Thesis submitted in partial fulfillment of the requirements for the degree of Master of Arts in Computational Linguistics

A feature-based neural model of sound change informed by global lexicostatistical data

Author:
Arne Rubehn

December 2022

First Examiner:
Dr. Johannes Dellert

Second Examiner:
Prof. Dr. Gerhard Jäger

Eberhard Karls Universität Tübingen
Philosophische Fakultät
Seminar für Sprachwissenschaft

Antiplagiatserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst habe, dass ich keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe, dass ich alle wörtlich oder sinngemäß aus anderen Werken übernommenen Aussagen als solche gekennzeichnet habe, dass die Arbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens gewesen ist, dass ich die Arbeit weder vollständig noch in wesentlichen Teilen bereits veröffentlicht habe, und dass das in Dateiform eingereichte Exemplar mit dem eingereichten gebundenen Exemplar übereinstimmt.

Tübingen, den _____ _____

Arne Rubehn

Abstract

Historical linguists have successfully reconstructed numerous unattested ancestral languages for over a century, mainly by applying the comparative method, a powerful procedure for recovering extinct languages and understanding how they developed into their modern daughter languages. With the exponential rise of computational power, scholars have been trying to develop computational solutions for tasks in historical linguistics for roughly two decades. The success of these methods, however, is limited to solving some individual tasks satisfyingly, while there are still no good solutions for other tasks. Part of the reason why scholars were not able to find good computational methods for some parts of the comparative method is that historical linguists often rely on their intuition and general linguistic knowledge when reconstructing ancestral languages, a component that computational models naturally lack.

This thesis presents a neural model that aims at bridging that gap by providing typological information about the likelihood of sound changes. The model was trained on large-scale global lexical data and is therefore able to assess whether a queried sound change is common or uncommon on a global scale. Since it operates on phonological features, it is able to process any given sound change between two arbitrary IPA symbols.

The model was trained on sound changes observed in Maximum Parsimony reconstructions on a large-scale global lexical dataset. The model was trained as a binary classifier in a noise-contrastive estimation setting, where the observed sound changes contributed positive training data which was weighed against randomly generated negative training data.

Applying a weighted version of Maximum Parsimony, in which the weights were derived from the model, produced better reconstructions for Proto-Austronesian and Proto-Oceanic than unweighted Maximum Parsimony reconstructions. That showed that the model was able to learn common sound changes, including the direction in which they tend to happen. While it requires further systematic testing, the model shows the potential to enhance tasks in computational historical linguistics by simulating implicit linguistic knowledge as a component of the comparative method.

Acknowledgments

My journey as an aspiring linguist began in the fall of 2015 when I enrolled to the General Linguistics program at the University of Tübingen. Since then, I have met so many fascinating people who have taught me important and interesting lessons, who helped me when I was struggling, who laughed and cried with me. I know that it sounds like a cliché, but I mean it: Every single person has contributed to who and where I am today, and I am very thankful for everyone who has accompanied me on my way. While this applies to way too many people to name them all, I would like to specifically thank some people who have contributed to my studies or my life in the best way possible. A special thank you to...

...Johannes, my boss and supervisor. I am not exaggerating when I say that I couldn't have done it without you. You provided me with invaluable advice, you always found time for a discussion or a debugging session whenever I needed it, you encouraged me to keep on going when nothing would work. Besides that, you gave me the possibility to work on interesting research questions in computational historical linguistics as part of an awesome team. It is no understatement to say that nobody at the SfS has influenced me nearly as much as you did. Thank you for everything.

...all the teaching staff at the SfS for introducing me to a plethora of interesting concepts and giving me the opportunity to participate in many fascinating projects. Most importantly, however, I am very grateful for the friendly, respectful, and cooperative way everybody treats each other. Such a climate can by no means be taken for granted in academia.

...Susanne for proofreading this thesis and providing me with very good, detailed feedback and annotations.

...Daniela and Ben. You kept me company during the largest part of the writing process and kept me productive on a daily basis. I really appreciated each stupid joke, each nerdy linguistic discussion, and the good balance of distraction and productivity.

...my roommates Meike, Paul, and Konstantin. You really turn our flat into a home for me, which I can not value highly enough.

...my best friend Oliver for countless good talks, for all the good times we enjoyed together, and all the bad times we helped each other through.

...my younger brother Fynn for constantly holding up a mirror in which I can see myself (whether you like it or not).

...my older brother Nils for always taking your time to help me out, and especially for your valuable input to optimizing and debugging my code.

...Mom and Dad. You raised me to be the person I am today with unconditional love and an almost incomprehensible amount of patience. You always let me explore my interests at my own pace and allowed me to choose my own path. I can not be grateful enough for everything you have done for me.

Contents

Introduction

In the nineteenth century, a school of linguists who called themselves the Neogrammarians discovered that different sounds in related languages correspond to each other in a strikingly systematic way. Based on this observation, they claimed that sounds change in a completely regular manner without any exceptions, and that these changes can be expressed by sound laws. This claim gave rise to the comparative method, a powerful toolset that enables historical linguists to reconstruct ancient languages by comparing modern related languages to each other.

While the absolute claim of regularity has been falsified by now, the comparative method is still the most popular and reliable tool for reconstructing extinct languages – while not completely without exceptions, the majority of sound changes do follow regular patterns. For quite some decades, a many unattested proto-languages have been successfully reconstructed by applying the comparative method, which up to this day is "heralded as one of the major intellectual achievements of the nineteenth century" (Campbell, 2013).

With the rapidly increasing availability of computing power over the last decades, computational approaches towards language comparison and reconstruction have recently gained popularity and now form a relatively young line of research commonly referred to as *computational historical linguistics*. Most researchers aim at models that imitate the comparative method or parts of it. However, the field of computational historical linguistics is still in its infancy, and no model so far is able to fully automate the comparative method. Scholars have rather addressed individual sub-problems, such as cognate detection or ancestral state reconstruction.

For some of those tasks, researches have developed some techniques that work reasonably well. For other tasks, however, there are still no satisfying solutions. Only a few models have been proposed for ancestral state reconstruction, the task to reconstruct the phonological or phonetic value of a word in a given proto-language. None of these models has posed a convincing solution for this task, which has shown to be quite challenging.

But what makes ancestral state reconstruction so much harder than other tasks in computational historical linguistics? Naturally, there is no single answer to this question, but a major part of the answer is that linguists employ a good deal of implicit knowledge when reconstructing proto-languages. A trained linguist knows which sounds are commonly used in languages and has an intuition about which

kinds of sound changes are plausible and which ones are not. All this knowledge was acquired by experience with many different languages and language families – it can not be deducted from a piece of linguistic data alone.

Computational methods, however, completely rely on the data they are working on – it is the only source from which they can learn linguistic patterns! Therefore, they are obviously not able to apply patterns observed beyond the scope of the present data, which plays a major role in manual linguistic reconstruction. This thesis introduces a neural model which serves the purpose of bridging this gap by providing global information about the plausibility of sound change. By learning on large-scale global lexical data, it is able to capture general sound change tendencies and assess how common or uncommon a given sound change is. Since it operates on phonological features, rather than on discrete IPA sounds, it is able to process any sound and therefore even assess sound changes that have not been seen in the training data. Techniques for ancestral state reconstruction can benefit from such a model in two ways. First of all, it enables those techniques to reconstruct sounds that are not present in the given data, since the model is able to generalize over all kinds of sounds and sound changes. This alleviates a strong limitation that all techniques for this task suffer from so far, that they are only able to operate on a fixed alphabet of sounds that are present in the data. Furthermore, the model can serve as a typological prior for sound change probabilities, which can enhance models that focus on local patterns within the data.

In Chapter 1, I introduce the core concepts for phonological reconstruction. I explain the general workflow of the comparative method and discuss its theoretical foundation, strengths and limitations. I then elaborate how and why sound changes, and conclude the chapter by briefly discussing how realistic reconstructed forms and languages can be.

Chapter 2 gives an overview of related work in computational historical linguistics, discussing approaches to common individual sub-tasks and addressing open problems in the field.

Chapter 3 provides a detailed explanation of my methodology and describes the full workflow for developing and applying the model. I describe which lexical data was used, how it was processed in order to generate training data for the model, and how the model was trained and post-processed.

Chapter 4 describes the set-up for evaluating the model by applying it to existing techniques for ancestral state reconstruction. I discuss the results of this evaluation quantitatively and qualitatively in chapter 5. Finally, chapter 6 concludes this thesis and suggests immediate improvements to the workflow outlined in this thesis, as well as possible future work in the field.

1 Phonological Reconstruction

1.1 The Comparative Method

The comparative method is the most successful tool to reconstruct ancient languages that did not leave any records behind and is therefore central to historical linguistics. It relies on the assumption that sounds change in a completely regular way which can be predicted by stating according sound laws. These regular sound changes result in systematic sound correspondences in related modern languages. For example, compare the German words *Apfel, Pfad*, and *Pfennig* to their English counterparts *apple, path*, and *penny*: German consistently has a *pf* in places where English features a *p*.

Correspondences like this one are no coincidence – they can frequently be observed in related languages and provide strong evidence in favor of the regularity assumption. While this assumption in its strict sense is falsified by now, since there are clear instances of irregular sound changes, it still holds true for the vast majority of historical sound changes. This quasi-regularity of sound change enables linguists to reconstruct ancient language states by comparing documented languages that are related to each other (Campbell, 2013). If sound changes did not exhibit a high degree of regularity, it would not be possible for linguist to detect systematic sound correspondences in related languages in the first place, which then enables them to postulate sound changes and reconstruct proto-languages – which are essentially the core principles of the comparative method.

Throughout this chapter, I will demonstrate these principles and how they are applied, then I will discuss the regularity assumption and its implications, and finally I will turn to the limitations of the comparative method.

1.1.1 Procedure

Despite what its name suggests, the comparative method is not a uniform, fully defined pipeline of concrete techniques that should be applied in a certain order. It rather serves as an overarching term under which certain principles are collected

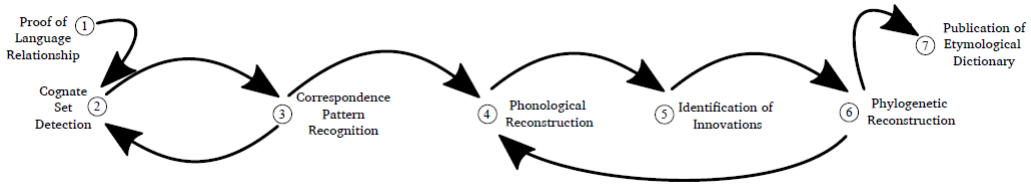

Figure 1.1: Workflow of the Comparative Method by Ross and Durie (1996), adapted and illustrated by List (2022).

that should be applied when reconstructing proto-languages by the means of comparing their extant daughter languages. While there is no clear consensus about the exact techniques to be used and their concrete ordering, linguists largely agree on some basic principles that are to be considered in comparative linguistics.

Figure 1.1 illustrates a possible workflow of the comparative method as it was outlined by (Ross and Durie, 1996). In this section, I will use this structure to explain the important principles for the comparative method, mainly because it nicely mirrors common subtasks in computational historical language comparison, as I will show in Section 2 (List, 2022). However, there are many other possibilities to declare tasks and techniques in the comparative method, which are not inherently more or less valuable (Crowley and Bowern, 2010; Campbell, 2013) – the structure chosen here is merely one of many ways to bundle numerous principles in subtasks of a workflow. Whichever exact approach one chooses, the comparative method is a highly iterative process, and each part of it needs to be constantly revisited and checked against new findings from other stages. An etymological scenario should therefore be constantly self-optimizing, rather than being a product of the subsequential application of a number of techniques – this is also indicated by the backwards-pointing arrows in Figure 1.1.

This section will use the workflow proposed by Ross and Durie (1996) to subsequently illustrate the core principles and ideas of the comparative method and how to successfully apply it.

1.1.1.1 Proof of relatedness

The first step outlined by Ross and Durie (1996) is often overlooked and seems counterintuitive at first glance – is it not the exact purpose of the Comparative Method to establish relationships between languages? Many scholars consider the successful application of the comparative method to prove the relatedness of the languages in question (Campbell and Poser, 2008). Nichols (1996) however argues that in fact, quite the opposite is the case: Determining the relatedness of languages is an independent task from reconstructing a proto-language by applying the later steps of the comparative method.

The core of this claim becomes more apparent when the notion "proof of relatedness" is rephrased: In order to successfully apply the comparative method, one needs to *assume* that the languages in question are related to each other in the

first place. The comparative method assumes a strict regularity of sound change which accounts for regular and systematic sound correspondences in the daughter languages, which are the basic building blocks for the comparative reconstruction of a proto-language (discussed in the following sections). If two languages are not related to each other, there is no systematic distribution of the languages' sounds with regard to each other – therefore, it would be impossible to find systematic sound correspondences and to apply the comparative method in a meaningful way.

In fact, the first applications of the comparative method for Indo-European languages did not really consider the question of relatedness – it was just apparent that the languages that were compared to each other were related. Slavonic philologists have been treating Slavic languages as a genetic unit ever since; their relatedness has never been disputed. Both, the fact that Slavic languages up to this day are quite similar to each other and are mutually intelligible to a certain degree, and that there is a traditional shared Slavic identity, lead to the conclusion that their genetic relatedness is self-evident and does not require any further formal proof (Jagić, 1910; Nichols, 1993, 1996).

Although not as obvious as for Slavic language, the same self-evident relatedness applies to Indo-European languages. Consider the well-known quote by Sir William Jones:

> The Sanskrit language, whatever may be its antiquity, is of wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either; yet bearing to both of them a stronger affinity, both in the roots of verbs and in the forms of grammar, than could have been produced by accident; so strong that no philologer could examine all three without believing them to have sprung from some common source, which, perhaps, no longer exists. There is a similar reason, though not quite so forcible, for supposing that both the Gothic and Celtic, though blended with a different idiom, had the same origin with the Sanskrit; and the old Persian might be added to the same family. (quoted from Nichols 1996)

The main innovation attributed to this work by Jones lies within the idea that related languages descend from a common ancestor language that does not exist anymore. At his time, it was already well known that languages are related to each other in some way, and philologists have been studying the structure of languages like Latin or Greek for a long time, but there was a prevalent conception of some languages representing newer stages of other languages that still exist – for example, it was assumed that Latin was a newer or even a "corrupted" form of Greek for a long time (Nichols, 1996). The concept of older language stages that no longer exist essentially built the foundation for reconstructing such extinct languages by comparing their modern descendants, i.e. the comparative method.

More striking in this context, however, is the strong claim about the relatedness of the languages. Note how Jones does not refer to lexical similarity at any given point, but rather bases his claim on structural properties, namely the "roots of

verbs and [...] the forms of grammar", which in modern terms refers to the inflectional morphology of verbs and nouns. More generally even, none of the well established language families today has been assembled merely based on lexical evidence. To safely group languages together as a family, they need to share certain structural features in function and form. Consider the parallel adjective paradigms from Latin and Ancient Greek in Table 1.1: Not only do both languages have a very similar case system, where the individual cases are used in similar ways; but the paradigms also show syncretisms at the same places, using the same suffix for the masculine accusative and both neuter nominative and accusative. There is also a clear correspondence between individual sounds, like Latin *u* and *m* corresponding to Greek *o* and *n* respectively. It is clearly visible how these paradigms in the two languages mirror each other both in function, the case system; and in form, that is the syncretisms and sound correspondences (Nichols, 1996).

	Masculine	**Feminine**	**Neuter**
Latin:			
Nominative	-us	-a	-um
Accusative	-um	-am	-um
Greek:			
Nominative	-os	*(*)-ā*	-on
Accusative	-on	*(*)-ān*	-on

Table 1.1: Reduced adjective paradigm in Latin and Ancient Greek, taken from Nichols (1996).

Lexical similarity on the other hand can only provide very weak evidence for language relatedness. On the one side, lexical items are borrowed at much more easily than morphological or syntactical features, so the vocabulary of a language is very prone to being affected and altered due to language contact. On the other side, semantic change happens unsystematically and is therefore unpredictable, so it is easy to find words that resemble each other by chance in the comparison of arbitrary languages if you allow for enough semantic variation. Macro-families like Nostratic (Dolgopolsky, 2008) or Amerind (Greenberg and Ruhlen, 2007) have been proposed based on such multilateral lexical comparisons, but it has been shown that their suggested evidence does not meet the criteria of statistical significance – it is more likely that the found similarities have developed independently from each other by mere chance (Ringe, 1992, 1999; Nichols, 1996).

In essence, the comparative method can only be successfully applied on languages that are related to each other. While Nichols (1996) claims that this relatedness between the languages has to been proven independently from the comparative method, her main criteria for determining language relatedness fall in line with what most scholars that see the comparative method as a proof for language relatedness agree on: Besides sharing a large part of their core vocabulary, languages should also exhibit systematic sound correspondences and similarities in morphology and syntax in order to be classified as related (Campbell and Poser, 2008).

In simpler terms, lexical evidence alone is not enough to establish language relatedness, which draws the line between established families like Indo-European, and speculative macro-families like Nostratic: "Dictionaries of groupings like Indo-European are compiled only after relatedness is assumed or proven, and they serve to reconstruct and subgroup, while dictionaries of the long-range groupings are offered as evidence of relatedness" (Nichols, 1996). The successful application of the comparative method can provide strong evidence supporting that certain languages are related, if it is applied to an extent where complete etymological scenarios, including inflectional and derivational morphology, can be reconstructed; however, language families that are proposed mainly based on lexical evidence like those mentioned above usually do not meet the criteria for safely establishing language relatedness – even if parts of the comparative method have been applied.

1.1.1.2 Detection of cognates and sound correspondences

Cognate sets are the basic building blocks of the application of the comparative method, since they reflect to which forms a single proto-word has evolved in the modern languages. Cognacy is used in a strict sense here – only words that are directly inherited from a common ancestor should be considered. Since the comparative method aims at reconstructing how the proto-language has changed and split into its daughter languages, only directly inherited forms that can tell this story should be considered. Borrowed words should be disregarded, even if they ultimately happen to trace back to the same origin: Loanwords follow different phonological rules than native vocabulary, and they have undergone different sound changes in their history. The English word *egg* for example should not be considered when reconstructing Proto-West-Germanic since it was borrowed from Old Norse – even though it originates from Proto-Germanic **ajjǫ*, just like the direct reflexes of other West Germanic languages (Kroonen, 2013).

Putative cognate sets can be assembled in the first place by comparing languages' basic vocabularies, which are generally considered to be quite stable and therefore less prone to borrowing and semantic shift than other parts of the vocabulary. Words in related languages that are similar in both meaning and form are unlikely to coexist next to each other by accident, or in other words, are likely to be cognates. Consider the following basic concepts from four Polynesian languages[1]:

Tongan	Samoan	Rarotongan	Hawai'ian	
tapu	tapu	tapu	kapu	'forbidden'
taŋaka	taŋaka	taŋaka	kanaka	'man'
maŋa	maŋa	maŋa	mana	'branch'
puhi	feula	pu'i	puhi	'blow'

Table 1.2: Cognate sets for basic concepts in Polynesian languages, taken from Crowley and Bowern (2010).

[1]The grapheme ' is used to express the glottal stop /ʔ/ in Polynesian languages.

It is obvious that the first three concepts are full cognate sets, since the words in all languages have the same meaning and resemble each other strongly. The same observation can be made for Tongan, Rarotongan, and Hawai'ian in the fourth set, however Samoan *feula* looks substantially different and is therefore likely not a cognate to the other three forms.

Naturally, the first assessment of cognacy can often be speculative to some degree, since only the similarity on the surface is considered. Cognate sets are therefore constantly revisited in the later steps of the comparative method – new findings regarding sound correspondences and sound laws can provide new evidence for or against certain cognacy judgements. English *egg* would probably be seen as probable cognate to German *Ei* in the initial stage, but since it would contradict sound laws found at a later stage, it can be safely classified as a borrowing at that point. Some true cognates on the other hand require a certain understanding of sound laws: Although Nganasan *bî'*, Nenets *ju'*, and Selkup *köt* have all regularly evolved from Proto-Samoyedic **wüt*, they appear to be quite different on the surface and would therefore not be recognized as cognates instantly (Daneyko, 2020).

Understanding sound correspondences and sound laws is therefore crucial towards showing whether certain words are actually cognates, but those can only be established by assembling putative cognate sets in the first place. The comparative method therefore is a strongly iterative process: New insights about sound laws can show or discard putative cognate sets; the newly assessed cognate sets in return can provide new sound correspondences, giving evidence for or against certain sound laws. Applying well-established sound laws can prove cognacy relations between words whose semantics seem too dissimilar to be considered cognates initially, like German *walken* 'to knead' and English *walk*, which both have regularly developed from Proto-Germanic **walkan* 'to roll' (Daneyko, 2020). By the same means, 'false friends' – words that look like they are cognates, but are in fact not – can be identified: English *have* and Latin *habēre* look very similar and share the same meaning. However, Grimm's law states that voiceless stops change into voiceless fricatives from Proto-Indo-European to Proto-Germanic, so the English initial *h-* traces back to Proto-Indo-European **k-*, which Latin has retained – *have* and *habēre* therefore can not be cognates! In fact, English *have* comes from Proto-Indo-European **keh₂p-*, which makes it a cognate of Latin *capere* 'to take' (Kroonen, 2013; De Vaan, 2018).

Whether having being established based on known sound laws or merely based on superficial similarities, as needs to be done in the first stage, cognate sets provide valuable information about sound correspondences – which sounds in language A regularly correspond to which sounds in language B. Looking at the cognate sets from Table 1.2, some interesting observations can be made: Hawai'ian *k* seems to systematically correspond to *t* in the other three languages, and likewise Hawai'ian seems to have *n* in the same places where the other three languages feature an *ŋ*. Following the convention by Crowley and Bowern (2010), these sound correspondences can formally be denoted as $t : t : t : k$ and $ŋ : ŋ : ŋ : n$ respectively, denoting which language uses which sound in a given correspondence (using the same order of languages as the table above).

A further sound correspondence that can be observed in the fourth example is that Rarotongan *'* corresponds to Tongan and Hawai'ian *h*. Since there is no Samoan reflex for this cognate set, however, it is impossible to know the corresponding Samoan sound! In cases like this, the sound correspondence has to be denoted as incomplete for the moment.[2] Throughout this thesis, I will follow the notation introduced by List (2019a) and denote incomplete sound correspondences like this as *h : ø : ' : h*. Unlike Crowley and Bowern (2010), I will use the symbol *ø* only for denoting a missing reflex in a sound correspondence. Indicating a gap in a sound correspondence, where a certain sound in a given language systematically corresponds to the absence of a sound in another language, will be denoted by the gap symbol -. For example, Tongan *'ahu* 'gall' is cognate with *au* in the other three languages that lack the consonants in this word. Therefore, Tongan *'* and *h* correspond to a gap in the other languages, which is denoted as *' : - : - : -* and *h : - : - : -* respectively. The gap symbol therefore indicates that the cognate sets show the systematic absence of a certain sound in a certain language, while *ø* is used to show that the corresponding sound of the respective language is unknown due to a missing reflex.

All sound correspondences that have been introduced so far were perfect correspondences, where each sound of a given language could be mapped to exactly one sound in another language. In many cases however, things are not as simple, and sound correspondences can be overlapping. Consider the following minimal example, consisting of two cognate sets from four Romance languages:

Italian	Spanish	Portuguese	French	
caro /karo/	*caro* /karo/	*caro* /karu/	*cher* /ʃɛr/	'dear'
colore /kolore/	*color* /kolor/	*côr* /kor/	*couleur* /kulœr/	'colour'

Table 1.3: Minimal example of overlapping sound correspondences in four Romance languages, taken from Campbell (2013).

The initial *k-* in Italian, Spanish, and Portuguese corresponds to French *ʃ* in the first example, but to French *k* in the second one; yielding the overlapping sound correspondences *k : k : k : ʃ* and *k : k : k : k*. Unlike the examples seen before, in this case it is impossible to predict the French form from the other languages – initial *k-* could either correspond to French *ʃ-* or *k-*. Overlapping sound correspondences like this one hint towards conditioned sound changes, whereas perfect sound correspondences usually are a result of unconditioned sound changes.

