element?
- GB320 Is paucal number regularly marked in the noun phrase by a dedicated phonologically free element?
- GB321 Is there a large class of nouns whose gender/noun class is not phonologically or semantically predictable?
- GB322 Is there grammatical marking of direct evidence (perceived with the senses)?
- GB323 Is there grammatical marking of indirect evidence (hearsay, inference, etc.)?
- GB324 Is there an interrogative verb for content interrogatives (who?, what?, etc.)?
- GB325 Is there a count/mass distinction in interrogative quantifiers?

GB326	Do (nominal) content interrogatives normally or frequently occur in situ?
GB327	Can the relative clause follow the noun?
GB328	Can the relative clause precede the noun?
GB329	Are there internally-headed relative clauses?
GB330	Are there correlative relative clauses?
GB331	Are there non-adjacent relative clauses?
GB333	Is there a decimal numeral system?
GB334	Is there synchronic evidence for any element of a quinary numeral system?
GB335	Is there synchronic evidence for any element of a vigesimal numeral system?
GB336	Is there a body-part tallying system?
GB400	Are all person categories neutralized in some voice, tense, aspect, mood and/or negation?
GB401	Is there a class of patient-labile verbs?
GB402	Does the verb for 'see' have suppletive verb forms?
GB403	Does the verb for 'come' have suppletive verb forms?
GB408	Is there any accusative alignment of flagging?
GB409	Is there any ergative alignment of flagging?
GB410	Is there any neutral alignment of flagging?
GB415	Is there a politeness distinction in 2nd person forms?
GB421	Is there a proposed complementizer in complements of verbs of thinking and/or knowing?
GB422	Is there a postposed complementizer in complements of verbs of thinking and/or knowing?
GB430	Can adnominal possession be marked by a prefix on the possessor?
GB431	Can adnominal possession be marked by a prefix on the possessed noun?
GB432	Can adnominal possession be marked by a suffix on the possessor?
GB433	Can adnominal possession be marked by a suffix on the possessed noun?
GB519	Can mood be marked by a non-inflecting word (""" auxiliary particle """)?
GB520	Can aspect be marked by a non-inflecting word (""" auxiliary particle """)?
GB521	Can tense be marked by a non-inflecting word (""" auxiliary particle """)?

GB522      Can the S or A argument be omitted from a pragmatically unmarked clause when the referent is inferrable from context ("pro-drop" or "null anaphora")?

---

Table 5: Table of Grambank fatures

## F Further details on the tree phylogeny

The tree from Gray et al. (2009) contains duplicates (see for example Nakanai). This is because it is a tree of word-lists for languages (doculects) rather than languages themselves. There are also some instances where multiple dialects of one language is included. For the analysis, only one tip per language was retained, based on which had best coverage in the underlying data for the tree (i.e. the Austronesian Basic Vocabulary Database, ABVD (Greenhill et al. 2008)). This means that duplicate languages were reduced to one, and only one dialect per language retained if there was more than one.

For both Maximum Parsimony and Maximum Likelihood the tree were first pruned down to only languages where there is data in Grambank. For the further Maximum Parsimony analysis, tips representing languages where the data for a particular feature was missing were dropped for the analysis of that feature. For the Maximum Likelihood analysis, the value at such tips was converted to ambiguous. This is necessary because otherwise the rate of change across the different features would not be comparable in the Maximum Likelihood analysis.

## G Further details on the Grambank coding of proto-languages

Another example of how information in the publications was turned into Grambank feature coding relates to verbal markers encoding subjects and objects, as proposed by Lynch et al. (2011) among others. In their book, there is a paper on reconstructions of grammar for Proto-Oceanic and in the section on the basic verb phrase we find the statement below:

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

Lynch et al. (2011:83)

This statement, together with a verb schema provided in the section, support the notion that Proto-Oceanic had subject proclitics and object enclitics. We can also infer from this publication as a whole that the authors believe Proto-Oceanic in fact did *not* have subject enclitics and object proclitics. This second prediction relies on absence of evidence and is less strong than the first, but given that the whole paper is void of any description of object proclitics or subject enclitics being a possibility (including the verb schema) and argument structure is well-discussed, we may dare to make this leap. This information can be translated into the Grambank questionnaire by positing absence and presence for the six relevant features that concern argument marking on the verb (where S stands for subject of intransitive, A for subject of transitive and O for object; see table 6).

Table 6: Example of predictions from historical linguistics as rendered in Grambank features.

Grambank ID	Question	Proto-language	Expert prediction	Reference
GB089	Can the S argument be indexed by a suffix/enclitic on the verb in the simple main clause?	Proto-Oceanic	Absent	Ross (2004:498-499), Lynch et al. (2011:83)
GB090	Can the S argument be indexed by a prefix/proclitic on the verb in the simple main clause?	Proto-Oceanic	Present	Ross (2004:498-499), Lynch et al. (2011:83)
GB091	Can the A argument be indexed by a suffix/enclitic on the verb in the simple main clause?	Proto-Oceanic	Absent	Ross (2004:498-499), Lynch et al. (2011:83)
GB092	Can the A argument be indexed by a prefix/proclitic on the verb in the simple main clause?	Proto-Oceanic	Present	Ross (2004:498-499), Lynch et al. (2011:83)
GB093	Can the P argument be indexed by a suffix/enclitic on the verb in the simple main clause?	Proto-Oceanic	Present	Ross (2004:498-499), Lynch et al. (2011:83)
GB094	Can the P argument be indexed by a prefix/proclitic on the verb in the simple main clause?	Proto-Oceanic	Absent	Ross (2004:498-499), Lynch et al. (2011:83)

## H Table of historical linguistics sources surveyed

Table 7: Table of historical linguistics publications used in this dissertation for Proto-Oceanic grammar