1.1.1.3 Phonological reconstruction and sound law inference

After having identified the systematic and frequent sound correspondences, a historical linguist will turn to reconstructing the respective proto-sounds and inferring the sound laws involved. Since understanding how sound changes work and how linguists use these pieces of information to "reverse-engineer" them is a central

[2]Later evidence can always be included to complete a sound correspondence. In this case, a cognate set like *ahi - afi - a'i - ahi* 'fire' can show the complete sound correspondence *h : f : ' : h*.

part of this thesis, I will discuss this topic in detail in Section 1.2. For the purpose of illustrating the work flow of the comparative method concisely, I will use this section to merely show the implications that reconstructed proto-sounds have on the workflow, assuming that the correct proto-sound was reconstructed.

First of all, there are some hard structural constraints on which proto-sounds can be reconstructed. For each sound correspondence that has been determined in the previous step, exactly one proto-sound has to be reconstructed, following the assumption that sound change is strictly regular (Campbell, 2013). For illustration purposes, I will just use the reconstruction principle of the *majority vote* and postulate the sound that is reflected in the majority of extant languages as proto-sound. As we will later see, this is only one of several rules of thumb to choose the most likely proto-sound, but it is sufficient for the examples we have seen so far.

By the means of this simple principle, **t* would be the best candidate for the Proto-Polynesian sound that is reflected in the sound correspondence *t : t : t : k*. Reconstructing **t* implies that there was a sound change from /t/ to /k/ in Hawai'ian, which can formally be denoted as /t/ > /k/. Likewise, the sound correspondence *ŋ : ŋ : ŋ : n* implies a sound change /ŋ/ > /n/ for Hawai'ian.

Since these sound correspondences are not overlapping, it is easy and straightforward to posit the respective sound changes. These sound changes for Hawai'ian are unconditioned, so they can be stated quite simply – every Proto-Polynesian /t/ is reflected as a /k/ in Hawai'ian. Things become a bit more complicated when dealing with overlapping sound correspondences, like the one observed in Table 1.3. Following the majority vote principle, we reconstruct the proto-sound **k* for both sound correspondences, *k : k : k : ʃ* and *k : k : k : k*. That leads to a problem: We need to come up with sound laws that explain how Latin /k/ in some cases turned into /ʃ/, but stayed /k/ in others.

As already briefly mentioned above, this is a case of *conditioned* sound change – /k/ became /ʃ/ in French only under certain conditions and stayed /k/ otherwise. Historical linguists usually denote such sound laws in the same way that phonological rules are defined by generative phonologists. In this case, the sound change is conditioned by the following vowel – /k/ only becomes /ʃ/ if followed by a front vowel. This rule is commonly written as:

$$/\text{k}/ \rightarrow /\text{ʃ}/ \ \backslash\_[\text{-cons,-back}]$$

Just like generative phonologists, historical linguists commonly make use of phonological feature representation to express structural similarities of sounds and bundle them together by those means – in that case, all vowels that are not back vowels (/i, e, ɛ, a/ in Old French) trigger the sound change. Sound laws have to be stated in a way that they derive the modern reflexes reliably and deterministically from the proto-forms: For each reconstructed proto-form, there can only be one possible reflex in the daughter languages by rigorously applying the sound laws, so there is no room for ambiguity.

Another parallel to derivational phonology is that some rules will affect the same

target sounds, in which cases a rule ordering has to be defined. The intuition behind rule ordering is actually a lot simpler in the case of historical linguistics: For derivational phonology, the idea is that one rule is applied before another one in an abstract, hypothetical cognitive derivation process – the interpretation for historical linguistics is just that a certain sound change happened at an earlier point in time than another one. For example, Old French /an/ changed to /ɑ̃/ at some point, deriving the modern surface form /ʃɑ̃t/ for *(je) chante* 'sing (1.sg.)' from /ʃant/, which is ultimately derived from Latin *cantō* /kantoː/. In combination, both sound laws derive the French form from the Latin form like this (ignoring the loss of the final vowel):

$$/\text{kantoː}/ > /\text{ʃant}/ > /\text{ʃɑ̃t}/$$

This derivation implies that /k/ has become /ʃ/ *before* /an/ has turned to /ɑ̃/ (Pope, 1934). If the rules were applied in the reversed order, the latter rule would take away the context for the first rule to apply by shifting the vowel to a back vowel, and /k/ would not change to /ʃ/ – which would be equivalent to *bleeding* in terms of derivational phonology. This however does not happen, since both rules do apply, which shows that the former sound change happened before the latter one.

That exemplifies how sound laws have to be inferred from the identified sound correspondences according to two important principles. The set of inferred sound laws needs to explain all regular sound correspondences by properly defining both the right conditions and the right order for the sound changes to apply. Coming from a reconstructed proto-form, this set of rules should be able to derive the extant forms in a deterministic, non-ambiguous manner.

1.1.1.4 Reconstruction of an etymological scenario

The final step of the comparative method is to collect all the evidence that has been gathered by recursively applying the aforementioned steps and to flesh out a full etymological theory. This theory – usually compiled in an etymological dictionary – should tell the story about how all extant languages have derived from the reconstructed proto-language (Ross and Durie, 1996). That includes not only providing systematic reconstructions for all cognate sets from the data, but also to to reconstruct grammatical morphemes in form and function. In essence, an etymological dictionary should systematically explain the heritage of the grand share of lexemes in the data, including how inflectional and derivational morphology has developed over time. For partial cognates, where for example two words in the extant languages share the same stem but include some other morpheme, the theory needs to be able to thoroughly explain which morphological processes have occurred that lead to the forms we find in the data. That means, that the etymological scenario would not only need to provide reconstructions for the shared stem, but also for the other (functional) morphemes.

In some cases however, it is not possible to fully reconstruct a proto-form, because its reflexes have not been retained in all daughter languages. This leads to an incomplete sound correspondence in the respective cognate set, as described pre-

viously in Section 1.1.1.2. Since sound correspondences can be overlapping, such incomplete correspondence sets can sometimes not be clearly mapped to one full sound correspondence. In such cases, it is common practice to underspecify the sound that can not be reconstructed with full certainty. Indo-European philologists for example regularly use the symbol *H* to denote any of the three laryngeals in cases where there is no evidence from Greek (List, 2019b).

Finally, an etymological theory also needs to account for forms that have not been passed down from mother to daughter language in a regular fashion, as the comparative method strictly assumes. That includes explaining how and why some irregular changes have happened that violate the rigorous assumption of regularity; these kinds of changes will be discussed in Section 1.1.2. Another obvious case that needs to be considered is borrowing: Loanwords (or any other piece of language that has been borrowed) also call for explanation. The theory should be clear about which words have been borrowed from which donor language within which timeframe. That requires the inclusion of some external information, ranging from linguistic knowledge about other language families to geographical, cultural, and social factors throughout history. Naturally, the theory therefore also needs a concise explanation when the population that spoke a given proto-language lived in which geographical area, and to which other communities they have had contact at which point in time. To frame it in an exaggerated example, it would be unreasonable to assume that the same community of speakers borrowed a word from a language spoken in South-East Asia, and another one from a South African language.

Ultimately, a good etymological scenario should provide some deeper understanding of the history of the respective language family. That can enable historical linguists to base new claims on that acquired knowledge, or to analyze new findings in the context of what is known about the family's history. Prominently, this refers to recognizing new languages as members of a family. Revisiting Jones' quote from earlier, he proposed that Gothic, Celtic, and Persian might as well belong to the Indo-European family – a hypothesis that could be confirmed at a later point, *after* already gaining some knowledge about how early stages of Indo-European must have looked like based on the evidence of Sanskrit, Ancient Greek, and Latin. Since the potential new family members could easily be derived from the proposed proto-language, it was evident that these languages were in fact Indo-European as well. This understanding of the structure of (Proto-)Indo-European also enabled researchers to identify the Tocharian languages as Indo-European after they were discovered (Ross and Durie, 1996; Nichols, 1996). After all, the comparative method, and in a broader sense linguistic reconstruction, is highly recursive in nature, and every new piece of evidence can change parts of a theory.

1.1.2 On regularity of sound change

In the previous sections, I have already hinted at the fact that the comparative method relies on some rigorous assumptions, most prominently that sound change occurs in a regular manner without any exceptions. Ross and Durie (1996) explain

that the comparative method is not to be seen as a tool on its own, but is strongly associated to the Neogrammarian hypothesis (Osthoff and Brugmann, 1878), which explicitly claims the strict regularity of sound change. This theory was a result of a heavy Darwinist thinking that was prevalent among contemporary scientists – Schleicher (1863) considered languages to be natural organisms that follow natural, exceptionless laws. He therefore claimed that linguistics were a natural science, since according to his view, it investigated natural organisms, just like for example biology. This idea also gave rise to depict the genetic relation of languages to each other by means of trees, analogously to the tree of life – a model that, despite its drawbacks and inaccuracies, is still widely used today.

The same holds true for the comparative method and its assumption of regular sound change. As of today, there is plenty of evidence that sound change is not completely regular. In fact, there are many reasons why sounds can change in a different way than sound laws would predict, which include analogy, onomatopoesia, sound symbolism, avoidance of homophony, tabooization, and bilingualism (Ross and Durie, 1996). According to Grimm's and Verner's Laws, English *father* should surface as **[fadə(r)], however the medial consonant has become a [ð] in analogy to *brother* which exhibits the same medial consonant (which has developed regularly in the latter case). Analogical sporadic sound change often occurs in words that are found in narrow semantic fields like kinship terms or lists like numerals (McMahon, 1994). Latin *lupus* 'wolf' has been borrowed from Sabellic **lupo-* in order to avoid the direct utterance of a taboo term; the direct Latin reflex would have been ***lucus* (De Vaan, 2018). Bantu languages of Southern Africa borrowed click consonants from adjacent Khoisan languages which would partially replace their native consonant inventory (Daneyko and Bentz, 2019). In Northern Estonian, final *n* was lost regularly, except in first person singular verb forms to avoid homophony with the singular imperative form; retaining forms like *kannan* 'I carry' (Campbell, 1996).

All of these examples violate the regularity assumption made by the comparative method and therefore can not be accounted for by applying it in a strict sense. Nevertheless, it is indisputable that sound changes exhibit a strong tendency towards being regular, and examples of sporadic changes tend to be the exception rather than the rule. Such cases can not be handled by the comparative method, but require case-by-case examination. While the claim of regularity of sound change without exceptions has to be falsified, the principle of regularity still holds true for most cases and allows historical linguists to systematically reconstruct older language stages. Campbell (1996) concludes that "although irregularities in sound change undoubtedly occur, we should not give up the basic concept of the regularity of sound change since many irregularities can be explained by linguistic and sociocultural factors which may 'interfere' with regularity but do not undermine the principle of regularity itself."

1.1.3 Limitations

Besides the regularity assumption discussed in the last section, the comparative method is further limited by strongly relying on the tree model: It assumes that languages change along the branches of a phylogenetic tree, and that there is exactly one uniform (proto-)language at each of its nodes (Campbell, 2013).

These assumptions should be treated in a similar way as the regularity assumption – representing languages as leaves and nodes on a tree is a model that can capture certain phenomena, but does not reflect the whole truth. The comparative method aims at reconstructing one uniform proto-language, disregarding any sort of internal variation. However, it is safe to assume that extinct languages behaved like modern languages and featured dialects, sociolects, and diastratic variation. Disregarding these layers of variation by assuming the uniformity of a language (which is also frequently done for modern languages) already constitutes an abstraction of the actual reality of the spoken language.

According to the tree model, a proto-language would split up into multiple daughter languages at a certain point in time, which then are located on different branches on the tree. This has two further implications: First of all, it implies that there is an abrupt, non-gradual change from mother to daughter language. Of course, no linguist would assume that language evolves this way, but it is worthwhile to note that the tree model is technically not able to model the gradual transition between different language stages – that requires further human interpretation outside of the model. The second implication however is the more important one for the application of the comparative method: Once a language has split into different daughter languages, these languages will evolve independently from each other. The tree model therefore is not able to account for horizontal contact between languages after they have split up – and in a broader sense, it is not able to model language contact and its effects at all. For example, Norwegian Bokmål is considered to be a descendant of Old West Norse, just like Faroese and Icelandic. Swedish and Danish on the other hand have evolved from Old East Norse – that implies that Norwegian is more closely related to Icelandic and Faroese than to Swedish and Danish (Hammarström et al., 2022). While this might reflect the historical evolution of the languages, it completely disregards the factors of geography and thus language contact: Norwegian has been in close contact with the mainland Scandinavian languages over the last centuries, while being isolated from its 'siblings' spoken on islands. As a result, Norwegian today is much more similar to Swedish and Danish, whose speakers can understand each other most of the times. Icelandic and Faroese on the other hand are now too different from Norwegian to retain mutual intelligibility. In order to successfully reconstruct respective proto-languages, a historical linguist is bound to deviate from the tree model in order to account for the geographic reality of these languages and the degree of contact they have had with each other.

Last but not least, the quality of reconstructions is naturally directly dependent of the availability of data. The more languages of a family are known and can provide data, the better a linguist can understand the history of the family and

reconstruct earlier language stages. If French and English were the only surviving Indo-European languages, it would not even be able to identify them as belonging to the same family (Meillet, 1958), leave alone to reconstruct a common ancestor. For Latin, most lexical items and phonological structures can successfully be reconstructed from the modern Romance languages, but there is not enough evidence for the extensive case system and verbal inflections, since no modern descendant has retained these features systematically (Campbell, 2013). Reconstructions for families with bad availability of data are therefore more speculative and less accurate in nature.

1.2 Sound change

1.2.1 How sounds change

A number of frameworks have extensively dealt with the question how sounds change over time and have come up with different explanations and models. Neogrammarians believed that sound changes arise at a given point in time and affect the whole lexicon simultaneously, following the rigorous principle of regularity. Structuralists focused on the phonological structure of a language, seeking the main motivation for sound change in the different functions that phonemes (should) have in a language to make its phonology efficient. According to this theory, languages tend to avoid asymmetric and therefore inefficient phonological systems. Generativists viewed diachronic sound change in the framework of generative phonology, concluding that every phonological change can be explained by a change in either phonological rules, their ordering, or the underlying representations (McMahon, 1994).

As of today, there seems to be a consensus among scholars that sound changes do not happen abruptly (as claimed by the Neogrammarians), but rather gradually. Any given sound change therefore starts as an innovation that affects a small part of the lexicon, and from there gradually spreads to other lexical items. Given enough time for the sound change to happen without other processes interfering, it will eventually affect the whole lexicon and therefore fulfil the Neogrammarians' regularity claim. However, this model of lexical diffusion can also explain residual effects, sporadic cases where a sound change did not affect individual words that should have been affected – basically, a sound change can just 'die' before having affected the whole lexicon (McMahon, 1994; Crowley and Bowern, 2010). This nicely accounts for the evident quasi-regularity of sound change with both truths that come with it, that there is an undeniable regularity on the one hand, but also some clear exceptions on the other hand. Likewise, this model can also account for sporadic sound changes, in which case that change has gone essentially unproductive before affecting the grand share of the vocabulary. If a sound change is rejected by the community of speakers early enough, it might even be reverted completely (Blust, 1996).

Sound changes do not only propagate gradually through the lexicon, but also through a community of speakers. Just like changes do not affect all lexical items

immediately at the same time, they are not instantly applied by all speakers of a community. Whether a speaker chooses to integrate an ongoing sound change to their idiolect can heavily depend on different sociolinguistic factors like age, gender, or social status; or simply based on where they live, i.e. whether a sound change has spread to their home region – a dimension that dialectologists frequently represent as isoglosses. Due to those factors, Ross and Durie (1996) argue for framing language change in a speaker-oriented paradigm rather than a language-oriented one. They claim that language should not be seen as a natural organism that lives and changes on its own, but as a human-made tool that can be applied and shaped to fit its speakers' needs, leaving space for both regularity and irregularity.

Although having established a reasonably good understanding of how innovations diffuse through the lexicon and the speakers over time and how this effectively changes languages, the question *why* sound changes arise in the first place still leaves much to be investigated. The Neogrammarians suggested that sound change can be attributed to physiological reasons – sounds or sound clusters that are hard to pronounce give way to those that take less effort to articulate. Latin obstruent clusters were systematically reduced to a geminate of its last segment, as in *septem* > *sette* 'seven' or *octō* > *otto* 'eight'. While this reasoning seems very attractive and should not be disregarded, it can not account for all sound changes: If that was the case, the same kind of sound changes would be applied in the same contexts in all languages of the world, which would make sound change essentially universally predictable (McMahon, 1994). Since that is not the case, there are necessarily some additional factors that lead to sound change. The most widely accepted explanation is that sounds that are frequently confused with each other are also likely to change into each other. This includes both errors on the speaker's and on the listener's side. The former one partially overlaps with the Neogrammarians' claim: Voiceless obstruents often become voiced intervocalically, because it is easier to articulate the whole vowel-obstruent-vowel sequence as voiced, rather than articulating only the obstruent unvoiced, which leads speakers to frequently mispronounce the obstruent as voiced. Including the role of the listener however is crucially different from attributing all sound changes to the ease of articulation – some differences are simply harder to perceive than others. Nasal consonants therefore are commonly subject to deletions, although they are not particularly hard to articulate, but they are hard to perceive due to their weak acoustic signal. If a listener therefore commonly misunderstands some words or sound sequences, they will naturally reproduce them how they have perceived them and therefore articulate some pieces of language differently than other speakers.

1.2.2 Phonemic and phonetic change

When applying the comparative method, linguists aim at providing a good phonological reconstruction of a proto-language. This term implies that the reconstructed sounds are actually phonemes, and not phones; representing the structure of the proto-language's phonological system rather than their exact phonetic realizations. By extension, that means that sound changes that only change the phonetic realization of a sound, without changing any part of the phonological

structure of the language, are usually not considered in reconstruction. Such changes are called phonetic or sub-phonemic changes (Crowley and Bowern, 2010; Campbell, 2013). An example of this is the English phoneme /r/, that used to be articulated as a trill [r] or a flap [ɾ], but in modern days has changed to an approximant [ɹ]. However, since this change does not interfere with any other phonemes, it does not change the phonological structure of English – and the underlying phoneme /r/ is still the same, because it still fulfils the exact same function in the phonological system (Crowley and Bowern, 2010). Allophonic change is another example of sound change that occurs on a sub-phonemic level and is therefore disregarded by phonological reconstruction. In many Spanish dialects, word-final /n/ surfaces as [ŋ]. Again, this does not change the phonological structure of the language and is perfectly predictable from synchronous data, so a historical linguist would not bother to explain where the surface [ŋ] comes from (Campbell, 2013). Since modern languages frequently display allophonic variation, it is safe to assume that languages in the past behaved in the same way – however, it is impossible to reliably reconstruct how different phones might have had the same structural function due to being allophones of the same phoneme. Strictly speaking, we can only reconstruct the function of a phoneme within the language system, but not its phonetic value(s); this issue will be elaborated further in Section 1.3. For now, it is sufficient to bear in mind that historical linguists are only concerned with sound changes that occur on a phonemic level and thus alter the phonological structure of the language.

The structuralist typology of phonemic change (Hoenigswald, 1960) essentially describes two types of sound change that impact the phonemic system: Splits describe the process of one phoneme splitting up in two different ones; whereas mergers involve two phonemes collapsing to a single phoneme (McMahon, 1994). As any other kind of sound change, splits and merges can be unconditioned or conditioned. So where does a sound change start changing the phonological system of its language?

Unconditioned mergers are the most intuitive and simple example to illustrate phonemic change. They describe sound changes where the phonetic realization of a phoneme changes in all environments, such that the new phonetic realization coincides with the phonetic value of another phoneme. Spanish *ll* used to denote the phoneme /ʎ/, but its phonetic realization has shifted to [j] in most dialects. Since Spanish already has a phoneme /j/ (written *y*) that has the same phonetic value, the phonemic distinction between /ʎ/ and /j/ has been lost. In other words, the two phonemes have merged into one phoneme /j/, resulting in the loss of the phoneme /ʎ/ (Campbell, 2013).

Splits on the other hand occur when a phoneme develops two different phonetic realizations, whose distributions can not be explained synchronously by allophony and phonological rules. Hoenigswald (1960) distinguishes between primary and secondary splits. The former always comes paired with conditional mergers, as illustrated in Figure 1.2: When two allophones stand in complementary distribution to each other, it is possible that one of them coincides with the phonetic realization of another phoneme. In Latin for example, intervocalic /s/ changed into [r]. Since

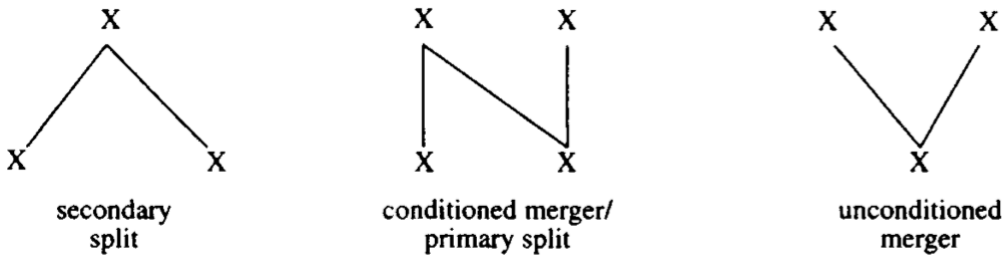

Figure 1.2: Illustration of mergers and splits, taken from McMahon (1994).

that was also the phonetic value of /r/, it made a [r] that is derived from /s/ indistinguishable from one that belonged to /r/ – in essence, all instances of the phone [r] were then reanalyzed as belonging to /r/. The intervocalic /s/ > [r] therefore first split off from /s/ (*primary split*), and then merged with /r/, *conditioned* by the context that triggered the phonological rule (McMahon, 1994).

In order to understand *secondary splits*, one needs to understand the axiom that *splits follow mergers* (Campbell, 2013). As already discussed, sound change is only considered phonemic when the surface forms can not be mapped back to their original phonemes deterministically, like for example the Latin [r] that could originally be derived from /r/ or /s/. Just as primary splits, secondary splits are also the result of a complementary distribution of allophones. However, as long as the phonetic value is cleary predictable by the context, there is clearly only one underlying phoneme. In order for this phoneme to split up in two phonemes, the *context* has to change, rather than the phonetic realization itself: When the context that originally determined the phonetic value of the phoneme is taken away or changed in a way that it loses its predictive power, the phonetic realizations suddenly become phonemically contrastive. For example, Nahuatl /s/ regularly surfaced as [ʃ] before [i] – obviously a clear case of allophony in complementary distribution. The phonological rule would apply for words like /sima/ > [ʃima] 'to shave', but not in words like /s̨ima/ > [s̨ima] 'to prepare plant leaves for extracting fibres'. However, Nahuatl has undergone an unconditioned merger that changed /ɨ/ to /i/, essentially losing the phoneme /ɨ/. Following that change, the previously introduced forms surfaced as [ʃima] and [sima], constituting a minimal pair and therefore creating a phonemic distinction between /s/ and /ʃ/. Note how the phonetic value of the sounds in question has not changed at all, but a *merge* in the context turned this allophonic to a phonemic distinction. The same principle applies when the context triggering an allophonic change disappears completely: If we postulate the deletion of the segment in question as a merger with the 'zero morphome' (that does not have any phonetic value), it again results in a split that was only brought to life by a previous merger (Campbell, 2013).

1.2.3 Guidelines for reconstructing proto-sounds

In Section 1.1.1.3, I have already hinted that there are some rules of thumb which help the linguist to determine the best proto-sound that reflects a given sound correspondence in a proto-language. Despite the fact that reconstructions can

only take place on a phonological level, which is naturally a certain abstraction of the phonetic reality of the spoken language, we still attempt to reconstruct a proto-sound that is as close as possible to how it was actually pronounced (Campbell, 2013). In order to do so, Campbell (2013) outlines some guidelines for phonological reconstructions, which I will summarize in this section.

The simplest principle is the *majority vote* principle, which was already introduced in Section 1.1.1.3. If no other factors suggest that one of the candidates seems more likely, we just let the majority win, as already exemplified in that section. The intuition behind that is quite simply that it is less likely for the same sound change to occur independently multiple times; so the reconstruction that implies less changes is considered the most likely.

Closely related to that is the principle of *economy*, which is based on the same idea of minimizing independent changes. Both principles are based on Occam's razor or the principle of parsimony, that processes should be explained with the fewest possible assumptions, i.e. that the simplest explanation is more likely to be true than more complex ones. The principle of *economy*, in contrast to the *majority vote*, is concerned with minimizing changes globally rather than locally. Since the comparative method assumes that languages change along the branches of a phylogenetic tree, its application implies the reconstruction of different intermediary nodes within the tree. At each node, the *majority vote* would strictly reconstruct the sound that is most commonly found in the direct descendants. The *economy* principle rather minimizes the number of independent sound changes across the whole tree: If the reconstructions at the different nodes of a tree imply that a given sound change has occurred multiple times within the language family, it is often more likely that this sound change has only happened once at an earlier stage, and the resulting sounds in the modern languages are a reflex of that sound change.

In cases where there are multiple different reflexes of the same proto-sounds in the extant languages, comparing the *shared phonological features* between the sounds can narrow down likely proto-sound candidates. Consider the cognate set for the word 'goat' in the four Romance languages that have been introduced previously; Italian *capra* /kapra/, Spanish *cabra* /kabra/, Portuguese *cabra* /kabra/, and French *chèvre* /ʃɛvr(ə)/. This cognate set contains the interesting sound correspondence *p : b : b : v* between the first vowel and the *r*. Based on the reconstruction principles that we have seen so far, there is no obvious choice for the best proto-sound, however all sounds in question are labial obstruents. Since all sounds in the correspondence share these phonological features, it is very likely that the proto-sound was a labial obstruent too, and it is highly unlikely for it to lack one of these features.

But how can we determine the best proto-sound for that sound correspondence? This is where the principle of *directionality* comes in handy: Certain sound changes are more common than others, which also implies that between pairs of sounds, a change in one direction is often more likely than the inverse sound change. The sound correspondence in our example occurs between a vowel and the trill

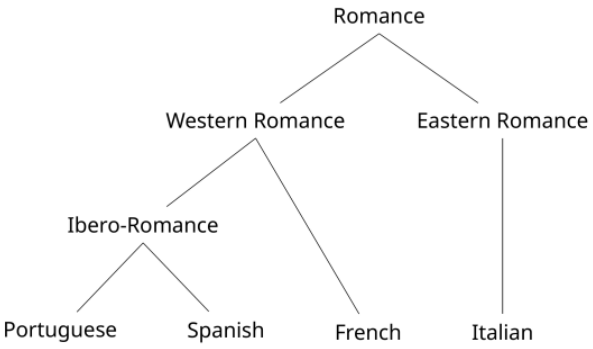

Figure 1.3: The internal structure of Romance, reduced to four exemplary languages.

/r/, both voiced sounds with high sonority. In this context, it is very likely for voiceless obstruents to become voiced, whereas the opposite case is quite unlikely. This tendency makes **p* the most likely proto-sound candidate, even though it is only retained in one language, Italian.

But wouldn't that violate the *majority vote* principle? At first sight, yes – we observe *b* in both Portuguese and Spanish, so it occurs more often than the other candidates. However, so far we have disregarded the internal structure of the Romance language and, for the sake of simplicity, pretended that they all were direct descendants of "Proto-Romance". Of course this is not true, as shown in Figure 1.3. Portuguese and Spanish are closely related to each other, and then are joined with French to form the Western Romance subgroup. Italian as the only Eastern Romance language in this sample is therefore more distantly related to the other languages than they are among each other. This has some important implications for the *economy* and the *majority vote* principles: Instead of trying to reconstruct Proto-Romance directly from the four extant languages at the same time, we now have intermediary nodes, allowing for a more precise understanding about which changes happened at which time. For the sound correspondence $p : b : b : v$, it does not matter which of the three sounds we postulate as proto-sound in terms of economy: Each option requires two sound changes along the tree; one from Romance to one of its children, and another one from Western Romance to either Ibero-Romance or French. Proposing **b* as the proto-sound therefore would require a sound change $b > p$ for Eastern Romance, and $b > v$ for French. In terms of economy, this has the same implication as proposing **p* as the proto-sound – either way, there are two different sound changes that happened along different branches of the tree. Due to the implications of *directionality* however, it is much more likely to reconstruct **p*, suggesting $p > b$ for Western Romance, than to reconstruct **b* which would imply the typologically marked devoicing between voiced sounds.

Earlier, we also accepted that **k* is the best reconstruction for the sound correspondence $k : k : k : ʃ$. Revisiting this correspondence, both *economy* and *directionality* provide additional evidence for reconstructing **k*: On the one hand, the reconstruction of **ʃ* would require the sound change $ʃ > k$ to happen twice independently

from each other; once for Eastern Romance, and once for Ibero-Romance. This violates the principle of *economy*. On the other hand, the palatalization of *k* to *ʃ* (usually via *tʃ*, which was also the case for French) is typologically quite common, while a *(t)ʃ* hardly ever becomes *k*. The sounds in question therefore exhibit a clear *directionality* which also shows **k* to be the preferable proto-sound.

The principle of *directionality* is therefore arguably the most important principle to reconstruct the most likely proto-sound. However, it poses a major problem: It completely relies on the experience and the intuition of the operating linguist. While the other principles can be applied directly on the data, the judgement about the plausibility of a sound change relies strongly on the linguist's intuition. Some phenomena occur frequently enough across different language families for linguists to agree that they are, in fact, very likely to happen; that includes virtually all sound changes that have been introduced so far, like k-palatalization, voicing of intervocalic obstruents, or loss of word-final nasals. For less frequent sound changes however, the intuition about the plausibility of a sound change might differ drastically from linguist to linguist, depending on their area of specialization and thus their experience (Kümmel, 2007). Most historical linguists are specialized on either a geographical area or a certain language family, and some sound changes can frequently occur within such a sample of languages while being infrequent outside of it. However, linguists are bound to rely on their intuition when assessing whether a sound change is likely or not – up to this day, there are hardly any frameworks that successfully quantify the typology of sound change. The closest approximation to that are two documents that exhaustively collect reconstructed sound changes for a given set of languages. Kümmel (2007) collects data about which sound changes can be found in which languages, and by extension provides information about which sound changes can be commonly observed in his data. However, his works mainly relies on Indo-European data, with the addition of some Uralic and Afro-Asiatic data, and therefore can not be regarded typologically independent. The second corpus is the Index Diachronica (2016) that collects a plethora of sound changes, indexed by language (sub-)family. While this technically would allow a reader to assess how common a given sound change is typologically, the indexing strategy would require them to go through each node of each family and to check whether the sound change in question is present there. It must also be noted that the Index Diachronica is an open source project whose contributors and editors are conlangers, an internet community of people who invent constructed languages for fun. It has therefore never been peer-reviewed by experts and is therefore likely to contain some inaccuracies. *Directionality* and by extension the typological markedness of a sound change in question therefore still mainly rely on the linguist's intuition and can hardly be backed quantitatively in a systematic way.

The last important factors to determine the ideal proto-sound reconstruction is what Campbell (2013) calls *phonological* and *typological fit*. Simply speaking, the set of reconstructed proto-sounds should result in a sensible phonological system for the proto-language. The *phonological fit* hereby is concerned with the structure of the phonology and exploits the strong tendency of languages to have a certain

symmetricity in their phoneme inventory. For example, if a language has a series of voiceless stops /p, t, k/ and if it has voiced stops, it is likely that they are /b, d, g/, mirroring the structure of the voiceless stops. This kind of symmetricity is often observed in languages all across the world and is therefore also likely to be observed in proto-languages. Consider the case where two related languages exhibit the sound correspondence *d : r*. Both sounds are quite likely to change into each other, so the principle of directionality can not help to determine which of these two sounds has been the proto-sound; and without further information from other hypothetical related languages, the other principles can not be applied either. However, if the reconstructed phoneme inventory contains the three voiceless stops **p, *t, *k*, as well as the voiced stops **b, *g*, it is very likely that the proto-language also contained **d*, making it the more probable candidate.

The closely related *typological fit* is concerned with choosing sounds that commonly occur in phoneme inventories across different languages over sounds that are not found in many languages: Typologically unmarked sounds should be preferred over typologically marked ones, and the resulting phoneme inventory should exhibit typologically common patterns. Structuralists have proposed some universals about the systematicity of phonological systems. For example, Jakobson (1958) claimed that the presence of aspirated voiced stops in a phonological system strictly implies that the language in question also contains plain (=unaspirated) voiced stops. While the strict implication does not hold true – there are in fact some languages that exhibit such a system – it still applies to the grand share of phonological systems across the world (Kümmel, 2007). A historical linguist therefore should refrain from reconstructing typologically marked sounds like aspirated voiced stops in absence of their unmarked (non-aspirated) counterparts, unless they have a very good reason for doing so.

Reconstructing the correct, or at least most likely proto-sound therefore often requires a delicate balance between evidence from the data and typological guidelines. In cases where these two factors contradict each other and suggest different reconstruction options, linguists have to make sensible case-by-case decisions: Either, they have to suggest typologically marked phenomena based on good evidence in the data, or they have to propose reconstructions that seem less plausible based on the data, but are typologically justified.

1.3 The abstractionist-realist debate

In previous sections, I have already made clear that reconstructions are of *phonological* rather than *phonetic* nature. They are therefore not able to represent how exactly a proto-language sounded like – just as synchronous phonology is already an abstraction of the actual, physical properties of the sounds and the language. Symbols like /t/ and /d/ are regularly used to denote coronal stops in Spanish and English, disregarding that they are usually articulated as dentals in the former and as alveolars in the latter language. Likewise, /r/ is often used to encompass all kinds of rhotic sounds, regardless of their exact articulatory realization – as long as there is no phonemic difference between them, phonologists are happy to attribute

phones like [r, ɾ, ɹ, ʁ] to the phoneme /r/, following the intuition that those are slightly different phonetic realizations of the same sound (Hayes, 2008).

Naturally, that raises the question how realistic phonological reconstructions can be, when we already encounter such abstractions in synchronous phonology. After all, the successful application of the comparative method in a strict sense only produces a set of contrastive units by accumulating sound correspondences and grouping non-contrastive units together. The *symbol* for each of these units however has to be chosen by the linguist (Anttila, 1989). For theory-created objects like these, there are two polar philosophical stances on how to interpret them: Realists argue that theories can generate realia, real world objects of some kind; whereas abstractionists view these generated objects as devices of a pure algebraic nature, that can and should be used for calculations within the theory, but do not have any real world value (Lass, 2017). In our particular case, abstractionists therefore see reconstructed proto-sounds as purely relational symbols that can express the phonemic structure of the proto-language, but are completely independent from any phonetic value. For realists on the other side, every reconstructed phoneme inherently holds a phonetic value that they try to reconstruct as faithfully as possible.

The abstractionalist stance has been strongly represented by the structuralist school. Famously, Kuryłowicz (1964) even argued that phonetics are not to be considered part of linguistics in a strict sense:

> Physiological speculations [...] do not grasp the *linguistic* essence of [...] changes, the shift of the *internal* relations of the elements in question being the only pertinent fact. Once we leave language *sensu stricto* and appeal to extralinguistic factors, a clear delimitation of the field of language research is lost. (cited from Lass 2017)

Essentially, Kuryłowicz proposes to separate phonology completely from phonetics. After all, the structural relations between phonemes – both synchronously and diachronically – can be expressed by any set of arbitrary symbols. However, reducing proto-sounds to such arbitrary symbols regardless of any phonetic value poses a problem: It completely disregards the phonetic reality of the data, the languages we base our reconstructions on (Lass, 1993, 2017). Consider a cognate set like {Sanskrit *pitar-*, Latin *pater*, Old English *fæder*, German *Vater*} for the concept father, where all reflexes have a labial obstruent as their initial sound. It would be unreasonable to disregard these obvious phonetic similarities! How could we explain that the reflexes of this sound correspondence regularly surface as labial obstruents, if we do not assume any phonetic value for the proto-sound (Lass, 2017)?

Cognacy is proven by successfully applying sound laws and demonstrating systematic sound correspondences. It is true that this can technically be done only by the means of relational symbols, an algebra linking abstract phoneme representations. However, that poses a major problem in the initial stage – at the first assessment of the language data, nothing about sound correspondences and sound laws is known. As described in Section 1.1.1.2, the first step must be to identify preliminary sets

of putative cognates. For this first cognacy judgement, phonetic similarity is required as a heuristic – disregarding phonetic information would essentially result in an infinite search space, where each pair of words can be compared in search for regular sound correspondences.

"Phonology without phonetics is perverse", states Lass (2017) and concludes that linguistic reconstructions therefore must be at least realist to a certain extent: They must resemble a language in a form that it might have been spoken at some point in time by a real community. Following the Uniformitarian Principle, we must assume that languages that have been spoken in the past behave just like languages that are spoken nowadays, and therefore observed tendencies of modern languages also hold true for extinct ones. Besides that, the comparative method crucially exploits how sound changes occur in a quasi-regular way, and how certain kinds of sound changes can frequently be observed in various languages across the world – an aspect that explicitly relies on the phonetic nature of sounds and can not be separated from it (Anthony and Ringe, 2015).

However, the weaker implication of the abstractionist stance must not be disregarded either. Reconstructed proto-sounds can not fully reflect the phonetic value of a proto-sound, it can provide an approximation at best, which can be more or less accurate depending on the circumstances. A reconstructed */k/ for example does not specify any exact phonetic information, like the location of the dorsovelar closure, or the voice onset timing (Lass, 1993), and this kind of information is arguably impossible to collect without recordings of the language (which can not be obtained unless time machines are invented). However, */k/ is not just an arbitrary symbol to label a given sound correspondence either – it suggests that the sound in question was probably realized as a velar (or at least dorsal) stop of some kind, based on the evidence from the modern languages.

As it is so often, the truth lies somewhere in between the two polar nodes, although the realist side in this case seems to have a stronger claim for their position (Anttila, 1989). All reconstructed symbols imply at least some phonological features, they contain some information about their probable phonetic realization. Every symbol that represents a proto-sound must be viewed as a broad symbol that encompasses many different phonetic realizations, just as in synchronous phonology. Some symbols however are broader than others (Lass, 2017): The phonetic value of the Indo-European laryngeals up to this day are highly disputed, which is why scholars use abstract symbols to represent them. However, all laryngeals must be viewed as realia, since they have had a phonetic value, although it is unknown to us. When reconstructing prehistoric languages, the goal must be to approximate the phonetic reality of these languages as closely as possible, while accepting that the exact physical and articulatory details can not be recovered. If we had a time machine which enabled us to talk to speakers of a reconstructed proto-languge, like Proto-Indo-European, we would therefore expect that the actual language would not sound exactly as the reconstruction, but resemble it enough to be intelligible to some degree (Lass, 1993, 2017). As Lass (1993) concludes: "Reconstruction does not give us back a language [...], but it is not an uninterpreted algebra without substantive content either."

2 Computational Approaches

With the rapidly expanding availability of computational power, *computational historical linguistics* have emerged as a new branch in computational linguistics over the past two decades, and the model described in this thesis is a further contribution to this relatively young line of research. Throughout this chapter, I will give an overview over the history and current state-of-the-art methods applied in computer-assisted language comparison and discuss their respective strengths and shortcomings. I will then introduce EtInEn, the framework within which this thesis and the corresponding model was developed, and finally I will describe which open problems my model addresses and thus its main innovations towards the field.

2.1 Computer-Assisted Language Comparison

For roughly two decades, scholars have been investigating the potential of enhancing historical linguistics by the means of computational methods (List, 2022). Most of these approaches aim at imitating the comparative method or parts of it by applying statistical models and machine-learning algorithms on multilingual lexical data. The majority of work in this field has focused on finding good computational solutions for sub-tasks of the comparative method, like detecting cognates or reconstructing proto-forms from *a priori* defined cognate sets. A few notable exceptions that actually intend to set up a computational pipeline that imitates the whole comparative method are described in Section 2.1.3 as well.

I will discuss current approaches to different tasks over the course of the next subsection, defining commonly addressed sub-tasks following List (2022), whose structure itself is based on the workflow outlined Ross and Durie (1996), which I have used in the previous section.

2.1.1 Proof of relatedness

Scholars are generally quite hesitant to employ computational methods when it comes to showing that languages are genetically related, and there is so far no statistical test that is generally accepted for that purpose. The reluctance to use such tests can be explained by three main factors: Firstly, all tests proposed so far only rely on lexical data. However, most scholars demand evidence beyond lexical similarities in order to accept a genetic relationship between languages, as discussed in Section 1.1.1.1. Even the best tools for identifying cognates could therefore not satisfy the requirement of offering structural evidence for a proposed genetic relationship. Secondly, the philosophical stance of classical historical linguists on identifying genetic relationships is quite distinct – the task is essentially treated like a mathematical proof, which is required to reconstruct a proto-language. A proof leaves no room for uncertainties; either something is proven, or it is not. A test on the other hand is fundamentally different in design, since its result lie in a continuous space of belief or significance. This merely offers approximations to a problem, resulting in false positives and false negatives – byproducts that the strict idea of a proof does not allow (List, 2022). At last, most of the code and/or data from the tests has not been made public, which makes it impossible to quickly replicate the results or to apply a proposed test on another dataset. That would require building the whole method from scratch, which would be an unreasonable effort, given that scholars usually work with already well-established language families (List, 2022).

The tests that have been designed for showing genetic relatedness all rely on a basic idea. They compare the basic vocabulary of the languages in scope with each other, typically using a Swadesh list (Swadesh, 1955) or some variation of it. They observe correspondences in this lexical data and cast them into probability distributions in order to calculate the probability of obtaining a distribution like the one observed in the data. The principle of statistical significance can then be applied: If the probability for randomly generating the correspondences found in the data is below a certain, quite low threshold, there is a statistically significant correlation, which indicates that the languages in question are related.

Different strategies for observing and counting correspondences between the languages have been employed. Ringe (1992) and Baxter and Manaster-Ramer (1996) only consider the first consonant in word pairs and calculate the probabilities of a given pair by estimating probability distributions from the observed pairs. Turchin et al. (2010) and Kassian et al. (2015) follow a similar approach and also compare only initial consonants to each other, but conflate them into 10 consonant classes defined by Dolgopolsky (1964) and only count exact matches. Blevins and Sproat (2021) infer similarity metrics from pairwise alignments and test the data against artificial wordlists that are randomly generated by lexical language models, which combine sounds to word forms following a probability distribution over n-grams (Miller et al., 2020). The former approaches only compare words that are identical in meaning, whereas the latter method also considers word pairs with similar, but not identical meanings, making use of colexifications obtained from the CLICS^3 database (Rzymski et al., 2020).

None of these methods however has found general acceptance among scholars so far, mainly due to the aforementioned drawbacks. List (2022) argues that this can only change as soon as some crucial prerequisites are fulfilled. He demands a uniform gold standard dataset paired with clear and sensible error metrics that can quantify how well a proposed test performs on different languages and language families. Furthermore, authors should release the source code for their tests in order to make it possible to measure their performance on other datasets. As soon as different statistical tests are available that can be used easily across different datasets, and as soon as there is a way of assessing the strengths, weaknesses, and overall quality of those tests, they have the potential to become a powerful tool to provide evidence for suggesting deep language relationships.

2.1.2 Detection of cognates and sound correspondences

The next step of the outlined workflow consists in detecting cognates and systematic sound correspondences. Both tasks were considered to be quite difficult to solve computationally, and while there are still very few approaches to automatically identifying sound correspondences, the task of automated cognate detection has received quite some attention over the last decade, resulting in some promising techniques (List, 2022).

The vast majority of workflows for cognate detection consists of two main stages. In the first stage, all words of a multilingual wordlist that express the same concept are compared to each other pairwise. For each form pair, some sort of similarity or distance metric is calculated, indicating how similar the two phonetic forms are to each other. In the second stage, the forms are then clustered in cognate sets based on the individual values. This is usually done by means of agglomerative clustering approaches like UPGMA (Sokal and Michener, 1958), joining forms and clusters together until a certain, typically user-defined threshold is reached (List, 2022).

The technicalities of different methods usually differ from each other in terms of how they calculate phonetic similarity, which clustering algorithm they employ, and which threshold is chosen. For calculating phonetic similarity, Hall and Klein (2010) employ a parameterized edit distance, which Hauer and Kondrak (2011) enhance by adding various string similarity features such as longest common prefix or the number of common bigrams. More recent statistical approaches are based on pairwise language-specific scoring schemes (List, 2014; Rama, 2016) or on pairwise similarities between sounds by the means of Pointwise Mutual Information (Jäger, 2013; Jäger et al., 2017; Dellert, 2018). Furthermore, there are some variants of this workflow that try to identify partial cognates or cognate morphemes (List et al., 2016b), search for cognates across different concepts (Wu et al., 2020), or employ methods for community detection instead of applying flat clustering algorithms (List et al., 2017).

A quick and easy alternative to calculating phonetic similarity metrics is again to resort to consonant classes or sound classes (Turchin et al., 2010; List, 2014; Rama and List, 2019). This approach is much faster and computationally less expensive

than those based on pairwise phonetic similarity, however they can not compete with them in terms of accuracy (List et al., 2017; List, 2022).

Hardly any reliable methods for identifying sound correspondence patterns have been proposed so far. The main challenge in identifying those correspondences is that most cognate sets in wordlists are not complete – there is not a reflex for every proto-form in every extant language that is reflected in the database. I have briefly discussed incomplete correspondence patterns and how they belong to full correspondence patterns in Section 1.1.1.2. When aligning incomplete cognate sets, such incomplete correspondence patterns are frequently found. All of these incomplete correspondence patterns must then be matched and clustered together to full correspondence patterns. List (2019a) employs a clique covering technique to partition all individual correspondences to the smallest possible number of 'cliques', where each clique only contains (incomplete) correspondences that are compatible with each other. Two correspondence patterns are compatible with each other when they either have the same sound for the same language, or at least one of them underspecifies it (has a gap for that language). $p : b : \varnothing$ therefore is compatible with $p : b : b$ or $p : \varnothing : b$, but not with $b : b : b$.

Bodt and List (2022) report that the approach by List (2019a) performs well for predicting missing reflexes, however List (2022) points out that this techniques still "needs to be tested and applied to more language families". Nonetheless, this technique currently states the only feasible solution for this task.

2.1.3 Phonological reconstruction and sound law inference

In the classical application of the comparative method, phonological reconstruction goes hand in hand with the induction of sound laws, as already described in Section 1.1.1.3. List (2022) states that scholars have not tried to imitate this classical iterative process, but instead have been focusing on the two related tasks of *supervised phonological reconstruction* and *ancestral state reconstruction*. The former task is solved in a supervised machine-learning setting by training a model to learn correspondences from source to target language, which it then applies to predict unseen forms in the target language. The latter task iterates over a phylogenetic reference tree and uses the synchronous forms from pre-defined cognate sets to reconstruct the most likely proto-forms.

In contrast to several approaches towards automated phonological reconstruction, the task of identifying sound laws has received little to no attention so far. Hruschka et al. (2015) employ a Markov Chain Monte Carlo model as used for investigating concerted evolution in biology, however, their model is only able to infer unconditioned sound laws. The only approach towards automatically inferring conditioned sound laws was presented by Daneyko (2020) who uses Probabilistic Soft Logic (PSL) and trigram models based on observed sound correspondences.

The task of supervised phonological reconstruction is technically identical to the task of reflex prediction (List, 2022), since in both cases, there is attested training data form the target language, from which the model should learn to generalize

and predict forms outside the scope of the training data. Whether that target language is extinct or still spoken today makes no difference to the technical set-up. The main difference lies in the availability of data – reflex prediction can be used for any modern language with a basic amount of available data, and Bodt and List (2022) argue that it can be a useful tool for extending lexical data of low resource languages. Reflex prediction is mostly viewed as a machine translation task (Beinborn et al., 2013; Dekker and Zuidema, 2020; Fourrier et al., 2021), an alternative approach is outlined by Bodt and List (2022) who identify systematic sound correspondences (using a clique-covering introduced by List 2019a) and use them to recover missing reflexes.

The only usage of such a technique for reconstructing proto-forms of an extinct language was proposed by Ciobanu and Dinu (2018) and expanded by Meloni et al. (2019), who reconstruct Latin forms from five modern Romance languages. In both papers, the supervised reconstruction is seen as a machine translation task, for which the earlier employs conditional random fields, while the later uses recurrent neural networks. However, there are hardly any extinct languages that are documented well enough to make a supervised approach feasible, and already having attested data from the proto-language essentially defeats the purpose of the comparative method, which is used for recovering ancient languages without surviving records. The application of supervised methods for reconstructing proto-languages is therefore quite limited.