Citation	Title	Proto-Languages	Domains
Pawley (1970)	Grammatical reconstruction and change on Polynesia and Fiji	Proto-Central Pacific	Verbal markers and aspect particles
Pawley (1973)	Some problems in Proto-Oceanic	Proto-Oceanic and Proto-Polynesian	Possession, noun phrase marking, negation, verbal markers, clusivity, word order

Citation	Title	Proto-Languages	Domains
Clark (1976)	Aspects of Proto-Polynesian syntax	Proto-Oceanic and Proto-Polynesian	Alignment, negation, word order, possession, noun phrase marking, voice
Chung (1978)	Case marking and grammatical relations in Polynesian languages	Proto-Polynesian	Alignment, word order, voice, noun phrase marking
Crowley (1985)	Common noun phrase marking in Proto-Oceanic	Proto-Oceanic	noun phrase marking, clusivity
Jonsson (1998)	Det polynesiska verbmorfemet - <i>Cia</i> ; om dess funktion i Samoanska	Proto-Polynesian	Verbal marker
March (2000)	Polynesian languages (in Facts About the World's Languages: An encyclopaedia of the world's major languages, past and present)	Proto-Central Pacific and Proto-Polynesian	Word order, verbal markers, possession, clusivity
Evans (2001)	A study of valency-changing devices in Proto Oceanic	Proto-Oceanic	Verbal markers
Ball (2007)	On ergativity and accusativity in Proto-Polynesian and proto-Central Pacific	Proto-Polynesian	Alignment, voice
Kikusawa (2001)	Rotuman and Fijian case-marking strategies and their historical development	Proto-Oceanic	Possession, pronominal number
Kikusawa (2002)	Proto Central Pacific ergativity: Its reconstruction and development in the Fijian, Rotuman and Polynesian languages	Proto-Central Pacific	Alignment, word order
Lynch et al. (2011)	The Oceanic Languages, paper 4: Proto-Oceanic	Proto-Oceanic, Proto-Central Pacific and Proto-Polynesian	Negation, word order, verbal markers, clusivity, possession, pronominal number, polar interrogation, nominalisations and more
Ross (2004) <sup>19</sup>	The morphosyntactic typology of Oceanic languages	Proto-Oceanic and Proto-Polynesian	alignment, word order, verbal markers, possession, noun phrase marking

## I Table of new predictions

<sup>19</sup>This paper makes statements about “canonical” Oceanic languages, which is technically different from *reconstruction* of Proto-Oceanic. However, the author does state that the “canonic type is probably also a reflection of the morphosyntax of Proto Oceanic” (Ross 2004:492) and has given personal approval for the paper to be included in this study in this manner.



GB130b	Proto-Eastern Polynesian	Present	Present
GB193b	Proto-Central Pacific	Present	Present
GB193b	Proto-Eastern Polynesian	Present	Present
GB203b	Proto-Oceanic	Present	Present
GB203b	Proto-Central Pacific	Present	Present
GB203b	Proto-Polynesian	Present	Present
GB203b	Proto-Eastern Polynesian	Present	Present
GB020	Proto-Eastern Polynesian	Present	Present
GB021	Proto-Eastern Polynesian	Present	Present
GB022	Proto-Eastern Polynesian	Present	Present
GB027	Proto-Oceanic	Present	Present
GB028	Proto-Central Pacific	Present	Present
GB028	Proto-Eastern Polynesian	Present	Present
GB031	Proto-Central Pacific	Present	Present
GB031	Proto-Eastern Polynesian	Present	Present
GB035	Proto-Central Pacific	Present	Present
GB035	Proto-Eastern Polynesian	Present	Present



GB124	Proto-Central Pacific	Present	Present	Present	Present	Present
GB124	Proto-Polynesian	Present	Present	Present	Present	Present
GB126	Proto-Oceanic	Present	Present	Present	Present	Present
GB126	Proto-Central Pacific	Present	Present	Present	Present	Present
GB126	Proto-Polynesian	Present	Present	Present	Present	Present
GB126	Proto-Eastern Polynesian	Present	Present	Present	Present	Present
GB126	Proto-Eastern Polynesian	Present	Present	Present	Present	Present
GB131	Proto-Oceanic	Present	Present	Present	Present	Present
GB134	Proto-Central Pacific	Present	Present	Present	Present	Present
GB134	Proto-Polynesian	Present	Present	Present	Present	Present
GB134	Proto-Eastern Polynesian	Present	Present	Present	Present	Present
GB134	Proto-Oceanic	Present	Present	Present	Present	Present
GB135	Proto-Central Pacific	Present	Present	Present	Present	Present
GB135	Proto-Polynesian	Present	Present	Present	Present	Present
GB135	Proto-Eastern Polynesian	Present	Present	Present	Present	Present
GB135	Proto-Oceanic	Present	Present	Present	Present	Present
GB139	Proto-Central Pacific	Present	Present	Present	Present	Present
GB139	Proto-Eastern Polynesian	Present	Present	Present	Present	Present





GB333	Proto-Eastern Polynesian	Present	Present	Present	Present	Present
GB421	Proto-Central Pacific	Present	Present	Present	Present	Half
GB433	Proto-Central Pacific	Present	Present	Present	Present	Absent
GB519	Proto-Central Pacific	Present	Present	Present	Present	Present
GB519	Proto-Eastern Polynesian	Present	Present	Present	Present	Present
GB520	Proto-Eastern Polynesian	Present	Present	Present	Present	Present
GB521	Proto-Polynesian	Present	Present	Present	Present	Present
GB521	Proto-Eastern Polynesian	Present	Present	Present	Present	Present
GB522	Proto-Eastern Polynesian	Present	Present	Present	Present	Present

Table 8: Table showing predictions where the six MP and ML methods agree on presence