The task of ancestral state reconstruction is conceptually closer to the comparative method, since it does not rely on previous expert annotations or reconstructions. These techniques only need a phylogenetic tree, along whose branches a proto-form has evolved into its modern reflexes, and cognacy judgements. Jäger and List (2016) employ some conceptually simple algorithms for reconstructing proto-forms, including all intermediary stages such that a proto-form is reconstructed for all nodes of the phylogenetic tree. Despite the conceptually simple approach, they report moderately good results, indicating the potential of employing more sophisticated methods for ancestral state reconstruction. One of the employed reconstruction principles, Maximum Parsimony (which is roughly equivalent to the reconstruction principle of economy, see Section 1.2.3), is explained in more detail in Section 3.1.2.

Jäger (2019) applies this approach to reconstruct Latin forms from modern Romance languages taken from the ASJP database (Wichmann et al., 2013), however he reports rather disappointing results. The underwhelming quality of the reconstructed forms can be explained by several factors: The model reconstructs Proto-Romance, but is evaluated against Classical Latin, which is not identical. Furthermore, the employed reconstruction method simply chooses the most probable proto-sound per column in the sequence alignment from a fixed alphabet. It is therefore neither aware of the context of the sound, nor of the phonology of the proto-language in question – both of which are important factors for a successful reconstructions, as previously discussed. Up to this day, the grand share of techniques for ancestral state reconstruction struggle with overcoming these problems and including these important pieces of information. Over the course of the thesis,

I will use the terms *naïve* and *language-agnostic* for describing reconstruction algorithms that do not process information about the context and the proto-language respectively.

A notable exception to this problem is the work by Bouchard-Côté et al. (2013) who use stochastic string transducers to reconstruct Austronesian proto-forms, based on lexical items from the Austronesian Basic Vocabulary Database (Greenhill et al., 2008). Following earlier experiments on Romance lexical data, they employ phoneme-level edit models conditioned on the immediate neighbors (Bouchard-Côté et al., 2007a,b) and language-specific n-gram models called markedness (Bouchard-Côté et al., 2009). The former feature allows the model to reconstruct proto-sounds with regard to the context they stand in, while markedness efficiently restricts the number of plausible proto-sounds and sound sequences. Furthermore, some parameters are shared across the branches of the tree, enabling the model to learn general tendencies (Bouchard-Côté et al., 2009; Bouchard-Côté et al., 2013). The authors report very good results for reconstructing Austronesian, which needs to be taken with a grain of salt nonetheless, since Austronesian seems to be one of the simpler test cases, as suggested by Jäger and List (2018) who compare the performance of ancestral state techniques across different datasets.

Following that line of research, He et al. (2022) recently proposed a neural edit model that learns the probability of edit operations (insertions, deletions, substitutions) for individual branches in an estimation-maximization setting, guided by a bigram model for the proto-language. They report improved results for reconstructing Latin from modern Romance language, however they concede that their model is focused on optimizing parameters along branches locally and comes short in propagating information across the tree. Their model assumes a flat language tree, with Latin as the root and the modern languages as the leaves, which is not feasible for phonological reconstruction with higher time depth.

Although the task of ancestral state reconstruction *per se* only requires the cognate sets and the language phylogeny, it is notable that both Bouchard-Côté et al. (2013) and Jäger (2019) outline a full pipeline that includes automated cognate detection; the latter work even includes methods for demonstrating language relatedness and phylogenetic inference. This is arguably the closest approximation to automating the comparative method so far.

2.1.4 Open problems

Due to its relatively young age, the field of computational historical linguistics still exhibits a notable discrepancy between well-established methods and problems that have hardly been successfully addressed at all. Among the major open problems, List (2022) names that there is no generally accepted procedure for testing relationships between languages, that there are hardly any well-established methods for cognate detection beyond full cognates between words with the same meaning, and that automated induction of sound laws has been addressed sporadically at best. Furthermore, the techniques for most individual tasks can still be substantially refined, as discussed in the previous sections.

No computational method so far has been able to produce full-fledged etymological scenarios. Besides refining techniques for imitating individual steps of the comparative method, that essentially includes to detect and explain borrowings. While the automated detection of borrowing events has received some attention – mainly by treating them analogously to horizontal gene transfer in bioinformatics – all techniques that are currently available produced rather disappointing results (Köllner, 2021). An alternative approach to the automated detection of loanwords was recently proposed by Blaschke et al. (2022), who define a set of heuristics and employ them as rules within a PSL (Probabilistic Soft Logic) framework. While their approach looks promising on first individual test cases, it requires further, more systematic testing in the future.

2.2 EtInEn

This thesis and the model described in it were developed as part of EtInEn (**Et**ymological **In**ference **En**gine). EtInEn is an interactive software for historical linguists which is currently under development at the Linguistic Department of the University of Tübingen. It "works and communicates with the user in classical terms, but is supported by a probabilistic model that is used to quantify strength of evidence" (Dellert, 2019).

EtInEn provides the user with state-of-the-art methods for many tasks in computational historical linguistics, like cognate detection, sound correspondence identification, phonological reconstruction, sound law inference, and loanword detection. While EtInEn can technically be used as a fully automated pipeline for suggesting an etymological scenario, it is designed to be used step-by-step in an interactive way so the user can explore different ideas. By automating many routine tasks, it is able to indicate whether certain ideas are coherent with the data and/or previously defined ideas. EtInEn is able to process custom user-defined ideas and to base its suggestions according to them.

The model presented in this thesis will be used in two places within EtInEn. It will be used to inform the sound law inference module (Daneyko, 2020) about the typological markedness of sound changes, assisting it in inferring phoneme inventories of proto-languages and sound laws. Furthermore, it can be used to generate naïve, language-agnostic reconstructions that can serve as a first, rough approximation to what a proto-language could have looked like.

2.3 Attempts at innovation

All models for ancestral state reconstruction that have been proposed so far are limited by the data they work on when predicting proto-sounds. Since the sounds from the input data are usually processed as atomic units, it is impossible for models to reconstruct any sound that has not been seen in the training data. List (2022) proposes that models could overcome this issue by either "learn[ing] common sound change processes from training datasets, or [...] turn[ing] to feature

representations of sounds". The core result of my thesis is a neural model that combines both ideas by learning common sound changes from global lexicostatistical data and operating on phonological feature representations, enabling it to predict the likelihood of sound changes between arbitrary sounds.

This comes with two crucial innovations. First of all, using phonological features makes the model very robust and enables it to process arbitrary sound changes – as long as both sounds in question can be represented by the means of phonological features. That drastically alleviates the problem that current models for phonological reconstruction have, which can only process a defined, finite alphabet of sounds that has been learnt.

Secondly, the model presented in this thesis constitutes a first step towards a quantitative typology of sound change. In Section 1.2.3 I have already outlined that there are no exhaustive frameworks that contain information about the frequency of certain sound changes and thus their likeliness. In classical historical linguistics, the assessment whether a sound change is plausible or not is mainly based on the linguist's intuition. Computational approaches naturally lack such an intuition – they can merely learn common sound correspondences from the present data. Since my model has been trained on large-scale global lexicostatistical data, it can simulate human intuition by informing reconstruction modules about the typological markedness of a given sound change.

3 Methodology

3.1 Preparing training data

3.1.1 Source datasets

The lack of quantitative data for sound changes that contain information about how often certain changes occur – and thus, which changes are typologically marked and which ones can be expected frequently – has been discussed in previous sections. That entails that a model can not be trained directly on such data, but that a workaround is required. Reconstructed proto-forms can serve that purpose: By aligning proto-forms with their respective reflexes in the daughter languages, it is possible to count the frequencies of transition between sounds.

The most straightforward way to generate data over sound transition counts would be to use either attested proto-forms (like Latin or Sanskrit) or expert reconstructions. However, this approach poses two problems: On the one hand, lexical data that includes reconstructions is still only sparsely available in a digital format, on the other hand, the little data that can be used is mostly limited to a few well-studies language families like Indo-European or Austronesian. Using such data would therefore result in a model that is likely to be poorly informed due to the low amount of training data, and even if it happened to learn well from the training data, it would be heavily biased towards sound changes that are specific to a few language families and therefore would not meet the requirement to be typologically relevant.

In order to come up with a good quantity of typologically diverse data, we must therefore turn to automatic reconstructions from synchronous lexical data. The amount of such data in digital formats is constantly increasing and convenient efforts towards a standardized data format have been made, enabling researchers to work on different databases with the same technical set-up. The most notable and resourceful collection of standardized lexical data is Lexibank (List et al.,

Dataset	Reference	Concepts	Varieties	Family / Genus
abvdoceanic	Greenhill et al. (2008)	191	418	Oceanic (Austronesian)
birchallchapacuran	Birchall et al. (2016)	125	10	Chapacuran
bodtkhobwa	Bodt and List (2019)	662	8	Kho-Bwa (Sino-Tibetan)
bowernpny	Bowern and Atkinson (2012)	344	190	Pama-Nyungan
carvalhopurus	de Carvalho (2021)	205	4	Purus (Arawakan)
chaconarawakan	Chacon (2017)	102	8	Arawakan
chaconbaniwa	Chacon et al. (2019)	243	14	Arawakan
chaconcolumbian	Chacon (2017)	128	69	Colombia*
constenlachibchan	Constenla Umaña (2005)	110	25	Chibchan
crossandean	Blum et al. (2021)	150	50	Andes*
davletshinaztecan	Davletshin (2012)	100	9	Uto-Aztecan
dravlex	Kolipakam et al. (2018)	100	20	Dravidian
dunnaslian	Dunn et al. (2013)	146	32	Aslian (Austro-Asiatic)
felekesemitic	Feleke (2021)	150	21	Semitic (Afro-Asiatic)
galuciotupi	Galucio et al. (2015)	100	23	Tupian
gerarditupi	Ferraz Gerardi and Reichert (2021)	244	38	Tupian
grollemundbantu	Grollemund et al. (2015)	100	424	Bantu (Atlantic-Congo)
hsiuhmongmien	Hsiu (2015)	315	12	Hmong-Mien
leeainu	Lee and Hasegawa (2013)	199	19	Ainu
leejaponic	Lee and Hasegawa (2011)	210	59	Japonic
leekoreanic	Lee (2015)	246	15	Koreanic
lionnetyotonahua	Lionnet (1985)	364	15	Uto-Aztecan
liusinitic	Liú et al. (2007)	203	19	Sinitic (Sino-Tibetan)
luangthongkumkaren	Luangthongkum (2019)	341	11	Karenic (Sino-Tibetan)
lundgrenomagoa	Lundgren (2020)	1,807	3	Tupian
mannburmish	Mann (1998)	391	7	Burmish (Sino-Tibetan)
mcelhanonhuon	McElhanon (1967)	140	14	Nuclear Trans New Guinea
meloniromance	Meloni et al. (2019)	5,419	6	Italic (Indo-European)
nagarajakhasian	Nagaraja et al. (2013)	200	6	Khasian (Austro-Asiatic)
northeuralex	Dellert et al. (2020)	1,016	107	Eurasia*
peirosaustroasiatic	Peiros (2004b)	100	109	Austro-Asiatic
peirosst	Peiros (2004a)	110	128	Sino-Tibetan
pharaocoracholaztecan	Pharao Hansen (2020)	100	9	Uto-Aztecan
ratcliffearabic	Ratcliffe (2021)	100	14	Arabic (Afro-Asiatic)
robinsonap	Robinson and Holton (2012)	398	13	Alor-Pantar
saenkoromance	Saenko (2015)	110	43	Romance (Indo-European)
sagartst	Sagart et al. (2019)	250	50	Sino-Tibetan
savelyevturkic	Savelyev and Robbeets (2020)	254	32	Turkic
sidwellbahnaric	Sidwell (2015)	200	24	Austro-Asiatic
sidwellvietic	Sidwell and Alves (2021)	116	33	Vietic (Austro-Asiatic)
simsrma	Sims (2020)	233	11	Qiangic (Sino-Tibetan)
starostinhmongmien	Starostin (2015a)	110	20	Hmongic (Hmong-Mien)
starostintujia	Starostin (2015b)	109	5	Tujia (Sino-Tibetan)
syrjaenenuralic	Syrjänen et al. (2013)	173	7	Uralic
utoaztecan	Greenhill et al. (2022)	121	46	Uto-Aztecan
walworthpolynesian	Walworth (2018)	210	31	Polynesian (Austronesian)
wichmannmixezoquean	Cysouw et al. (2006)	110	10	Mixe-Zoque
yanglalo	Yang et al. (2010)	1,000	8	Lalo (Sino-Tibetan)
zhangrgyalrong	Zhang et al. (2019)	120	7	Sino-Tibetan

Table 3.1: Overview of Lexibank datasets used for generating training data. Entries in the last column marked with an asterisk (*) do not refer to phylogenetic classifications, but to geographical areas from which the respective dataset encompasses genetically diverse data.

2022a), which currently[1] contains 147 multilingual wordlists in a standardized *Cross-Linguistic Data Format* (CLDF; Forkel et al. 2018). Besides providing data in a standardized format, Lexibank also adds relevant meta-data to the databases by linking its contents to external frameworks: Language varieties are provided with references to Glottolog (Hammarström et al., 2022), concepts are linked to Concepticon (List et al., 2016a), and sounds are further standardized by a reference to the *Cross-Linguistics Transcription System* (CLTS; Anderson et al. 2018).

Lexibank is therefore a good and convenient resource to accumulate lexical data from different sources that meets the requirement of being typologically and genetically as diverse as possible. Table 3.1 lists the Lexibank datasets that I used to compile my dataset, based on which the training data was produced later. All of these databases fulfil two relevant criteria, as they are cognacy-annotated and include phonetic transcriptions in IPA. The need for the latter criterion is straightforward, since IPA representations are required to describe sounds in a standardized way and thus count transitions between them on a global scale. Cognacy annotations on the other hand are not a technical requirement per se, since there are relatively good methods for automated cognate detection (see Section 2.1.2); however adding another layer of automation would add a higher degree of noise in the data due to errors of the algorithm. Since the automatic reconstructions are already expected to be quite noisy (the details will be discussed in Section 3.1.2), and there is already a substantial amount of cognacy-annotated data that can be obtained from Lexibank, datasets that lack cognacy annotations were not regarded.

The basic idea of merging the source databases into one large database is simple enough, however, there are some special cases to consider when different source datasets overlap with regard to the languages they investigate – after all, it is not desired to process the same data from the same language multiple times. Two important questions therefore had to be answered: How should duplicate languages (that are included in more than one source database) be handled? And should lexical items that stem from different datasets even be compared to each other?

It is probably best to discuss these questions by means of a concrete example where the scopes of two datasets overlap. The `lundgrenomagoa` database (Lundgren, 2020) investigates the closely related Tupian languages of Kokama, Omagua, and Tupinambá – all of which are also included in `gerarditupi` (Ferraz Gerardi and Reichert, 2021), which is a collection of lexical data from 38 Tupian languages. The languages included in `lundgrenomagoa` are therefore a proper subset of those in `gerarditupi`; however, the former dataset encompasses a much larger amount of concepts. Of course, we would like to exploit both the concept depth of `lundgrenomagoa` and the language coverage of `gerarditupi` – but at the same time, we want to avoid having the same information twice in the final dataset.

There are two thinkable approaches to handling this situation, and which one to use depends on how the second of the questions stated above is answered. If it is

[1] accessed on Nov 3rd, 2022

desired that words from different source databases are directly compared to each other, it is technically possible to merge the datasets via their Concepticon ID. Although the concept IDs within the individual databases naturally differ from each other, one can assume that two concepts (from different databases) that map to the same Concepticon ID are identical. By analogy, that means that lexical items that map to the same Concepticon ID and the same Glottocode (the unique identifiers within Glottolog) are identical in meaning and language, and should therefore refer to the same lexeme. Removing duplicates after merging two partially overlapping datasets, so that there is only one entry per language and concept, is technically feasible.

If all of these duplicates were actually identical (i.e., containing the same phonetic transcriptions and referring to the same cognate sets), there would be no conceptual problem with merging partially overlapping datasets, while retaining all information from both. However, given the non-discrete nature of sounds in human speech, different sources often provide slightly different transcriptions for the same sounds or words. If we wanted to merge two datasets where some identical forms are transcribed differently, the question arises: Which form should be used and which form should be disregarded? Choosing one dataset over the other in such cases of "merge conflict" might lead to systematic inconsistencies in the transcriptions of the merged dataset, but is arguably still better than choosing a random form each time, which would lead to asystematic inconsistencies. But even with the inconsistencies being systematic, we would encounter many cases where they would suggest sound correspondences (and, via analogy, also sound changes) between two sounds in seperate languages that actually refer to the *same* sound, which is just transcribed differently in different sources. While it might be true that sounds pairs (or sets) where this happens are usually phonetically so similar to each other that they can easily change into one another, we are only interested in capturing phonemic sound changes rather than subphonemic ones. Since the model's main purpose is to assist in phonological reconstruction, the training data is designed to emphasize phonemic sound changes over merely phonetic ones.

In order to avoid the aforementioned challenges that can arise from differences in phonetic transcription, I decided to include data from only one dataset per language. In cases of multiple datasets covering the same language, only data from the dataset with the most concepts was preserved, while data from the other datasets for the language in question was disregarded. In the case of our examples, the data from `lundgrenomagoa` was used for Kokama, Omagua, and Tupinambá, while the `gerarditupi` data for those three languages was disregarded. The latter dataset was however still used as a data source for all other languages it contains. As discussed previously, it is also not desirable to reconstruct proto-forms from data stemming from different sources; therefore, merging cognate sets via Concepticon was deliberately left out. In practice, that means that the three languages from `lundgrenomagoa` are excluded when reconstructing Proto-Tupian from `gerarditupi` data – neither do we want to include data from `lundgrenomagoa` due to the aforementioned risk of inconsistent transcriptions distorting the observed sound changes; nor do we want to end up with duplicate information, which would

happen when including data for those languages from `gerarditupi`.

The core principle that results from these choices is that information from different source should be kept separated in the combined dataset. This choice was made in order to avoid observing sound changes that are not phonemic, but rather a result of different transcription practices. For every language, only information from one source can be retained, for which the source dataset that contains the most concepts is chosen, in order to retain as much data as possible. Merging all datasets listed in Table 3.1 according to these principles resulted in a large, combined dataset that contains a total of 379,407 lexemes and 97,564 cognate sets (36,878 of which are singletons), while including 1,942 languages from all continents and 51 different language families.

3.1.2 Estimating sound transitions

In order to count sound transitions and by that make claims about frequency or markedness of sound changes, it is necessary to reconstruct a proto-form at *every* relevant node within a language tree. The cognate sets from the lexical data described in the previous section, paired with a reduced phylogenetic tree for the languages in question, builds the foundation for obtaining these reconstructions.

The initial reconstructions follow the simple and intuitive principle of Maximum Parsimony, denoting that the optimal phylogenetic tree should contain the least possible changes. Originating from evolutionary biology, the principle was first outlined by Farris (1970) and Fitch (1970) and was originally applied to find the optimal phylogenetic tree from unclassified sequences. Sankoff (1975) then devised an algorithm that was able to reconstruct missing forms in a given, finite tree, according to the same maxim. Whether applied for phylogenetic inference or for reconstruction, the target of the principle stays the same: The resulting tree should be the one with the lowest possible sum of branch lengths. Branch lengths can be measured by a distance metric between a parent and a daughter node; in the simplest case, it is just the normal edit distance (Levenshtein, 1966). A Maximum Parsimony tree therefore globally minimizes the distances across its branches, which is equal to the total number of edits when employing edit distance.

Sankoff's algorithm requires all input sequences – the words within a cognate set – to be aligned to each other, because it needs the information about which sounds correspond to each other. Multiple sequence alignments were obtained using the implementation of T-Coffee (Notredame et al., 2000) contained in EtInEn, an algorithm that aligns multiple sequences based on all pairwise alignments between the relevant sequences. Those pairwise alignments were generated using Information-Weighted Sequence Alignment (IWSA; Dellert 2018), a modified version of the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). Gaps that naturally result from aligning sequences of unequal lenghts are henceforth treated as regular symbols, as if it were a usual sound. Based on the aligned cognate sets, Sankoff's algorithm was applied to reconstruct proto-forms according to the Maximum Parsimony princple.

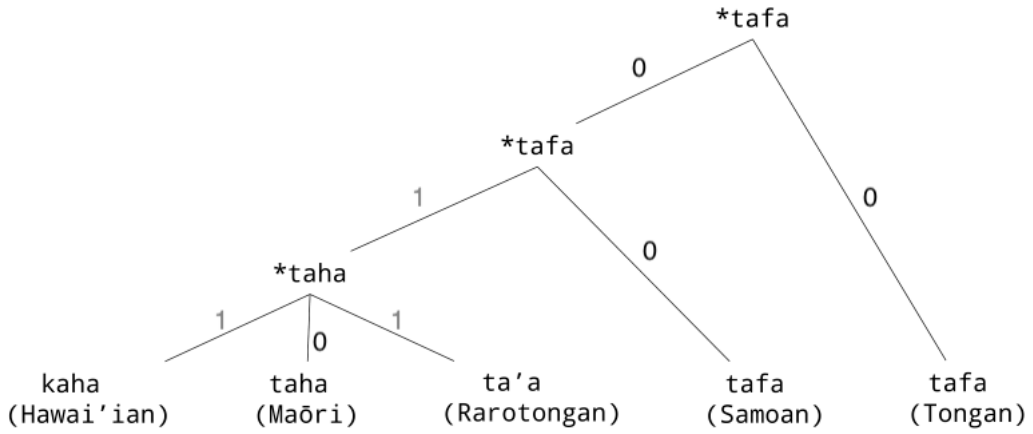

Figure 3.1: Illustration of a Maximum Parsimony reconstruction based on a cognate set from five Polynesian languages.

To illustrate what Maximum Parsimony reconstructions look like in practice, consider the cognate set given in Figure 3.1. Those words in the five chosen Polynesian languages mean *side* or *edge* and are clearly cognates (Campbell, 2013). The figure shows how a minimal language tree is reduced from Glottolog (Hammarström et al., 2022) such that its only leaves are the languages contained in the cognate set in question – in this case, the five Polynesian languages of Hawai'ian, Maōri, Rarotongan[2], Samoan, and Tongan. Hawai'ian, Maōri, and Rarotongan all belong to the Central-Eastern Polynesian group (the lowest node), which itself belongs to the Nuclear Polynesian group (the middle node), just like Samoan. The root node of the minimal tree is always the lowest common ancestor of the languages reflected in a cognate set, which in this case is Proto-Polynesian. Of course, there are many more linguistic subdivisions of the Polynesian languages than those reflected in this minimal tree; however the tree's purpose is only to reflect the relatedness between the languages in a cognate set.

The tree in Figure 3.1 contains the optimal reconstruction at each node to fulfil the requirement of Maximum Parsimony – having the least possible changes in the whole tree. The numbers on the branches indicate the number of changes required to convert the proto-form at the parent node to the form at the daughter node. A total of three changes can be observed in the tree: Medial [f] changes into [h] in Central-Eastern Polynesian, initial [t] to [k] in Hawai'ian, and medial [h] to [ʔ][3] in Rarotongan. This nicely illustrates how the Maximum Parsimony principle minimizes homoplasy, i.e. observing the same innovation multiple times in a tree. Each innovation (the sound changes described above) can only be observed at one single branch of the tree, which falls in line with the reconstruction principle of economy discussed in Section 2 – it is more likely that a certain sound change happened once in a family (and the affected proto-language passed it on

[2]called *Cook Island Maōri* in Glottolog

[3]The letter ' used in the figure denotes the glottal stop [ʔ]. All other used letters are identical to their IPA representations.

to its descendants), than for it to happen multiple times within the same family independently.

This rules out other possible reconstructions for the intermediary nodes. Only looking at its daughters, **tafa* and **taha* would both be equally valid reconstructions for Proto-Nuclear Polynesian, since both proto-forms would require one change to one of its daughters. Assuming the proto-form **taha* however would assume that the sound change [h] to [f] happened twice in the tree – once for Tongan and once for Samoan. Since that option would also increase the total number of innovations in the tree to four, this reconstruction is ruled out by the Maximum Parsimony principle. This demonstrates how valuable it is that Sankoff's algorithm manages to find the best reconstructions globally, and not only locally for each node, which would imply only minimizing changes from the node to its direct descendants.

In this case, Maximum Parsimony is actually sufficient to find the correct proto-form (Greenhill and Clark, 2011). Of course, that is not always the case, since there is much more to linguistic reconstruction than merely minimizing the number of innovations. However, it is able to provide good approximations in many cases, and given the enormous amount of lexical data that it is applied on, it is expected to roughly quantify which sound changes tend to occur more often than others. Naturally, sounds that often correspond to each other in the synchronous data will be frequently observed changing into each another. Maximum Parsimony reconstructions however can also give some information about the directionality of those changes, as can be seen in the minimal example above: Debuccalization is a commonly observed sound change, and therefore it is more likely for a [f] to change into a [h] than vice versa. The sound change observed in the example follows that assumption and provides quantitative evidence for this asymmetric directionality.

Maximum Parsimony reconstructions were performed for all cognate sets in the data, and all observed transitions were stored in a transition matrix. This matrix is of shape $n \times n$, where n is the size of the alphabet Σ – the set of all different sounds observed in the training data, including the gap symbol. More formally, the transition matrix $M \in \mathbb{N}_0^{n \times n}$, where $n = |\Sigma|$, contains the information about how often i transitioned into j for every possible $i, j \in \Sigma$. M was filled by inspecting each branch of the tree and iterating over the columns of the sequence alignment, counting each transition from parent to daughter language. For example, for the branch between Nuclear Polynesian **tafa* and Central-Eastern Polynesian **taha*, the four transitions $t \rightarrow t$, $a \rightarrow a$, $f \rightarrow h$, and $a \rightarrow a$ would be counted and added to M.

Table 3.2 exemplifies the shape of such a transition matrix based on the example that was discussed earlier. The rows indicate the source sounds, and likewise the columns correspond to the target sounds. Each row therefore contains the information about how often a certain sound developed *into* other sounds – looking at the second row for example, it is visible that [f] changed into a [h] once and stayed [f] three times. Likewise, the third column indicates that [h] originated

	a	f	h	k	t	ʔ
a	14	0	0	0	0	0
f	0	3	1	0	0	0
h	0	0	2	0	0	1
k	0	0	0	0	0	0
t	0	0	0	1	6	0
ʔ	0	0	0	0	0	0

Table 3.2: Transition matrix for reconstructions in Figure 3.1.

from [f] once and was passed on unchanged twice.

The transition matrix resulting from the dataset contains information about a total of 9,840,192 transitions between 1,184 distinct sounds, including the gap. In an estimation-maximization fashion (Moon, 1996), reconstructions were re-generated in three iterations. Those reconstructions do not measure the branch lengths of the tree by plain, but by weighted edit distance. The substitution cost (c) for a given sound change was inferred from the transition matrix of the previous iteration (M) by substracting the probability of that sound change from 1:

$$c(i, j) = 1 - \frac{m_{ij}}{\sum M}$$

Penalizing uncommon sound changes more than common ones is expected to refine the quality of Maximum Parsimony reconstructions. Finally, both transition matrices – the one obtained after the estimation-maximization iterations and the one before – are used to generate training data for different models, which are evaluated against each other in Section 4.

3.2 Training the model

The transition matrix described in the previous section contains information about the frequency of sound changes within the finite set of sounds observed in the data. Despite the matrix spanning over more than a thousand sounds, there is still a plethora of possible sound changes about which there is either insufficient information that can be deduced from the data, or no information at all. Therefore, a model that has the goal of robustly estimating some kind of typological plausibility value for every possible sound transition must be able to generalize, it must be able to predict probabilities for unknown sound changes based on information about known changes. I will first discuss how representing sounds in terms of phonological features enables this kind of generalization, then I will explain how training data was generated based on a transition matrix, and lastly I will outline the model architecture and the parameters used for training.

3.2.1 Phonological feature representations

Expressing structural patterns in sound by the means of phonological features is almost universal among linguists (Mortensen et al., 2016), and as seen in Section 1.1.1.3, historical linguists regularly make use of them when expressing soundlaws. Feature vectors are therefore an obvious choice to represent sounds as numerical vectors that can serve as input for a neural model, which is then able to process any possible combination of features in a continuous space and predict an outcome value. Learning tendencies regarding sound change likelihood based on phonological features, rather than on discrete symbols, thus allows the model to generalize observed patterns to unseen sounds.

The feature representation used in this thesis is an extended version of PanPhon (Mortensen et al., 2016), a database that is able to represent more than 5,000 simple and complex IPA segments in terms of 21 articulatory features. Each feature can be either present (+), absent (-), or not applicable (0). Strictly speaking, that contradicts the idea of binary feature encoding (where a feature can either be present or absent), but especially for using feature vectors for computation, it is worthwhile to distinguish whether a feature is absent because it does not apply in a given context, or whether it would actually be possible for that feature to be present in this context. Explicitly marking when a feature does not apply (like consonantal features in vowels) draws a stronger line between certain sound classes and therefore emphasizes the significance of an applicable, but absent feature.

In order to account for polyphthongs and complex tones as well, I expanded the PanPhon feature inventory by 13 additional features, tailored towards these two sound classes that can not be encoded in PanPhon. Since the base PanPhon features are well documented in the original paper, I will only discuss the features that were added for encoding polyphthongs and complex tones.

The six features that were added to describe polyphthongs are listed below. Since they were developed for diphthongs initially and their interpretation for triphthongs is less intuitive, the short descriptions provided below will only consider diphthongs; I will later discuss how these features are used to encode triphthongs. Every polyphthong is based on its first vowel feature-wise, so it defines all monophthongic vowel features for the whole polyphthong.

backshift. Is the second vowel further back than the first vowel?

frontshift. Is the second vowel more fronted than the first vowel?

opening. Is the second vowel more open than the first vowel?

closing. Is the second vowel more closed than the first vowel?

centering. Is the second vowel a central or centered vowel?

longdistance. Is there a long vertical trajectory of the diphthong?

secondrounded. Is the second vowel rounded?

The first four features, as well as the last one, are straightforward and do not require any further explanation. **Centering** is applied when the second vowel is

either a "true" central vowel (like [ə]) or the centralized version of a peripheral vowel (like [ɪ] or [ʊ]). The feature is therefore used for distinctions between diphthongs like [o͡ə] and [o͡ɛ] in the first case, and likewise for distinguishing for example [a͡ɪ] from [a͡i] in the second case. More precisely, a diphthong is [+closing] if one of the following conditions is true:

- The second vowel is the central vowel [ə].
- The second vowel is an open-mid central vowel.
- The second vowel is a close-mid central vowel.
- The second vowel is a near-open vowel and the first vowel is not a near-open vowel.
- The second vowel is a near-close vowel and the first vowel is not a near-close vowel.
- The second vowel is an open-mid vowel and the first vowel is a (near-)open vowel.
- The second vowel is an close-mid vowel and the first vowel is a (near-)close vowel.

The second feature that calls for further explanation is **longdistance**, which is applied when there is a long vertical trajectory. Its main purpose is to distinguish between diphthongs like [a͡i] and diphthongs like [a͡e]. A diphthong is [+longdistance] if one of the following conditions is true:

- The first vowel is an open-mid or near-open vowel and the second vowel is a closed vowel.
- The first vowel is a close-mid or near-close vowel and the second vowel is an open vowel.
- The first vowel is an open vowel and the second vowel is a (near-)close vowel.
- The first vowel is a closed vowel and the second vowel is a (near-)open vowel.

Triphthongs are basically treated as the combination of two diphthongs, and are therefore encoded by the union of positive features that characterize these diphthongs. For example, the triphthong [ɑ͡i͡ə] is treated as the combination of [ɑ͡i] and [i͡ə] – the latter diphthong makes the triphthong [+backshift,+opening,+centering], the former one adds [+frontshift,+closing,+longdistance] to the triphthong's feature representation. Combined with the base vowel features of [ɑ], that feature vector implies a triphthong that started at the open back vowel and followed a trajectory that would first close and front the vowel, and then retracting it again to a more central position.

The only exception to that logic is how to encode rounding within a triphthong. Since [a͡ʊa] seems phonetically closer to [ʊ͡a] than to [a͡ʊ], we would like to encode the rounded element in the fashion of the former diphthong rather than the latter one. Therefore, if the second vowel of a triphthong is rounded, the triphthong will be [+rounded] – [+secondrounded] is only applied when the last vowel is rounded as well.

The second shortcoming of the PanPhon feature inventory is the lack of possibilities to express complex tones which include different pitch contours. Three features have been added in order to encode complex tones, displayed as the concatenation of up to three canonical IPA tones:

rising. Does the tone start with a rising pitch?

falling. Does the tone start with a falling pitch?

contour. Does the tone include a contour (i.e. falling after rising pitch or vice versa)?

Similar to the handling of polyphthongs, the first tone of a complex tone serves as a base and defines its other phonological features. **Rising** and **falling** are then assigned according to whether the second tonal segment is higher or lower than the first one. For tones that are represented by three tonal pitch segments, **contour** applies when tone first rises, then falls again, or vice versa.

Using this expanded PanPhon feature representation, each sound can be represented by means of 34 phonological features. In order to be processed by a neural network, feature vectors have to be numericalized, which is done by assigning the values of 1, -1, and 0 for present, absent, and non-applicable features respectively. A sound change between two arbitrary sounds is represented by the concatenation of their individual feature vectors, with the target sound vector following the vector of the source sound. The model therefore expects input vectors with a length of 68.

3.2.2 From transition counts to training data

The simplest and most intuitive approach would be to train a model that is able to predict the raw frequency of a given sound change, such that it would learn from the cell values of the transition matrix and predict a frequency for a prompted sound change. This approach however does not work due to the fact that such a generalization could only be based on the similarity of certain sounds, which can be inferred from their feature representation; the typological markedness of a certain sound in contrast can not be deduced from its feature and would thus be disregarded by such an approach. The markedness of a sound however plays a large role in how frequent certain sound changes are – [a] to [ɑ] will be observed much more often than [a̰] to [ɑ̰]. It would be unreasonable to assume that the second pair is phonetically much more dissimilar than the first one; the reason that the first sound change appears more frequently is simply that plain vowels are more common than the typologically marked creaky vowels.

For that reason, aiming at *absolute* predictions about the frequency of a sound change is not very feasible – it makes more sense for the model to learn how likely a certain sound change is *in relation* to other candidates. In the end, the designed use case of the model is to obtain some sort of probability distribution over a set of candidate sound changes – usually with the goal of finding out where a given sound comes from, or what a given sound can turn into. That means that for the usual application, only one sound is considered as source or target, and only the values within the corresponding row or column from the transition matrix are relevant for estimating relevant sound change likelihoods. Viewing the transition matrix as an unnormalized joint probability distribution, the model should learn to generalize from its conditional distributions, rather than from the entire distribution.

Algorithm 1 Generating a random transition matrix for negative training data

$N \leftarrow$ zero matrix of shape M

for row in M **do**
  for i in $sum(row)$ **do**
    $r \leftarrow random$ from $\{0, 1, ..., row.length - 1\}$
    $N_{row,r} \leftarrow N_{row,r} + 1$
  end for
end for

for $column$ in M **do**
  for i in $sum(column)$ **do**
    $r \leftarrow random$ from $\{0, 1, ..., column.length - 1\}$
    $N_{r,column} \leftarrow N_{r,column} + 1$
  end for
end for

Still the question remains how the model can learn this relative information from the absolute counts. This thesis poses as solution a binary classifier that weighs the observed evidence against statistical noise, or more intuitively speaking, it quantifies how the observed frequency of a given sound changes relates to its expected value, assuming that sound changes followed a completely random distribution. Conceptually, this is an application of noise-contrastive estimation (Smith and Eisner, 2005; Gutmann and Hyvärinen, 2010), a technique where observed positive evidence is normalized by artificially generated negative data.

Algorithm 1 illustrates how a second transition matrix N is generated by iterating over the rows and columns of the true transition matrix M and generating random conditional distributions. By iterating over the rows and columns of M and generating corresponding conditional distributions, N implicitly contains information about the frequency (and thus markedness) of a sound. N therefore is an artificially generated transition matrix of a form that would be expected under the assumption that sound change happened completely randomly, with respect to the general frequency of the individual sounds. Even if there were no structural or phonetic processes involved in sound change, we would still expect to see [a] → [ɑ] more often than [a̯] → [ɑ̯] just based on the raw frequencies of the individual sounds!

Since the process that generates N iterates twice over M, a negative example is generated for each observed transition, leading to the fact that $\sum N = 2 \sum M$, meaning that N contains exactly twice as many data points than M. Since the number of negative and positive data points should be the same, M is scaled up by the factor 2.

Containing the same amount of data points, M and N are used to generate training data for a binary classification task. For each possible sound change i, j (indicating that i changes to j), m_{ij} positive data points (with $y_{ij} = 1$) and n_{ij} negative data

points (with $y_{ij} = 0$) are added to the training data. The input feature vector X_{ij} is generated based on the phonological features of i and j as described in Section 3.2.1.

Sampling training data according to the transition counts from M and N like this makes the model learn on contradicting data – for virtually every possible sound change (and its corresponding feature vector serving as the model input), there will be multiple data points, some of which claim to belong to the positive class, others labeled as negative. This forces the model to compromise between the contradicting data points in order to minimize its loss – the ideal (loss-minimizing) prediction for a certain input is not 0 or 1 (as in usual applications of binary classifiers), but a value between 0 and 1, the ratio of positive to negative examples. Given the loss function used to train this model is binary cross-entropy, which is defined as

$$\mathcal{L}(\hat{y_i}) = -(y_i log(\hat{y_i}) + (1 - y_i) log(1 - \hat{y_i}))$$

for a single data point. We can generalize that function to calculating the aggregated loss for a given sound change i, j, given the training data contains m_{ij} positive and n_{ij} negative examples for the corresponding feature vector:

$$\mathcal{L}(\hat{Y_{ij}}) = -(m_{ij} log(\hat{y}_{ij}) + n_{ij} log(1 - \hat{y}_{ij}))$$

It can analytically be shown that the loss function has a minimum at $\frac{m_{ij}}{m_{ij}+n_{ij}}$, which is the ratio of positive examples in the training data for i, j. The derivation of the loss function is:

$$\mathcal{L}'(\hat{Y_{ij}}) = \frac{n_{ij}}{1 - \hat{y}_{ij}} - \frac{m_{ij}}{\hat{y}_{ij}}$$

$\mathcal{L}$ has a minimum where $\mathcal{L}' = 0$, which is at $\hat{y}_{ij} = \frac{m_{ij}}{m_{ij}+n_{ij}}$, as shown below.

$$\begin{aligned}
\frac{n_{ij}}{1 - \hat{y}_{ij}} - \frac{m_{ij}}{\hat{y}_{ij}} &= 0 && \mid + \frac{m_{ij}}{\hat{y}_{ij}} \\
\frac{n_{ij}}{1 - \hat{y}_{ij}} &= \frac{m_{ij}}{\hat{y}_{ij}} && \mid * \hat{y}_{ij} \\
\frac{n_{ij}\hat{y}_{ij}}{1 - \hat{y}_{ij}} &= m_{ij} && \mid * (1 - \hat{y}_{ij}) \\
n_{ij}\hat{y}_{ij} &= m_{ij}(1 - \hat{y}_{ij}) \\
n_{ij}\hat{y}_{ij} &= m_{ij} - m_{ij}\hat{y}_{ij} && \mid + m_{ij}\hat{y}_{ij} \\
n_{ij}\hat{y}_{ij} + m_{ij}\hat{y}_{ij} &= m_{ij} \\
(n_{ij} + m_{ij})\hat{y}_{ij} &= m_{ij} && \mid : (n_{ij} + m_{ij}) \\
\hat{y}_{ij} &= \frac{m_{ij}}{m_{ij} + n_{ij}}
\end{aligned}$$

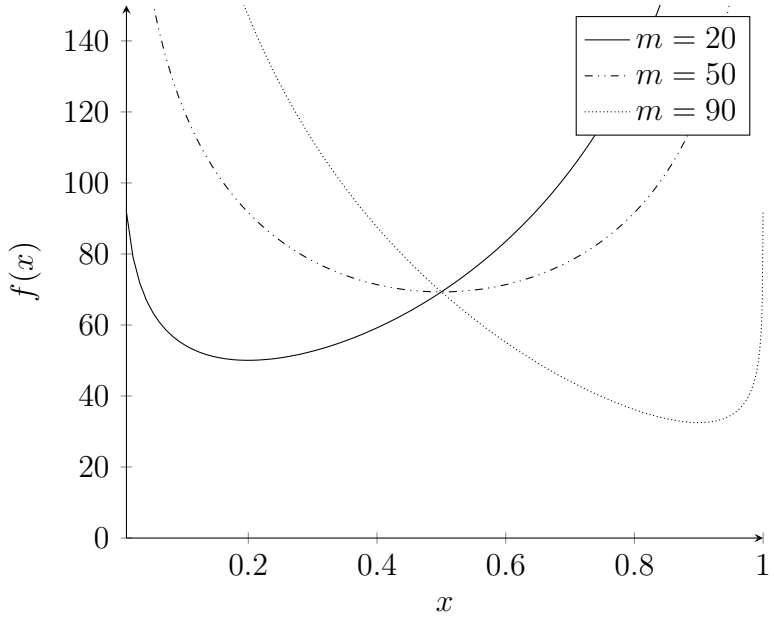

Figure 3.2: Total loss for different values of m (positive examples) with $(m + n) = 100$

Figure 3.2 exemplifies how the loss functions behaves for different amounts of positive data points, given a constant number of total training points. The previously discussed minima when $\hat{y}_{ij}$ corresponds to the positive rate in the data points is clearly visible. The more this loss-minimizing value approaches the vertical asymptotes at $\hat{y}_{ij} = 0$ and $\hat{y}_{ij} = 1$, the steeper the slope of the function becomes, such that wrong predictions for datapoints with strong evidence for being very likely or unlikely get penalized more than training examples where the observed sound change frequency is similar to the randomly generated one.

This adaptation of noise-contrastive learning produces a model which is able to predict pseudo-probabilities for certain sound changes, which can be used directly to inform a reconstruction module. Alternatively, the model can be used to query a probability distribution for a set of sound changes, which usually would be a conditional distribution based on one source or target sound. To obtain such a distribution, one can easily use add a softmax layer on top of the models' logits, as is standard practice for many classification tasks with multiple classes.

3.2.3 Model architecture

A simple feed-forward artificial neural network[4] with 3 hidden layers and a hidden size of 128 is applied to learn pseudo-probabilities as described in the previous section. The activation function employed to the hidden layers is GELU (Gaussian Error Linear Units; Hendrycks and Gimpel 2016) which slightly outperformed other activation functions; employing dropout layers did not seem to have any beneficial effects to the model's performance. The model is trained according to

[4]Artificial neural networks are considered to be a standard technique in the domain of machine learning and therefore not explained in detail. The interested reader is referred to introductory books like Haykin (1999) or Hertz et al. (2018).

standard practices for binary classifiers, applying a sigmoid activation function to the output layer and using binary cross-entropy as the loss function, the effects of which in this particular set-up have been thoroughly discussed in Section 3.2.2. The model was trained over two epochs, which was sufficient due to the large amount of training data points within one single epoch, using the Adam optimizer (Kingma and Ba, 2014). 90% of the generated data was used as training data, the remaining 10% was used for testing the model's performance. The entire development of the model was done in Python, using the TensorFlow package (Abadi et al., 2015) for conveniently wrapping all parameters and architecture decisions discussed above.

3.3 Post-Processing

3.3.1 Integration to EtInEn

The weights and the biases of the model that was trained in Python were exported to a text file in order to make them readable for other frameworks. Since EtInEn is written completely in Java, I had to manually implement classes for a feed-forward neural network and its individual layers, mirroring the previously discussed architecture of the model trained in Python. Since such simple networks are nothing else than mere matrix multiplications with an activation function on top, implementing an already trained model was straightforward.

Within EtInEn, the model's primary use is to serve as a typological prior for phonological reconstructions, that is quantifying how likely certain candidate sound changes are in general before applying language-specific layers like sound laws or phonotactics. The model is used in a reconstruction module that generates language-agnostic reconstruction candidates, which then are evaluated against the inferred sound laws from the proto-language to its respective daughter language. While for example there might be several reconstruction candidates that are generally justifiable (without applying specific knowledge about the languages in question), some candidates might violate an established sound law and therefore be considered unlikely. Inversely, good candidates that also follow the sound laws would be considered even more probable.

k	a	h	a	(Hawai'ian)
t	a	h	a	(Maōri)
t	a	ʔ	a	(Rarotongan)

Table 3.3: Example of a Multiple Sequence Alignment.

As is standard practice in automated reconstruction algorithms, the EtInEn reconstruction module takes the forms of the immediate descendants of the node that shall be reconstructed as input, and requires those forms to be aligned. Table 3.3 illustrates such a multiple sequence alignment for forms that have been previously introduced in Figure 3.1. The node in question, for which a proto-form should be

reconstructed in this case, is Central-Eastern Polynesian; therefore the input forms come from the immediate descendants[5] Hawai'ian, Maōri, and Rarotongan.

Given the multiple sequence alignment, the algorithm iterates over each column. The first input column therefore provides the sounds *(k, t, t)*, which we denote as a tuple of target sounds $T = \langle$[k], [t], [t]$\rangle$. Based on these input or target sounds, a probability distribution over a set of candidate source sounds S can be inferred by means of the following formula:

$$P(s) = \frac{\prod_{t \in T} p(s,t)}{\sum_{s' \in S} \prod_{t \in T} p(s',t)} \forall s \in S$$

S can be every arbitrarily chosen set of sounds. The probability of a particular sound change between a source sound s and a target sound t is denoted as $p(s,t)$ and is predicted by the neural model. For normalizing those frequency estimations into distributions, I implemented methods for obtaining conditional distributions over either source or target sounds; in these cases, $p(s,t)$ would actually be $p(s|t)$ or $p(t|s)$ respectively. Since the probability mass has to be normalized again (as seen in the function), it is also possible to work with the raw probability estimations the model was trained on, which is the final logit (for a single sound change) normalized by the sigmoid function. Those three possibilities to query the model and to normalize probabilities naturally lead to different posterior probability distributions. In my evaluation, I use $p(s|t)$ in order to obtain a probability distribution over possible source sound for each observed sound in the column. While this is the most intuitive approach for determining the most likely proto-sound for a set of observed sounds, there exist other possibilities to obtain probabilities from the model, which unfortunately could not be explored due to the computational expenses connected to the reconstruction technique.

Either way, the algorithm calculates a probability distribution over the candidate sounds S by obtaining probabilities for each sound change $\langle s,t \rangle \in S \times T$ and normalizing them as stated in the formula above. After calculating a probability distribution for every column of the alignment, the best reconstruction consists of the most likely proto-sound for each column.

The full-fledged EtInEn workflow for reconstructing proto-forms will finally include other pieces of information, like inferred sound laws (Daneyko, 2020), potential reconstructions of parent nodes, or belief values of reconstruction candidates in the descendants when moving higher up the tree. Since my workflow only produces naïve, language-agnostic reconstructions that disregard this information, the details of how EtInEn includes these factors are not discussed here.

3.3.2 Bias tuning for boundary cases

Before applying the evaluation, the model's performance was tested by manually analyzing automatically generated reconstructions that were only informed

[5]The three languages in question are immediate descendants according to the reduced tree for the exemplary language sample introduced in Section 3.1.2; not in a strict linguistic sense.

by the probabilities predicted by the model. Additional information that would be added later within the EtInEn workflow was not regarded, as discussed in the previous section. Manually inspecting some test cases has the advantage of gaining a first impression about how well the model was able to learn certain transition probabilities, and check the model's tendencies against the human linguistic intuition. This intermediary step between the training of the model and its evaluation also made it possible to manually add some biases for boundary cases, in which the model did not behave exactly as intended. I used the NorthEuraLex database (Dellert et al., 2020) for my test cases, automatically reconstructing Proto-Germanic, Proto-Balto-Slavic, Proto-Uralic, and Proto-Turkic. Two boundary cases that required some manual adjustment became apparent: Transitions between equivalent symbols and transitions including gaps. Both problems were addressed by manually adding a defined bias value to the logit of the respective sound change.

The first boundary case simply involves the predicted probability for a sound staying the same, i.e. *not* changing at all. Two issues that were related to these values were sporadically observed: Firstly, the model would sometimes suggest sound changes without any need, despite having the same sound in the entire column of the respective multiple sequence alignment, violating the principle of economy. This occasionally happened for sounds that are generally considered to be unstable, like nasal consonants or [h], which on the side leads to the pleasant conclusion that the model was able to learn that those sounds tend to change rather frequently. However, it should obviously not assume that it is less likely for a sound itself to derive from another sound, rather than from itself. The second related case occurred when two proto-sound candidates shared the same feature encoding, which sometimes happens between closely related sounds like [r] and [ɾ], despite the use of a rich phonological feature representation. Naturally, in cases like this where two sounds (and therefore the sound changes involving one of those two sounds) have the same feature representation, the model has no chance to tell them apart and will predict the same value. The principle of economy applies here again in quite an intuitive way – it is much more likely that there was no change of the articulatory and phonetic properties of the sound, than assuming that some sort of change happened, without having solid arguments for proposing such a change. In order to address these two problems, I added a bias of $+\mathbf{2}$ to each logit where the string representation of the input symbol is equivalent to the output symbol string.

The second boundary cases concerns sound changes including the gap symbol, which has so far been treated as if it were a regular segment. This does obviously not reflect the nature of the gap symbol, since it does not represent a physicial sound, but is only generated to indicate that some part of a sequence does not correspond to any part of another sequence. More concretely speaking, correspondences between sounds and gaps in sequence alignments used for reconstructions (assuming that those alignments are correct) imply that either an insertion or a deletion must have happened.

The special role of the gap symbol has been addressed already in the infancy of

automated sequence comparison for historical linguistics (Covington, 1996; Kondrak, 2000) and has been continuously receiving attention ever since (List, 2014). For most proposed automated methods towards sequence comparison or even ancestral state reconstruction, gaps are bound to cause problems if they are not specifically addressed. In my test cases, the model would frequently reconstruct gaps in cases of alignment columns that include both gaps and actual sounds; considering epentheses more likely than elisions. However, in most circumstances the opposite should be the case, since sounds generally seem to be more likely to be lost than to appear "out of the blue" (Kümmel, 2007; Campbell, 2013). The Maximum Parsimony reconstructions that were used to generate the training data are naturally not aware of such tendencies, and furthermore the training data contained much noise from inflectional affixes (e.g. the German infinitive ending *-en* which has been lost in English), leading to many gap alignments that are results of morphological rather than phonological processes. Moreover, since no information about phonotactics or language models were included in the reconstruction algorithm, no case-by-case distinction could be made whether the resulting proto-form after removing a vowel would still consist of valid syllables or commonly found n-grams. These factors lead to many reconstructions where syllable nuclei were removed, partially producing proto-forms that did not contain any vowels anymore. Albeit not as apparent as in vowels, the same tendency to overgenerate epentheses could also be observed for consonants. In order to avoid the excessive reconstruction of gaps, I drastically increased the threshold of reconstructing gaps from vowels by adding a bias of **-4** to the logits of respective sound changes. For consonants, a slight bias towards favoring deletions over insertions was applied by adding **-1** to the logits of sound changes from gap to consonant.

Resulting from these biases were improved automated reconstructions like Proto-Uralic ***voko* and ***sopok*, which had been previously reconstructed as ***vk* and ***spk*. The reconstructions in the development set also exposed that the automatic inference of proto-sound inventories by SoInEn (Soundlaw Inference Enginge; Daneyko 2020) did not work as intended in this setting. It produced way too small proto-inventories, which of course was responsible for some rather poor reconstructions. For the evaluation therefore, a global alphabet of possible proto-sounds had to be defined beforehand. Given the at times questionable proto-sound inventories in the development set, applying the biases lead to significantly better reconstructions, where the model was usually able to choose a reasonable proto-sound.

4 Evaluation

In order to evaluate the model's performance, two different reconstruction algorithms are employed that make use of probabilities obtained by the model. The first of these techniques is to recursively generate naïve and language-agnostic reconstructions up the tree. To reconstruct a given proto-language, every intermediary node in the language tree between the leaves (the modern languages) and the proto-language is reconstructed, traversing the tree in a bottom-up fashion. Therefore, proto-forms of nodes that are located lower in the tree, and therefore temporally closer to the modern languages, are reconstructed first, and then serve as reconstruction inputs for nodes higher up in the tree, i.e. for older proto-languages. As described in Section 3.3.1, the reconstruction algorithm employed here only works with the direct descendants of the respective node as input, disregarding information from forms that are located at other points in the tree. For each node, all proto-forms that have reflexes in at least one daughter language are reconstructed before moving on to reconstructing the next node in the tree.

As described in Section 3.3.1, the probability distributions for this bottom-up reconstruction are inferred by applying a softmax layer over the logits obtained for a given target sound. This yields conditional probability distributions that contain information about the most probable sources of a certain target sound. Recursive bottom-up reconstructions are computationally quite expensive, and it was therefore not possible to explore other distributions within the timeframe of this thesis.

Besides this recursive bottom-up reconstruction technique, another set of proto-forms is reconstructed by employing weighted Maximum Parsimony. The principle of Maximum Parsimony has already been discussed in detail in Section 3.2.2 where it was used to generate training data. In the weighted variant, the neural model is used to dynamically calculate substitution costs between sounds. The branch length between a daughter and a parent node therefore represents the phonetic similarity between the two forms, rather than just the number of required edit operations.

For the weighted Maximum Parsimony reconstructions, substitution costs between all possible symbols need to be inferred from the probability distribution obtained by the model. However, simply substracting the inferred probability from 1 does not yield useful transition costs, because the vast majority of the resulting probabilities is very close to 0, especially for larger alphabets. In order to counteract this phenomenon and to obtain a signal that is strong enough to actually impact the algorithm, two measures are taken to convert the information obtained from the model into useful transition costs.

As introduced in Section 3.3.1, I use S to describe the defined set of possible source sounds, whereas T denotes the set of target sounds, i.e. the sounds that are present in a given column of the multiple sequence alignment. For employing weighted Maximum Parsimony reconstructions, a global alphabet of possible source sounds Σ is defined *a priori*. For each column in the alignment, S is defined as $\Sigma \cup T$; the union of the global alphabet and the sounds present in the column.

To obtain sensible substitution costs, a stochastic transition matrix over the symbols $S \times S$ is generated, containing probability distributions of sound changes that are conditioned by the source sound. Essentially, conditioned probability distributions $P(s'|s) \forall s', s \in S \times S$ are inferred by applying a softmax layer on the model's output logits. The first measure that is taken to emphasize differences in probabilities is to exclude pairs of identical sounds from the output distribution before employing the softmax layer. This measure is taken because the vast majority of the probability mass is otherwise assigned to the sound staying the same. With reasonably sized alphabets (with at least 20 symbols), this identity transition usually amounts over 99% of the probability mass, while all other transitions share the last percent among themselves. Due to the resulting small probabilities, the model's prediction about how likely a given sound change is *in relation* to others is therefore essentially lost. By removing the largest attractor of probability mass, much higher probability values are assigned to transitions that involve actual sound change. The cost for a sound not changing is set to 0 globally.

This measure already emphasizes the relative differences between the predicted transition probabilities. However, most of the resulting probabilities are still too close to 0 to contribute a meaningful signal to the algorithm. The resulting probability distribution resembles a power distribution: The bulk of the probability mass is shared among the few most probable target sound, while there is a long tail of unlikely target sounds with probabilities near 0. This distribution shape is addressed by employing a logarithmic transformation: The closer probabilities are to 0, the stronger they are affected by the transformation. The cost c for a given sound change $s \rightarrow t$ is calculated by applying such a log-transformation[1] and substracting the log-normalized probability from 1:

$$c(s,t) = \begin{cases} 1 - (-\frac{1}{log_2(0.5p(t|s))}) = 1 + \frac{1}{log_2(0.5p(t|s))}, & \text{if } s \neq t \\ 0, & \text{otherwise} \end{cases}$$

[1]Rudimentary experiments on employing such a log-transformation for the bottom-up reconstructions showed no beneficial effect.

These scaling measures make it possible to extract costs with substantial and meaningful differences from the probabilities provided by the model. The obtained cost enable Sankoff's algorithm to use the model's power to quantify how likely certain sound changes are in relation to each other. However, adding this information to enhance existing reconstruction algorithms is only one of many possible use cases for the model – the same information can be valuable for other tasks, such as the detection of sound correspondences or the induction of sound laws. I would like to emphasize again that providing better automated reconstructions is not the core contribution of this thesis, but merely an application to showcase the model's potential for improving computational models of language change.

4.1 Evaluation dataset

Up to this day, computational historical linguists still struggle with a shortage of expert reconstructions that are available digitally. That severely limits the number of datasets that can provide a gold standard for automated reconstructions, essentially reducing it to a handful of datasets that contain information about proto-forms across all concepts covered (List et al., 2022b).

For my evaluation, I use the Austronesian basic vocabulary database (ABVD; Greenhill et al. 2008), which contains 325,947 lexical items from 1,692 languages spoken throughout the Pacific region. Based on ABVD, Proto-Austronesian and Proto-Oceanic are reconstructed and evaluated against respective gold standard reconstructions contained in ABVD. Proto-Austronesian expert reconstructions are provided by Blust (1999); for Proto-Oceanic, ABVD actually contains two sets of gold standard reconstructions, one by Blust (1993) and another one by Andrew Pawley.[2] Evaluating the automatically produced reconstructions against these lists of expert reconstruction mirrors the evaluation employed by Bouchard-Côté et al. (2013) and makes my results directly comparable to theirs. In order to keep this symmetry, ABVD-based evaluations are limited to the set of reconstructions discussed by Bouchard-Côté et al. (2013), which are listed in their appendix. Furthermore, the same global alphabet of proto-sound candidates was used, which was defined as the union of the sound inventories of their subset of ABVD.

4.2 Evaluation metrics

The automatically produced reconstructions are evaluated against their corresponding gold standards by the means of Normalized Edit Distance and B-Cubed F-Scores, both indicating how closely a produced reconstruction resembles the actual proto-form.

Normalized Edit Distance is obtained when dividing the plain edit distance (Levenshtein, 1966) of a pairwise alignment by the length of the alignment and has

[2]There seems to be no publication explicitly covering the reconstructions for the ABVD concepts; however there is a strong overlap with reconstructions discussed in Ross et al. (2007).

been used frequently to evaluate the quality of automatic reconstructions (Bouchard-Côté et al., 2009; Bouchard-Côté et al., 2013; Ciobanu and Dinu, 2018; Hall and Klein, 2010; Jäger, 2019; Meloni et al., 2019). Normalized Edit Distances quantify by simple means how different two sequences are, making it an obvious choice for measuring to which extend an automatic reconstruction deviates from its gold standard. Both its easy application and its intuitive interpretation have made (normalized) edit distance the most popular metric to evaluate ancestral state reconstruction methods.

Using B-Cubed scores (Amigó et al., 2009) for evaluating the quality of reconstructions, on the other hand, has recently been proposed by List (2019b) in order to measure structural similarity between two sets of reconstructions, rather than pure phonetic similarity. This accounts for the fact that reconstructions are by nature abstract to a certain degree, since the exact phonetic value of a reconstructed sound can not be predicted with full certainty, but some of their articulatory properties can be estimated from the data. This trade-off between reconstructing the structural function of a phoneme and its phonetic representation is known as the abstractionist-realist debate which has been summarized in Section 1.3.

Edit distances come with the drawback of not being able to identify the phonological structure within a set of reconstructions. If, for example, one scholar chooses to reconstruct a certain phoneme as [a], while another scholar reconstructs the same phoneme as [ɑ], the reconstructions of the two scholars would be structurally identical (and both equally valid). Comparing these two reconstruction systems by means of edit distance, however, would count a mismatch for each [a - ɑ] correspondence, counting differences as errors that are arguably too fine-grained to be really considered as such. B-Cubed scores on the other hand measure how well one cluster can predict another cluster – in this case, where [a] and [ɑ] perfectly correspond to each other, one can predict with full certainty that there will be an [ɑ] in the second reconstruction system wherever an [a] is encountered in the first one. In contrast to edit distances, B-Cubed scores therefore don't penalize subphonemic differences, but quantify the structural similarity of two sets of reconstructions.

When comparing two clusters for structural similarity, a *precision* and a *recall* score are obtained. The former quantifies how well the first cluster predicts the second cluster, while the latter one measures the inverse relation. When evaluating an automatically generated set of forms against a gold standard, *precision* usually denotes how well the generated forms predict the reference forms (and likewise *recall* is used to measure how well the generated forms can be predicted from the gold standard forms). To illustrate how this relation is not symmetrical, consider the sequences [A B C D] and [A A B B]: The first sequence can perfectly predict the second sequence, since each of its symbols map to exactly one symbol from the second sequence. However, both symbols of the second sequence map to two different symbols of the first sequence respectively – it is not possible to predict the first sequence from the second one. This minimal example therefore would yield a perfect *precision*, but a rather low *recall* (List, 2019b). The overall mutual predictive strength – and therefore structural similarity – between two clusters

is denoted by the B-Cubed F-Score, which is defined as the harmonic mean of precision and recall.

Due to its initial purpose of evaluating clustering tasks (Amigó et al., 2009), B-Cubed scores have been introduced early to automated language comparison when it comes to the task of automatic cognate clustering (Hauer and Kondrak, 2011; Jäger et al., 2017; List et al., 2017). The main innovation by List (2019b) consists in understanding ancestral state reconstruction as a partitioning or clustering task, emphasizing the abstractist, structural component of reconstructions over the realist, phonetic component. Since then, B-Cubed scores have been used as evaluation metrics for the closely related task of reflex prediction (Celano, 2022; Dekker and Zuidema, 2020; Kirov et al., 2022; List et al., 2022c) with essentially the same intuition, that the predicted forms should be as similar to the gold standard as possible. Apart from List et al. (2022b), however, I am not aware of any further publications that make use of B-Cubed score for evaluating the quality of automatic reconstructions, which can arguably be partially attributed to the aforementioned lack of good gold standards for the vast majority of proto-languages.

4.3 Baseline models

All reconstructions generated by my model are evaluated against the respective gold standards using the metrics described above. To compare those metrics, Maximum Parsimony reconstructions (as described in Section 3.1.2) serve as a simple baseline model. Furthermore, the performance on ABVD is compared to the reconstructed forms reported by Bouchard-Côté et al. (2013).

Since Bouchard-Côté et al. (2013) employ more sophisticated reconstruction algorithms that include some kind of awareness for a specific language and the context of a certain sound correspondence, it can not be expected that my model will outperform theirs. Comparing my reconstructions against their model has the primary goal of understanding to which extent language- and context-agnostic reconstructions can approximate those generated by models that take this information into account. After all, the core of this thesis is a model that can predict markedness for arbitrary sound changes, and not an improved reconstruction algorithm; therefore it is neither my intention nor my expectation to outperform state-of-the-art models.

5 Results and Discussion

Tables 5.1 and 5.2 show the overall performances of the different reconstruction algorithms by means of the two metrics introduced in Section 4.2, the average normalized edit distance and the B-Cubed F-Score. The two algorithms that were informed by the neural sound change model, Naïve Bottom-Up Reconstruction and Weighted Maximum Parsimony, are compared to the model by Bouchard-Côté et al. (2013) and to unweighted Maximum Parsimony Reconstructions that serve as a simple baseline. Over the course of this chapter, I will use the abbreviations (W)MP, BU, and BC to refer to the automated reconstructions produced by (Weighted) Maximum Parsimony, Bottom-Up reconstruction, and the model by Bouchard-Côté respectively.

The first observation is that employing an Estimation-Maximation (EM) technique for generating the training data, as described in Section 3.1.2, does not seem to be substantially beneficial to the model's performance – for the Proto-Oceanic reconstructions, it was even the model that was trained without any EM component that achieved the best results. However, these differences between the different models and the resulting reconstructions are barely noticeable. This can be attributed to two factors in the set-up of the EM module: Firstly, the substitution costs were directly inferred from the probabilities without any scaling or further modifications. In Section 4, I have discussed the necessity of scaling costs away from 1 in order to generate a signal that is strong enough for the algorithm to process – a challenge that was addressed by employing a logarithmic transformation to the probabilities for the WMP reconstructions. Such an adjustment of costs, however, was not employed to the EM module. The resulting cost matrix would therefore have values very close to 0 along the diagonals (wherever a sound does not change) and thus close to 1 everywhere else. A cost matrix of this shape is not able to provide Sankoff's algorithm with a useful signal, and a first exploratory run on ABVD even showed that WMP reconstructions that rely on such a matrix even produce slightly worse results than plain unweighted MP reconstructions.

The second issue with the EM module is that all transitions are counted equally along every branch in the tree. However, that disregards that different branches

Algorithm	Edit Distance	B-Cubed F-Score
Weighted Maximum Parsimony		
after 0 EM iterations	0.33	0.64
after 1 EM iteration	0.31	0.66
after 2 EM iterations	0.33	0.65
after 3 EM iterations	0.31	0.66
Bottom-Up	0.36	0.57
Maximum Parsimony	0.32	0.63
Bouchard-Côté et al. (2013)	0.15	0.8

Table 5.1: Evaluation metrics for Proto-Austronesian reconstructions against the gold standard reconstructions by Blust (1999).

of the phylogenetic tree can have vastly different time depths. The larger the temporal distance between a daughter and a parent node, the more innovations are expected to occur on this branch. While it is not really feasible to measure the actual time depth of a branch in question, it is trivial to quantify its innovativeness – which is the actually relevant piece of information. Observing a given sound change on an otherwise highly conservative branch is a stronger evidence for its commonness than observing it on a branch that is generally innovative. Daneyko (2020) proposes a method to count transitions weighted by the innovativeness of the respective branch, which could easily be implemented in my workflow as well.

Addressing these two problems in future work could significantly improve the EM module by providing it with useful substitution costs with respect to the innovativeness of individual branches. An improved EM module would in turn generate better training data, which would naturally result in better models. For now, I have to leave these ideas up to future work, and report that the EM module did not significantly impact the resulting models for better or worse. Throughout this chapter, I will focus on discussing the reconstructions generated by the model trained after the last EM iteration, since that model produced the best WMP reconstructions for Proto-Austronesian. Using that model furthermore enables me to compare WMP directly to BU, since it was informed by the same model. Since BU reconstructions are computationally much more demanding, it was not feasible for me to compare different models' performances for that technique within the timeframe of this thesis.

Shifting back to the reported metrics, another striking observation can be made: The MP reconstructions, that are intended to serve as a baseline, outperform the BU reconstructions. This might seem surprising at fist sight, however, MP has a conceptual advantage over BU, since it optimizes all the reconstructions *globally*, whereas BU is only able to reconstruct forms *locally*, based on the immediate descendants of the node in question. That bears the risk of propagating false or misleading information up the tree that can not be corrected afterwards by considering information from more distantly related languages and nodes higher

Algorithm	**ED (B)**	**B^3 (B)**	**ED (P)**	**B^3 (P)**
Weighted Maximum Parsimony				
after 0 EM iterations	0.3	0.7	0.28	0.72
after 1 EM iteration	0.31	0.69	0.29	0.72
after 2 EM iterations	0.33	0.67	0.31	0.69
after 3 EM iterations	0.3	0.69	0.28	0.71
Bottom-Up	0.38	0.55	0.37	0.56
Maximum Parsimony	0.31	0.68	0.3	0.7
Bouchard-Côté et al. (2013)	0.24	0.75	0.23	0.75

Table 5.2: Evaluation metrics for Proto-Oceanic reconstructions against the gold standard reconstructions by Blust (B) and Pawley (P). ED = Average normalized Edit Distance, B^3 = B-Cubed F-Score.

up the tree – which implicitly happens when optimizing a whole tree rather than recursively reconstructing individual nodes. The larger and deeper a language tree is, the stronger this effect is expected to happen – a tendency from which BU severely suffers in this case, since Austronesian is the second largest language family in the world (measured by the number of modern languages; Hammarström et al. 2022).

Due to the poor performance of BU, I will focus on WMP reconstructions when discussing some Proto-Austronesian and Proto-Oceanic reconstructions in detail. After all, employing techniques for ancestral state reconstruction merely serves the purpose of getting an impression of the neural model's predictive power, as measured by its ability to contribute to good reconstructions.

5.1 Proto-Austronesian

Since the metrics for Proto-Austronesian reconstructions generated by MP and WMP only differ slightly, it does not come as a surprise that most of these reconstructions are identical. Nevertheless, there are some interesting differences, part of which can be attributed to the refined substitution costs which were deduced from the neural model. With an average normalized edit distance of 0.32 and a B-Cubed F-Score of 0.63, however, MP obtained surprisingly good results for a conceptually simple baseline. That strengthens the conclusions drawn by Jäger and List (2018) that ABVD is a relatively simple dataset for ancestral state reconstruction.

For Proto-Austronesian, MP struggled to reconstruct word-initial and word-final */l/ and reconstructed **/r/ instead. Informed by the fact that /l/ frequently changes into /r/, however, WMP was able to reconstruct */l/ correctly in these cases. This can be observed in forms like */likud/ 'back' (MP: **/rikus/; WMP: **/likus/), /liqeʁ/ 'neck' (MP: **/ril/; WMP: /lil/), or */qebel/ 'smoke' (MP: **/ʔəbər/; WMP: **/ʔəbəl/).

In other cases, the information from the sound change model could be used to interpolate between different vowel qualities and reconstruct a better proto-sound. This was the case for */beli/ 'to buy', which was correctly reconstructed by WMP, but predicted as **/bali/ by MP. The extant forms frequently feature different (close-)mid vowels like /e, ə, o/. The neural model is able to predict reasonably high transition probabilities between those vowels based on their common features. The unweighted MP algorithm on the other hand has no access to this information and renders all of these sounds as different, completely unrelated symbols. Due to that, **/a/ is reconstructed as the proto-sound that generates a word tree with the least possible changes, whereas WMP successfully reconstructs */e/ based on the information of the vowel similarities. Another case where the sound change model lead to a better reconstruction of vowels was */ma-buʀaq/ 'rotten', which was reconstructed */buruk/ by WMP and */baruk/ by MP.

The last example already exhibits the two major shortcomings of both MP and WMP. Firstly, all reconstruction algorithms struggled with reconstructing */q/, probably due to its typological markedness, and instability throughout the Austronesian languages. Even Bouchard-Côté et al. (2013) report that this proto-sound was their main systematic source of errors, although their model still manages to reconstruct */q/ significantly better than my models. For example, the final segment of */biʀaq/ 'leaf' could not be reconstructed correctly by either model: BC reconstructs **/bela/, BU predicts **/biaɦi/, whereas both MP and WMP produce **/bia/. Especially MP and WMP consistently struggled to reconstruct */q/, reconstructing either a gap or */k/ instead.

The second issue that can be noticed is that the prefix **ma-* wreaked havoc for both models, since it has been lost in many of the extant languages. The models' inability to reconstruct affixes in such cases can also be seen in forms like */ma-lawas/ 'wide' or */i-kamu/ 'you', which were reconstructed as **/lawa/, **/kamu/ (WMP), and **/l̥awa/, **/kamu/ (MP) respectively.

In general, both MP and especially WMP tend to overpredict gaps in the proto-form, which constantly leads to shorter proto-forms than desired. The overprediction of gaps is the one major weakness that WMP has compared to MP, and it comes to no surprise: In Section 3.3.2, I discussed how the neural model tends to overestimate the probability of insertions, and how I addressed this issue by modifying the corresponding logits. However, this measure was aimed at the Bottom-Up reconstruction process, where conditional probability distributions are inferred with respect to the given target sounds. WMP on the other hand is informed by a stochastic cost matrix, that is based on distributions conditioned by the source sound. Therefore, the softmax layer in this case is only applied to all the logits predicted to insertions – which all are affected by the applied bias. Essentially, the cost matrix therefore loses the intended bias against insertions in favor of deletions.

While this overprediction of gaps in WMP can be observed in many places, it is often not detrimental to the metrics in comparison to MP, since the latter algorithm often reconstructed the wrong sound in such cases. Examples for that

phenomenon are */ŋajan/ 'name' (MP: **/ŋaŋan/; WMP: **/ŋaan/), */qabaʀa/ 'shoulder' (MP: */abafa/; WMP: **/abaa/), or */ma-baqeʀu/ 'new' (MP: **/baru/; WMP: **/bau/). For forms that are predicted too short by WMP, however, it can often be observed that BU generates better reconstructions that usually mirror the syllable structure of the gold proto-form, indicating that the bias against insertions works as intended for its original purpose. Adjusting the costs for insertions and deletions definitely would be one major point that future work on the WMP approach should address.

Due to its bias against reconstructing gaps without strong evidence, BU reconstructions however suffered from the opposite effect, especially in cases where the cognate set contained many partial cognates with additional affixes. Many modern reflexes of Proto-Austronesian */mula/ 'to plant' are formed together with a derivational affix, like Bolaang Mongondow *mo-mula* or Malagasy *mam-bòle/mam-bòly*. The fact that many of the modern forms were only partial cognates which could not be properly aligned made the model reconstruct the unreasonably long proto-form **/mamamamoah/, trying to avoid reconstructing gaps.

A further complication for my reconstructions lies within the shape of the source database. ABVD contains different expert cognacy judgements that can be competing, which is displayed by linking the same word to different cognate sets in case of conflict. This caused issues when importing the database to EtInEn, which expects that each word is only part of one single cognate set. Due to that, only the first assignment to a cognate set for each lexeme was considered, ignoring all others. That could lead to some incomplete cognate sets in cases where some forms had been assigned to another cognate set earlier.

ABVD furthermore does not contain gold tokenizations for the data, which forced me to rely on automated tokenization techniques. While that seemed to work well in most cases, there were some faulty tokenizations where diacritica were considered to be individual segments. That was the case for words like Tasmate /taʔari ʔi/ or Letemboi /rɛgɛⁿdao/ 'small'. Ideally, automated reconstructions should be based on consistent cognacy annotations with good tokenizations and alignments. While none of these factors had a noticeable effect in the end, they have definitely caused some noise in the data.

In order to realistically compare automated reconstructions to Bouchard-Côté et al. (2013), it would be necessary to use the exact same data instead of using the Lexibank version. On the one side, Bouchard-Côté et al. (2013) employed some pre-processing to clean up the data, addressing the aforementioned problems. On the other side, ABVD is still being expanded and edited, so the version that I use contained slightly other data than the one used by Bouchard-Côté et al. (2013).

5.2 Proto-Oceanic

Most of the observations made on Proto-Austronesian reconstructions also apply to the corresponding Proto-Oceanic reconstructions – all models struggled to

reconstruct */q/, WMP severely overpredicted gaps, while BU was especially vulnerable to trailing affixes and noisy alignments. The aforementioned issues with the source data naturally affected the Proto-Oceanic reconstructions in the same way as was the case for Proto-Austronesian. Nonetheless, there are some further interesting cases to discuss that are specific to Proto-Oceanic.

Besides */q/, MP and WMP also consistently failed to reconstruct */ʀ/ and predicted a gap instead. Both issues are reflected in the faulty reconstruction **/ua/ which actually should be */ʀuqa/ 'neck'. Another instance of that issue is */waiʀ/ 'water', which is reconstructed as **/wai/. While this tendency can also be occasionally observed in Proto-Austronesian reconstructions, it is very prominent in Proto-Oceanic.

Generally, word-final consonants tend to cause problems, since they have disappeared in many modern languages. */manuk/ 'bird' is reconstructed as **/manu/ by all algorithms (including BC), and analogically */nanuk/ 'mosquito' lacks the final */k/ in all automatic reconstructions as well.

On a more positive note, there were again some phenomena that were not reconstructed correctly with unweighted MP, but could be handled successfully with substitution costs generated by the neural model. Intervocalic */p/ for example was correctly reconstructed by WMP, but predicted as */f/ by MP. The form */api/ 'fire' was therefore correctly reconstructed by WMP, the same correspondence was successfully rendered in */mapanas/ 'warm' which was reconstructed as **/mapana/. Unweighted MP on the other hand reconstructed **/afi/ and **/mafana/ respectively. That suggests that the neural model was actually capable of predicting that the lenition of /p/ to /f/ is more likely than the inverse change.

Despite its overall tendency to overpredict gaps, WMP was able to retain some segments that were not reconstructed in MP. In particular, that refers to the second vowel in bisyllabic words like */kutu/ 'louse' or */patu/ 'stone' – for both words, the final */u/ is only predicted by WMP. While the former form is reconstructed correctly, the latter one is reconstructed as **/vatu/. In this case, the algorithm fails to capture another instance of spirantization, which it successfully did in the previous case for intervocalic */p/. The MP reconstructions for the two words in question here are identical to the WMP reconstructions apart from the lacking final */u/.

It is notable that the BC reconstructions for Proto-Oceanic are substantially worse than those for Proto-Austronesian, and many of their faulty reconstructions exhibit similar issues as my reconstructions. The challenge to reconstruct */q/ in word-final positions is much more apparent in Proto-Oceanic, as seen in predictions like **/mama/ (instead of */mamaq/ 'to chew') or **/pana/ (instead of */panaq/ 'to shoot'). BC generally seemed to struggle with word-final obstruents in Proto-Oceanic to a similar extent as my reconstruction models.

While the BC reconstructions for Proto-Oceanic are still better than the WMP reconstructions – which is not surprising at all, given the different complexities

of the models – the gap in performances is significantly smaller. In some cases, WMP even produced better reconstructions than BC, which was hardly ever the case for Proto-Austronesian. Instances of this are */piliq/ 'to choose' (WMP: **/pili/; BC: **/vili/), */puŋa/ 'flower' (WMP: */puŋa/; BC */vuŋa/), or */pulan/ 'moon' (WMP: **/pula/; BC: **/vula/). In all of these cases, WMP correctly reconstructed word-initial */p/ instead of **/v/. For the aforementioned example of */patu/ 'stone', however, the opposite is the case: BC correctly reconstructs the stop, where WMP predicts the fricative.

The fact that these two opposing observations coexist within the relatively small sample data has two implications. First of all, it shows how purely probabilistic models have no concept of regularity or symbolism – otherwise, it would not be possible to reconstruct different proto-sounds for the same sound correspondence. It furthermore shows the value of a good model for assessing the likelihood of sound changes: From the data alone, both */p/ and */v/ seem to be likely proto-sound candidates. One of the main reasons why linguists reconstruct */p/ is their implicit knowledge of directionality! The aformentioned set of forms is a good example to show how ancestral state reconstruction techniques can benefit from both, a more systematic notion of sound correspondences, and a model that contains quantitative information about the typology of sound changes.

6 Summary and Outlook

In this thesis, I have presented a neural model that is able to predict probabilities for arbitrary sound changes. The model contributes two innovations to the field of computational historical linguistics: It is able to provide general information about the directionality and typological markedness of sound changes, factors where classical historical linguists usually rely on their intuition and implicit knowledge rather than on rules or empirical methods. The model therefore simulates such an intuition about which sound changes are generally likely or unlikely on a global scale. Furthermore, the model is able to process any IPA symbol, since it operates on phonological feature representations. Current methods in computational historical linguistics can only process sounds that have been learned, limiting them to an alphabet of a fixed size. Representing individual sounds in terms of their phonological features bears great potential to overcome this limitation and build robust models that are able to process arbitrary IPA strings.

To showcase the potential of this model, ancestral state reconstructions were performed for Proto-Austronesian and Proto-Oceanic, based on ABVD, a dataset that contains lexical data from many modern Austronesian languages. Two techniques that were informed by the neural model were used for this task, namely Weighted Maximum Parsimony and Naïve Bottom-Up Reconstruction. The latter technique, however, was severely limited by the fact that it could only optimize reconstructions locally for each node. It therefore came short even to the Unweighted Maximum Parsimony baseline model which had the advantage of globally optimizing reconstructions over the whole tree. In order to counteract this major weakness, Bottom-Up reconstructions could benefit from iterative processes that propagate information along the tree in both directions, from bottom to top and from top to bottom. This way, reconstructions at each node could factor in information from all languages and not only from its immediate descendants.

The Weighted Maximum Parsimony reconstructions performed better and produced reasonably good reconstructions in many cases. While many reconstructions were identical to those predicted by the unweighted baseline, there were some interesting cases where the weighted model could actually make use of the

information obtained from the neural model, which lead to improved reconstructions. However, these efforts showed that it is a challenging task itself to deduct reasonable substitution costs from the probabilities provided by the neural model, and the method that was applied for this thesis is certainly not the optimal conversion function. Weighted Maximum Parsimony reconstructions could benefit even further from improved substitution costs in the future.

The workflow presented in this thesis can be immediately refined for future applications. When generating the training data, sound transitions were counted in an unweighted fashion, disregarding the distance between a daughter and a parent node. Such a distance could either represent the temporal depth of a branch or its degree of innovation. Weighted transition counts with respect to these factors could emphasize interesting sound changes in the training data, resulting in a model that is better suited to predict these changes.

The Estimation-Maximization component was designed to provide better training data with a stronger signal for frequent sound changes as well, however, it showed no substantial effect. The reason for this again lays within the conversion from transition probabilities to substitution costs, the input format that is needed to generate Weighted Maximum Parsimony Reconstructions. In Section 4, I described why it is necessary to employ non-linear transformations to produce sensible costs. Such a transformation was not employed for generating good costs within the EM iterations. The costs therefore all were very close to 0 (in case of identity) or 1, which does not constitute a workable signal for Sankoff's algorithm. Nearly all of the potential that lies within the EM module was therefore missed. Future work on this approach should address this problem and apply a better cost conversion in order to explore the actual potential of EM within the workflow.

Since the model presented in this thesis is able to process any pair of sounds, it makes it technically possible to reconstruct proto-sounds that can not be attested in the descendant languages. However, the workflow presented in this thesis still relies on a defined alphabet of target sounds, limiting the potential of a model that can generalize over the space of all possible IPA symbols. Techniques that can dynamically infer proto-sound inventories and/or propose sound laws could synergize well with this generalized model of sound change.

Generally, techniques in computational historical linguistics, and especially for ancestral state reconstruction mostly disregard the quasi-regular nature of sound change, since probabilistic models tend to perform better than rule-based, symbolic approaches. While probabilistic approaches are able to learn tendencies from the data quantitatively, they lack the ability to express rules and strictly regular patterns. Reconnecting to the concept of regularity of sound change could enhance methods in computational historical linguistics, since it was the idea that crucially enabled linguists to reconstruct proto-languages in the first place.

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486.

Anderson, C., Tresoldi, T., Chacon, T., Fehn, A.-M., Walworth, M., Forkel, R., and List, J.-M. (2018). A cross-linguistic database of phonetic transcription systems. In *Yearbook of the Poznan Linguistic Meeting*, volume 4, pages 21–53. De Gruyter Open.

Anthony, D. W. and Ringe, D. (2015). The Indo-European homeland from linguistic and archaeological perspectives. *Annual review of linguistics*, 1(1):199–219.

Anttila, R. (1989). *Historical and Comparative Linguistics*, volume 6. John Benjamins Publishing Company.

Baxter, W. H. and Manaster-Ramer, A. (1996). Review: On Calculating the Factor of Chance in Language Comparison. By Donald A. Ringe, Jr. *Diachronica*, 13(2):371–384.

Beinborn, L., Zesch, T., and Gurevych, I. (2013). Cognate production using character-based machine translation. In *Proceedings of the sixth international joint conference on natural language processing*, pages 883–891.

Birchall, J., Dunn, M., and Greenhill, S. J. (2016). A combined comparative and phylogenetic analysis of the Chapacuran language family. *International Journal of American Linguistics*, 82(3):255–284.

Blaschke, V., Daneyko, T., Kaparina, J., Gao, Z., and Dellert, J. (2022). Navigable atom-rule interactions in PSL models enhanced by rule verbalizations, with an

application to etymological inference. In *Prcoeedings of the 31st International Conference on Inductive Logic Programming, ILP2022 @ IJCLR*.

Blevins, J. and Sproat, R. (2021). Statistical evidence for the Proto-Indo-European-Euskarian hypothesis: A word-list approach integrating phonotactics. *Diachronica*, 38(4):506–564.

Blum, F., Barrientos Ugarte, C., Poirier, Z., and Ingunza, A. (2021). Una aproximación filolingüística a la clasificación interna del Quechua. Talk to be presented at Red Europea para los Estudios Andinos, Tübingen.

Blust, R. (1993). Central and central-eastern Malayo-Polynesian. *Oceanic Linguistics*, pages 241–293.

Blust, R. (1996). The Neogrammarian Hypothesis and Pandemic Irregularity. In Durie, M. and Ross, M., editors, *The comparative method reviewed: Regularity and irregularity in language change*. Oxford University Press.

Blust, R. (1999). Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. In *Selected papers from the eighth international conference on Austronesian linguistics*, volume 1, pages 31–94.

Bodt, T. A. and List, J.-M. (2019). Testing the predictive strength of the comparative method: an ongoing experiment on unattested words in Western Kho-Bwa languages. *Papers in Historical Phonology*, 4:22–44.

Bodt, T. A. and List, J.-M. (2022). Reflex prediction: A case study of Western Kho-Bwa. *Diachronica*, 39(1):1–38.

Bouchard-Côté, A., Griffiths, T. L., and Klein, D. (2009). Improved reconstruction of protolanguage word forms. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 65–73.

Bouchard-Côté, A., Hall, D., Griffiths, T. L., and Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Bouchard-Côté, A., Liang, P., Griffiths, T. L., and Klein, D. (2007a). A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896.

Bouchard-Côté, A., Liang, P. S., Klein, D., and Griffiths, T. (2007b). A probabilistic approach to language change. *Advances in Neural Information Processing Systems*, 20.

Bowern, C. and Atkinson, Q. (2012). Computational phylogenetics and the internal structure of Pama-Nyungan. *Language*, pages 817–845.

Campbell, L. (1996). On Sound Change and Challenges to Regularity. In Durie, M. and Ross, M., editors, *The comparative method reviewed: Regularity and irregularity in language change*. Oxford University Press.

Campbell, L. (2013). *Historical linguistics*. Edinburgh University Press.

Campbell, L. and Poser, W. J. (2008). *Language classification: History and method*. Cambridge University Press.

Celano, G. (2022). A Transformer architecture for the prediction of cognate reflexes. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 80–85.

Chacon, T. (2017). Arawakan and Tukanoan contacts in Northwest Amazonia prehistory. *PAPIA: Revista Brasileira de Estudos Crioulos e Similares [Brazilian journal of creole and similar studies]*, 27:237–65.

Chacon, T., Garcia Gonçalves, A., and Ferreira da Silva, L. (2019). A diversidade linguística Aruák no Alto Rio Negro em gravações da década de 1950. *Forma y Función*, 32(2):41–67.

Ciobanu, A. M. and Dinu, L. P. (2018). Ab initio: Automatic Latin proto-word reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1604–1614.

Constenla Umaña, A. (2005). ¿Existe relación genealógica entre las lenguas misumalpas y las chibchenses? *Estudios de Lingüística Chibcha*.

Covington, M. A. (1996). An algorithm to align words for historical comparison. *Computational linguistics*, 22(4):481–496.

Crowley, T. and Bowern, C. (2010). *An introduction to historical linguistics*. Oxford University Press.

Cysouw, M., Wichmann, S., and Kamholz, D. (2006). A critique of the separation base method for genealogical subgrouping, with data from Mixe-Zoquean. *Journal of Quantitative Linguistics*, 13(2-3):225–264.

Daneyko, T. (2020). Automated Sound Law Inference Using Probabilistic Soft Logic. Master's thesis, University of Tübingen.

Daneyko, T. and Bentz, C. (2019). Click languages tend to have large consonant inventories: Implications for language evolution and change. *Modern Human Origins and Dispersal*.

Davletshin, A. (2012). Proto-Uto-Aztecans on their way to the Proto-Aztecan homeland: linguistic evidence. *Journal of Language Relationship*, 8(1):75–92.

de Carvalho, F. O. (2021). A comparative reconstruction of Proto-Purus (Arawakan) segmental phonology. *International Journal of American Linguistics*, 87(1):49–108.

De Vaan, M. (2018). *Etymological dictionary of Latin and the other Italic languages*, volume 7. Brill, Leiden, Boston.

Dekker, P. and Zuidema, W. (2020). Word prediction in computational historical linguistics. *Journal of Language Modelling*, 8(2).

Dellert, J. (2018). Combining information-weighted sequence alignment and sound correspondence models for improved cognate detection. In *Proceedings of the 27th international conference on computational linguistics*, pages 3123–3133.

Dellert, J. (2019). Interactive Etymological Inference via Statistical Relational Learning. In *Workshop on Computer-Assisted Language Comparison at SLE-2019*. Leipzig.

Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H. I., Baysarova, Z., et al. (2020). NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. *Language resources and evaluation*, 54(1):273–301.

Dolgopolsky, A. B. (1964). Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]. *Voprosy jazykoznanija*, 2:53–63.

Dolgopolsky, A. B. (2008). *Nostratic Dictionary*. McDonald Institute for Archaeological Research, Cambridge.

Dunn, M., Kruspe, N., and Burenhult, N. (2013). Time and place in the prehistory of the Aslian languages. *Human Biology*, 85(1/3):383–400.

Farris, J. S. (1970). Methods for computing Wagner trees. *Systematic Biology*, 19(1):83–92.

Feleke, T. L. (2021). Ethiosemitic languages: Classifications and classification determinants. *Ampersand*, 8:100074.

Ferraz Gerardi, F. and Reichert, S. (2021). The Tupí-Guaraní language family: A phylogenetic classification. *Diachronica*, 38(2):151–188.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic zoology*, 19(2):99–113.

Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., and Gray, R. D. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, 5(1):1–10.

Fourrier, C., Bawden, R., and Sagot, B. (2021). Can Cognate Prediction Be Modelled as a Low-Resource Machine Translation Task? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 847–861.

Galucio, A. V., Meira, S., Birchall, J., Moore, D., Gabas Júnior, N., Drude, S., Storto, L., Picanço, G., and Rodrigues, C. R. (2015). Genealogical relations and lexical distances within the Tupian linguistic family. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas*, 10:229–274.

Greenberg, J. H. and Ruhlen, M. (2007). *An Amerind Etymological Dictionary*. Stanford University.

Greenhill, S. J., Blust, R., and Gray, R. D. (2008). The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:EBO–S893.

Greenhill, S. J. and Clark, R. (2011). POLLEX-Online: The Polynesian lexicon project online. *Oceanic Linguistics*, pages 551–559.

Greenhill, S. J., Haynie, H. J., Ross, R. M., Chira, A. M., List, J.-M., Campbell, L., Botero, C. A., and Gray, R. (2022). A Recent Northern Origin for the Uto-aztecan Family.

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., and Pagel, M. (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, 112(43):13296–13301.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Hall, D. and Klein, D. (2010). Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1030–1039.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2022). Glottolog 4.6. Max Planck Institute for Evolutionary Anthropology, Leipzig. URL: `https://glottolog.org/`, DOI: `https://doi.org/10.5281/zenodo.6578297`.

Hauer, B. and Kondrak, G. (2011). Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of 5th international joint conference on natural language processing*, pages 865–873.

Hayes, B. (2008). *Introductory phonology*, volume 7. John Wiley & Sons.

Haykin, S. (1999). *Neural Networks. A Comprehensive Foundation*. Prentice-Hall, Upper Saddle River NJ, 2 edition.

He, A., Tomlin, N., and Klein, D. (2022). Neural unsupervised reconstruction of protolanguage word forms. *arXiv preprint arXiv:2211.08684*.

Hendrycks, D. and Gimpel, K. (2016). Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*.

Hertz, J., Krogh, A., and Palmer, R. G. (2018). *Introduction to the theory of neural computation*. CRC Press.

Hoenigswald, H. M. (1960). *Language Change and Linguistic Reconstruction*. University of Chicago Press.

Hruschka, D. J., Branford, S., Smith, E. D., Wilkins, J., Meade, A., Pagel, M., and Bhattacharya, T. (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, 25(1):1–9.

Hsiu, A. (2015). The classification of Na Meo, a Hmong-Mien language of Vietnam. Handout prepared for SEALS 25 (Chiang Mai, 2015/05/27-29).

Index Diachronica (2016). Index Diachronica v.10.2. URL: `https://chridd.nfshost.com/diachronica/`.

Jäger, G. (2013). Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291.

Jäger, G. (2019). Computational historical linguistics. *Theoretical Linguistics*, 45(3-4):151–182.

Jäger, G. and List, J.-M. (2016). Investigating the potential of ancestral state reconstruction algorithms in historical linguistics. In Bentz, C., Jäger, G., and Yanovich, I., editors, *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*, Tübingen.

Jäger, G. and List, J.-M. (2018). Using ancestral state reconstruction methods for onomasiological reconstruction in multilingual word lists. *Language Dynamics and Change*, 8(1):22–54.

Jäger, G., List, J.-M., and Sofroniev, P. (2017). Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multilingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216.

Jagić, I. V. (1910). *Istorija slavjanskoj filologii*. Enciklopedija slavjanskoj filologii, vol. 1. Academy of Sciences, St. Petersburg.

Jakobson, R. (1958). Typological studies and their contribution to historical comparative linguistics. In Sivertsen, E., editor, *Proceedings of the Eighth International Congress of Linguists*, pages 17–35. Oslo.

Kassian, A., Zhivlov, M., and Starostin, G. (2015). Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies*, 43(3-4):301–347.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirov, C., Sproat, R., and Gutkin, A. (2022). Mockingbird at the SIGTYP 2022 Shared Task: Two Types of Models forthe Prediction of Cognate Reflexes. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 70–79.

Kolipakam, V., Dunn, M., Jordan, F. M., and Verkerk, A. (2018). DravLex: A Dravidian lexical database.

Köllner, M. (2021). *Automatic Loanword Identification Using Tree Reconciliation*. PhD thesis, Universität Tübingen.

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Kroonen, G. (2013). *Etymological Dictionary of Proto-Germanic*. Brill.

Kümmel, M. (2007). *Konsonantenwandel: Bausteine zu einer Typologie des Lautwandels und ihre Konsequenzen für die vergleichende Rekonstruktion*. Reichert Verlag.

Kuryłowicz, J. (1964). On the methods of internal reconstruction. In Lunt, H., editor, *Proceedeings of the Ninth International Congress of Linguists, Cambridge, Massachusetts*, pages 9–36. Mouton, The Hague.

Lass, R. (1993). How real(ist) are reconstructions? In Jones, C., editor, *Historical linguistics: Problems and perspectives*, pages 156–189.

Lass, R. (2017). Reality in a soft science: the metaphonology of historical reconstruction. *Papers in Historical Phonology*, 2:152–163.

Lee, S. (2015). A sketch of language history in the Korean Peninsula. *PLoS One*, 10(5):e0128448.

Lee, S. and Hasegawa, T. (2011). Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proceedings of the Royal Society B: Biological Sciences*, 278(1725):3662–3669.

Lee, S. and Hasegawa, T. (2013). Evolution of the Ainu language in space and time. *PLoS One*, 8(4):e62243.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Lionnet, A. (1985). Relaciones internas de la rama sonorense. *Amerindia*, 10:25–58.

List, J.-M. (2014). *Sequence Comparison in Historical Linguistics*. Düsseldorf University Press.

List, J.-M. (2019a). Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*, 45(1):137–161.

List, J.-M. (2019b). Beyond Edit Distances: Comparing linguistic reconstruction systems. *Theoretical Linguistics*, 45(3-4):1–10.

List, J.-M. (2022). Computational Approaches to Historical Language Comparison. [Preprint not peer reviewed]. Max Planck Institute for Evolutionary Anthropology, Leipzig.

List, J.-M., Cysouw, M., and Forkel, R. (2016a). Concepticon: A resource for the linking of concept lists.

List, J.-M., Forkel, R., Greenhill, S. J., Rzymski, C., Englisch, J., and Gray, R. D. (2022a). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):1–16.

List, J.-M., Forkel, R., and Hill, N. W. (2022b). A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. *arXiv preprint arXiv:2204.04619*.

List, J.-M., Greenhill, S. J., and Gray, R. D. (2017). The potential of automatic word comparison for historical linguistics. *PloS one*, 12(1):e0170046.

List, J.-M., Lopez, P., and Bapteste, E. (2016b). Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Berlin. Association of Computational Linguistics.

List, J.-M., Vylomova, E., Forkel, R., Hill, N. W., and Cotterell, R. D. (2022c). The SIGTYP 2022 shared task on the prediction of cognate reflexes. In *Proceedings of the 4th Workshop on Computational Typology and Multilingual NLP (SIGTYP 2022)*, pages 52–62. Association for Computational Linguistics.

Liú, L., Wáng, H., and Bǎi, Y. (2007). Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]. *Nánjīng: Fènghuáng*.

Luangthongkum, T. (2019). A view on Proto-Karen phonology and lexicon. *Journal of the Southeast Asian Linguistics Society*, 12(1):i–lii.

Lundgren, O. (2020). A phonological reconstruction of Proto-Omagua–Kokama–Tupinambá. Master's thesis, Lund University.

Mann, N. W. (1998). *A phonological reconstruction of Proto Northern Burmic*. The University of Texas at Arlington.

McElhanon, K. A. (1967). Preliminary observations on Huon Peninsula languages. *Oceanic Linguistics*, 6(1):1–45.

McMahon, A. M. S. (1994). *Understanding language change*. Cambridge University Press.

Meillet, A. (1958). *Linguistique historique et linguistique generate*. Librairie Honore Champion, Paris.

Meloni, C., Ravfogel, S., and Goldberg, Y. (2019). Ab Antiquo: Neural proto-language reconstruction. *arXiv preprint arXiv:1908.02477*.

Miller, J. E., Tresoldi, T., Zariquiey, R., Beltrán Castañón, C. A., Morozova, N., and List, J.-M. (2020). Using lexical language models to detect borrowings in monolingual wordlists. *Plos one*, 15(12):e0242709.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. (2016). Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.

Nagaraja, K. S., Sidwell, P., and Greenhill, S. J. (2013). A lexicostatistical study of the Khasian languages: Khasi, Pnar, Lyngngam, and War. *Mon-Khmer Studies*, 42:1–11.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Nichols, J. (1993). The linguistic geography of the Slavic expansion. In Maguire, R. A. and Timberlake, A., editors, *American Contributions to the Eleventh International Congress of Slavists, Bratislava, August-September 1993: Literature, Linguistics, Poetics*, pages 377–91. Slavica, Columbus.

Nichols, J. (1996). The Comparative Method as Heuristic. In Durie, M. and Ross, M., editors, *The comparative method reviewed: Regularity and irregularity in language change*. Oxford University Press.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–217.

Osthoff, H. and Brugmann, K. (1878). *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*, volume 1. Hirzel.

Peiros, I. J. (2004a). Comparative Sino-Tibetan Wordlist.

Peiros, I. J. (2004b). Genetičeskaja klassifikacija avstroaziatskix jazykov. *Moskva: Rossijskij gosudarstvennyj gumanitarnyj universitet (doktorskaja dissertacija)*.

Pharao Hansen, M. (2020). ¿Familia o vecinos? Investigando la relación entre el proto-náhuatl y el proto-corachol. [Family or neighbors? Investigating the relation between Proto-Náhuatl and Proto-Corachol]. In Rosales, R. H. Y., editor, *Lenguas yutoaztecas: historia, estructuras y contacto lingüístico: homenaje a Karen Dakin*. Universidad de Guadalajara.

Pope, M. K. (1934). *From Latin to Modern French with especial consideration of Anglo-Norman: Phonology and morphology.* Number 6. Manchester University Press.

Rama, T. (2016). Chinese restaurant process for cognate clustering: A threshold free approach. *arXiv preprint arXiv:1610.06053.*

Rama, T. and List, J.-M. (2019). An automated framework for fast cognate detection and Bayesian phylogenetic inference in computational historical linguistics. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 6225–6235. Association for Computational Linguistics.

Ratcliffe, R. R. (2021). The glottometrics of Arabic: Quantifying linguistic diversity and correlating it with diachronic change. *Language Dynamics and Change*, 11(1):1–29.

Ringe, D. A. (1992). On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society*, 82(1):1–110.

Ringe, D. A. (1999). How hard is it to match CVC-roots? *Transactions of the Philological Society*, 97(2):213–244.

Robinson, L. C. and Holton, G. (2012). Internal classification of the Alor-Pantar language family using computational methods applied to the lexicon. *Language Dynamics and Change*, 2(2):123–149.

Ross, M. and Durie, M. (1996). Introduction. In Durie, M. and Ross, M., editors, *The comparative method reviewed: Regularity and irregularity in language change.* Oxford University Press.

Ross, M., Pawley, A., and Osmond, M. (2007). *The Lexicon of Proto Oceanic: The culture and environment of ancestral Oceanic society.* ANU Press.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., Gast, V., Bodt, T. A., Hantgan, A., Kaiping, G. A., et al. (2020). The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):1–12.

Saenko, M. (2015). Annotated Swadesh wordlists for the Romance group (Indo-European family). In Starostin, G., editor, *The Global Lexicostatistical Database.* RGU, Moscow.

Sagart, L., Jacques, G., Lai, Y., Ryder, R. J., Thouzeau, V., Greenhill, S. J., and List, J.-M. (2019). Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences*, 116(21):10317–10322.

Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28(1):35–42.

Savelyev, A. and Robbeets, M. (2020). Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution*, 5(1):39–53.

Schleicher, A. (1863). *Die Darwinsche Theorie und die Sprachwissenschaft*. Weimar.

Sidwell, P. (2015). Austroasiatic dataset for phylogenetic analysis: 2015 version.

Sidwell, P. and Alves, M. (2021). Vietic 116 item phylogenetic lexicon (First version (26 Aug 2021)eng) [Data set]. In *9th International Conference on Austroasiatic Linguistics (ICAAL 9)*. Lund.

Sims, N. A. (2020). Reconsidering the diachrony of tone in Rma. *Journal of the Southeast Asian Linguistics Society*, 13(1):53–85.

Smith, N. A. and Eisner, J. (2005). Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438.

Starostin, G. S. (2015a). Annotated Swadesh wordlists for the Hmong group (Hmong-Mien family). In Starostin, G. S., editor, *The Global Lexicostatistical Database*. RGU, Moscow.

Starostin, G. S. (2015b). Annotated Swadesh wordlists for the Tujia group. In Starostin, G. S., editor, *The Global Lexicostatistical Database*. RGU, Moscow.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Syrjänen, K., Honkola, T., Korhonen, K., Lehtinen, J., Vesakoski, O., and Wahlberg, N. (2013). Shedding more light on language classification using basic vocabularies and phylogenetic methods: A case study of Uralic. *Diachronica*, 30(3):323–352.

Turchin, P., Peiros, I., and Gell-Mann, M. (2010). Analyzing genetic connections between languages by matching consonant classes. *Journal of Language Relationship*, (3):117–126.

Walworth, M. (2018). Polynesian segmented data (version 1). *Jena: Max Planck Institute for the Science of Human History,(doi: 10.5281/zenodo. 1689909)*.

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., Holman, E. W., et al. (2013). The ASJP Database (version 16). Leipzig. Max Planck Institute for Evolutionary Anthropology.

Wu, M.-S., Schweikhard, N. E., Bodt, T., Hill, N. W., and List, J.-M. (2020). Computer-Assisted Language Comparison: State of the Art. *Journal of Open Humanities Data*, 6(2):1–14.

Yang, C. et al. (2010). *Lalo regional varieties: Phylogeny, dialectometry, and sociolinguistics*. La Trobe University.

Zhang, S., Guillaume, J., and Lai, Y. (2019). A study of cognates between Gyalrong languages and Old Chinese. *Journal of Language Relationship*, 17(1-2):73–92.

A Reconstructions

Tables of all automatic Proto-Austronesian and Proto-Oceanic reconstructions, generated by Weighted Maximum Parsimony (**WMP**), Naïve Bottom-Up Reconstructions (**BU**), Maximum Parsimony (**MP**), and the model described in Bouchard-Côté et al. (2013) (**BC**). For Proto-Oceanic, both expert reconstructions by Blust (**B**) and Pawley (**P**) are given as gold standard.

A.1 Proto-Austronesian

Cognate ID	Gold Standard	WMP	BU	MP	BC
hand-1	*lima	*lima	*lima	*lima	*lima
left-1	*ka-wiʀi	*kiri	*kairi	*kiri	*kawiʀi
legfoot-1	*qaqay	*kai	*kakay	*kay	*qaqay
roadpath-1	*zalan	*daan	*dalan	*daan	*zalan
back-1	*likud	*likus	*likus	*rikus	*likud
breast-1	*susu	*susu	*susu	*susu	*susu
shoulder-1	*qabaʀa	*abaa	*qabana	*abafa	*qabaʀa
[1]	*bajaq	*avaa	*mavaka	*avaza	*mafana?
tofear-1	*ma-takut	*atakut	*matakut	*mtakut	*matakut
blood-1	*daʀaq	*tara	*tara	*cara	*daʀaq
head-1	*qulu	*ulu	*ulu	*ulu	*qulu
neck-1	*liqeʀ	*lil	*lig	*ril	*liqeʀ
mouth-2	*ŋusu	*ŋusu	*mgonou	*ŋusu	*ŋuju
tooth-1	*nipen	*nipən	*nipun	*nipən	*nipen
tovomit-1	*utaq	*muta	*mutah	*muta	*utaq
toeat-1	*kaen	*kan	*kaman	*kan	*kman
tochew-2	*qelqel	*məlqə	*fiəməwqər	*əmərqər	*qmelqel
tocook-1	*tanek	*talak	*munomanuk	*talak	*tanek
tobite-1	*kaʀat	*kat	*kamat	*kat	*kaʀat
tosuck-1	*sepsep	*səpsep	*supsep	*səpsep	*sepsep
tosee-1	*kita	*kita	*mkmita	*kita	*kita
tosleep-1	*tuduʀ	*maturu	*maturu	*matur	*tuduʀ

Continued on next page

[1] toknowbeknowledgeable-2

Proto-Austronesian – continued from previous page

Cognate ID	Gold Standard	WMP	BU	MP	BC
tostand-1	*diʀi	*miri	*mairi	*miri	*diʀi
womanfemale-1	*bahi	*babin	*babaian	*babin	*vavaian
mother-1	*t-ina	*ina	*ina	*ina	*tina
father-1	*t-ama	*tama	*ama	*tama	*tama
house-1	*ʀumaq	*umaq	*lumaq	*umaq	*ʀumaq
name-1	*ŋajan	*ŋaan	*nanan	*ŋaŋan	*ŋajan
needle-1	*zaʀum	*daum	*daum	*daum	*zaʀum
toshoot-1	*panaq	*pana	*panah	*pana	*panaq
tohit-1	*palu	*palu	*palu	*palu	*palu
tolivebealive-1	*ma-qudip	*murip	*maurip	*murip	*maqudip
toscratch-1	*kaʀaw	*kakao	*akakao	*kakao	*kaʀaw
tocuthack-1	*taʀaq	*taca	*taxa	*taca	*taʀaq
tocuthack-3	*tektek	*tatak	*tatak	*tatak	*tektek
towork-1	*qumah	*muma	*maumah	*mumah	*quma
toplant-1	*mula	*mama	*mamamamoah	*mawaʔ	*mula
tochoose-1	*piliq	*pili	*pili	*pili	*piliq
toswell-1	*baʀeq	*baa	*bnana	*baga	*baʀeq
toswell-26	*ʀibawa	*mlibawa	*malibawa	*mlibawaʔ	*abeh
tosqueeze-1	*peʀeq	*peqa	*pəmah	*peqa	*pereq
tohold-1	*gemgem	*gumkem	*komkom	*gumkem	*gemgem
todig-1	*kalih	*kali	*kali	*kali	*kali
tobuy-1	*beli	*beli	*vani	*bali	*beli
tobuy-23	*baliw	*baliv	*mabaliw	*baliu	*taiw
topoundbeat-20	*tutuh	*tutu	*nutu	*tutu	*tutu
dog-1	*wasu	*asu	*masu	*asu	*vatu
bird-2	*qayam	*ayam	*awam	*ayam	*qayam
tofly-2	*layap	*mayap	*mayap	*mayap	*layap
rat-1	*labaw	*kulabaw	*kulabaw	*kulabaw	*kulavaw
meatflesh-31	*isi	*isi	*isi	*isi	*isi
tail-1	*ikuʀ	*iku	*iku	*iku	*ikuʀ
rotten-1	*ma-buʀaq	*buruk	*buruk	*baruk	*maburuq
leaf-2	*biʀaq	*bia	*biaɦ	*bia	*bela
fruit-1	*buaq	*bua	*bua	*bua	*buaq
stone-1	*batu	*batu	*batu	*batu	*batu
sand-1	*qenay	*ənay	*onay	*ənay	*qenay
toflow-1	*qaluʀ	*mŋalir	*mŋalir	*mŋalir	*qaluʀ
salt-1	*qasiʀa	*sia	*sila	*sira	*qasiʀa
salt-2	*timus	*timu	*timuh	*timu	*timus
lake-1	*danaw	*danaw	*djanaw	*danaw	*danaw
star-1	*bituqen	*bituan	*bituqan	*bituan	*bituqen
thunder-3	*deʀuŋ	*səruŋ	*zoruŋ	*səruŋ	*deruŋ
wind-2	*bali	*bari	*bali	*bari	*beliu
smoke-1	*qebel	*ʔəbəl	*kobal	*ʔəbər	*qebel
ash-1	*qabu	*abu	*qabu	*abu	*qabu
green-1	*mataq	*mataa	*maataha	*matacha	*mataq
small-2	*kedi	*kaɖi	*kati	*kati	*kedi
big-1	*ma-ʀaya	*maya	*mawa	*ʀaya	*maʀaya
long-1	*inaduq	*naruq	*inaruq	*naruq	*anaduq
wide-1	*ma-lawas	*lawa	*malawa	*ɭawa	*malabeʀ
new-1	*ma-baqeʀu	*bau	*bahu	*baru	*vaquan
night-1	*beʀŋi	*beŋi	*beŋi	*beŋi	*beʀŋi
day-1	*qalejaw	*adaw	*kalaw	*adaw	*qalejaw
when-1	*ija-n	*aida	*kaiza		*pijan

Continued on next page

Proto-Austronesian – continued from previous page

Cognate ID	Gold Standard	WMP	BU	MP	BC
at-1	*i	*i	*i	*i	*i
at-20	*di	*di	*di	*di	*id
ininside-1	*i-dalem	*lalum	*lalom	*dalam	*idalem
this-1	*i-ni	*ini	*mini	*ini	*ani
where-1	*i-nu	*inu	*hinu	*inu	*ainu
i-1	*i-aku	*aku	*kaku	*aku	*iaku
heshe-1	*si-ia	*sia	*saia	*sia	*siia
we-1	*i-kita	*ita	*ita	*kita	*kita
we-2	*kami	*kami	*kami	*kami	*kami
you-1	*i-kamu	*kamu	*mamu	*kamu	*kamu
what-2	*n-anu	*nanu	*nanuh	*nanu	*anu
who-2	*si-ima	*tima	*tima	*tima	*tima
other-1	*duma	*ruma	*zuma	*ruma	*duma
all-1	*amin	*mamin	*mamin	*gamin	*amin
and-1	*ka	*ka	*ka	*ka	*ka
and-2	*mah	*ma	*ma	*ma	*ma
if-1	*ka	*ka	*ka	*ka	*ka
if-2	*nu	*nu	*nu	*nu	*nu
how-1	*kuja	*akua	*makua	*kakua	*kua
nonot-3	*ini	*ini	*ini	*ini	*ini
one-1	*isa	*sa	*isa	*sa	*isa
three-1	*telu	*turu	*tolu	*turu	*telu
five-1	*lima	*lima	*lima	*lima	*lima

A.2 Proto-Oceanic

Cognate ID	Gold (B)	Gold (P)	WMP	BU	MP	BC
hand-1	*lima	*lima	*lima	*lima	*lima	*lima
left-1	*mawiʀi	*mawiʀi	*mail	*mauoni	*mail	*mawii
legfoot-1	*waqe	*waqe	*kae	*kae	*kae	*waqe
towalk-2	*pano	*pano	*van	*ahanoa	*van	*vano
dust-1	*qapuk	*qapuk	*afu	*kahu	*afu	*avu
skin-1	*kulit	*kulit	*kuli	*kunina	*kuli	*kulit
liver-1	*qate	*qate	*ate	*katena	*ate	*qate
breast-1	*susu	*susu	*susu	*susu	*susu	*susu
shoulder-1	*paʀa	*qapaʀa	*avala	*kabaha	*avala	*vara
tofear-1	*matakut	*matakut	*matau	*mataku	*matau	*matakut
head-1	*qulu	*qulu	*ulu	*ulun	*ulu	*qulu
neck-18	*ʀuqa	*ʀuqa	*ua	*wuna	*ua	*ua
nose-1	*isuŋ	*ijuŋ	*isu	*phun	*isu	*isu
tobreathe-1	*manawa	*manawa	*manawa	*manawa	*manawa	*manawa
tovomit-1	*mumutaq	*mumuta	*muta	*momumua	*muta	*muta
tovomit-8	*luaq	*luaq	*lua	*lua	*lua	*lua
tospit-14	*qanusi	*qanusi	*anusu	*kamusu	*anusu	*aŋusu
toeat-1	*kani	*kani	*an	*kaani	*an	*kani
tochew-1	*mamaq	*mamaq	*mama	*mama	*mama	*mama
tocook-9	*tunu	*tunu	*tunu	*tunu	*tunu	*tunu
todrink-1	*inum	*inum	*inu	*munum	*inu	*inum
tobite-1	*kaʀat	*kaʀati	*kat	*kaai	*kat	*karat
tohear-1	*roŋoʀ	*roŋoʀ	*roŋo	*loŋo	*roŋo	*roŋo

Continued on next page

Proto-Oceanic – continued from previous page

Cognate ID	Gold (B)	Gold (P)	WMP	BU	MP	BC
eye-1	*mata	*mata	*mata	*mata	*mata	*mata
tosee-1	*kita	*kita	*ite	*kite	*ite	*kita
toyawn-1	*mawap	*mawap	*mama	*mamama	*mama	*mawa
toliedown-1	*qinop	*qeno	*eno	*weno	*eno	*eno
tosit-16	*nopo	*nopo	*nofo	*noho	*nofo	*nofo
tostand-2	*tuqur	*taqur	*tu	*tuuu	*tu	*tuqu
[2]	*taumataq	*tamwata	*tamata	*tamaa	*tamata	*tamata
manmale-1	*mwaʀuqane	*taumwaqane	*mane	*mamamane	*mane	*mwane
father-1	*tama	*tamana	*tama	*taman	*tama	*tama
thatchroof-1	*qatop	*qatop	*ato	*noto	*ato	*qato
name-1	*ŋajan	*qajan	*asa	*akan	*ara	*qasa
rope-1	*tali	*tali	*tal	*tali	*tal	*tali
needle-1	*saʀum	*saʀum	*saum	*saum	*saum	*sau
toshoot-1	*panaq	*pana	*pan	*pana	*fan	*pana
tostabpierce-8	*soka	*soka	*soka	*soaka	*soka	*soka
tolivebealive-1	*maqurip	*maqurip	*mauri	*mauli	*mauri	*maquri
toscratch-44	*karu	*kadru	*karu	*karui	*karu	*kadru
stickwood-1	*kayu	*kayu	*kai	*kai	*kai	*kai
toplant-2	*tanum	*tanom	*tan	*tano	*tan	*tanəm
tochoose-1	*piliq	*piliq	*pili	*hilii	*pile	*vili
togrow-1	*tubuq	*tubuq	*tupu	*tobu	*tupu	*tubu
todig-1	*keli	*keli	*keli	*keli	*keli	*keli
tobuy-1	*poli	*poli	*voli	*boli	*voli	*voli
topoundbeat-2	*tutuk	*tuki	*tuki	*tuki	*tuki	*tutuk
bird-1	*manuk	*manuk	*manu	*manu	*manu	*manu
egg-1	*qatoluʀ	*katoluʀ	*tolu	*nakolun	*tolu	*tolu
feather-1	*pulu	*pulu	*fulu	*mulnun	*fulu	*vulu
meatflesh-1	*pisiko	*pisako		*pikikon		*kiko
snake-12	*mwata	*mwata	*mata	*mawa	*mata	*mwata
louse-1	*kutu	*kutu	*kutu	*kuu	*kut	*kutu
mosquito-1	*namuk	*namuk	*namu	*namo	*namu	*namu
fish-1	*ikan	*ikan	*ika	*nika	*ika	*ikan
leaf-1	*raun	*rau	*rau	*louena	*rau	*dau
root-2	*wakaʀa	*wakaʀ	*waka	*wokana	*waka	*waka
flower-1	*puŋa	*puŋa	*puŋa	*puŋa	*puŋa	*vuŋa
fruit-1	*puaq	*puaq	*vua	*huanana	*ua	*vua
stone-1	*patu	*patu	*vatu	*batu	*vat	*patu
water-2	*waiʀ	*waiʀ	*wai	*wai	*wai	*wai
toflow-1	*tape	*tape	*tafe	*tahe	*tate	*tave
sky-1	*laŋit	*laŋit	*laŋ	*laaŋi	*laŋ	*laŋi
moon-1	*pulan	*pulan	*pula	*bula	*pula	*vula
star-1	*pituqun	*pituqon	*pitu	*bihuuu	*pitu	*vetuqu
fog-1	*kaput	*kaput	*kapu	*kabu	*kapu	*kabu
rain-1	*qusan	*qusan	*usa	*uwa	*uha	*usa
wind-1	*aŋin	*mataŋi	*aŋi	*mannen	*aŋi	*mataŋi
warm-1	*mapanas	*mapanas	*mapana	*mahana	*mafana	*mavana
dry-11	*maca	*masa	*mamaa	*mamama	*mamasa	*mamasa
heavy-1	*mamat	*mapat	*mata	*mamama	*mata	*mamava
fire-1	*api	*api	*api	*waihi	*afi	*avi
smoke-2	*qasu	*qasu	*asu	*kahu	*asu	*qasu
thick-3	*matolu	*matolu	*matolu	*matolu	*matolu	*matolu

Continued on next page

[2] personhumanbeing-1

Proto-Oceanic – continued from previous page

Cognate ID	Gold (B)	Gold (P)	WMP	BU	MP	BC
narrow-1	*kopit	*kopit	*kopiti	*kopiui	*kopiti	*kapi
old-1	*matuqa	*matuqa	*matua	*matua	*matua	*matuqa
badevil-1	*saqat	*saqat	*sa	*maha	*sa	*saqa
night-1	*boŋi	*boŋi	*boŋ	*boni	*boŋ	*boŋi
year-1	*taqun	*taqun	*tau	*tau	*tau	*taqu
when-1	*ŋaican	*ŋaijan	*naisa	*nanisa	*naisa	*ŋisa
tohide-1	*puni	*puni	*muni	*mmuni	*muni	*vuni
toclimb-2	*sake	*sake	*sak	*make	*sae	*cake
this-1	*ne	*ani	*ni	*anee	*ni	*eni
near-9	*tata	*tata	*tata	*athata	*tata	*tata
far-1	*sauq	*sauq	*sau	*mao	*sau	*sau
where-3	*pai	*pea	*ve	*ahea	*ve	*vea
i-1	*au	*au	*au	*nau	*au	*yau
thou-1	*ko	*iko	*o	*iko	*ko	*kou
heshe-1	*ia	*ia	*ia	*nina	*ia	*ia
we-2	*kamami	*kami	*ami	*mami	*ami	*kami
they-1	*ira	*ira	*ira	*kila	*ira	*sira
what-1	*sapa	*saa	*aa	*aha	*aha	*sava
how-1	*kua	*kuya	*eɣua	*nekua	*eɣua	*kua
one-1	*sakai	*tasa	*sa	*esaa	*sa	*sa
two-1	*rua	*rua	*rua	*rua	*rua	*rua
three-1	*tolu	*tolu	*tolu	*tolu	*tolu	*tolu
four-1	*pani	*pat	*vat	*hasa	*vat	*vati
five-1	*lima	*lima	*lima	*lima	*lima	*lima