Probabilistic Evaluation of Comparative Reconstructions

by

Andrei Munteanu

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Linguistics
University of Toronto

Probabilistic Evaluation of Comparative Reconstructions

Andrei Munteanu

Doctor of Philosophy

Department of Linguistics
University of Toronto

2023

Abstract

This dissertation introduces Wordlist Distortion Theory, a framework for the probabilistic evaluation of comparative reconstructions in historical linguistics. The framework estimates the likelihood that a randomly generated wordlist merits the same type and number of diachronic transformations (e.g. sound changes, replacements, etc.) as required by the reconstruction. This is the first probabilistic evaluation of comparative reconstruction of its kind.

Wordlist Distortion Theory is primarily intended as a platform for objective and accessible debate surrounding spurious reconstruction in historical linguistics. Additionally, the framework can be used as a tiebreaker between conflicting reconstructions for the same data. Finally, the framework allows for probability-based theoretical arguments in historical linguistics about the interaction of synchronic and diachronic factors with reconstruction reliability. For example, this dissertation argues that the effect of semantic change on reconstruction reliability is relatively minor and the effect of contrast-preserving sound change almost negligible.

Wordlist Distortion Theory can also feature as part of a machine learning algorithm, allowing for the stochastic generation and evaluation of comparative reconstructions. This dissertation presents the results of a case study conducted on 74 Austronesian languages and 5 proto-languages, where, for each pair of language and proto-language, the algorithm suggested sound changes with the goal of decreasing the probability of a random match as calculated by Wordlist Distortion Theory. The resulting reconstructions are in line with general knowledge about sound change and Austronesian historical linguistics. Additionally, Wordlist Distortion Theory was used to evaluate the putative connection between the Austronesian and Ongan language families. Automated reconstructions from Proto-Ongan-Austronesian to members of either family are not conclusive.

Acknowledgments

After years of sailing through the capricious and eerie waters of grad school, land is finally in sight. On the surface it seems like a miracle and a half that I am able to set my foot ashore. However, this miracle is in fact the result of many hours of hard work, in large part not my own.

Peter Jurgec, my supervisor, was truly a godsend. When I was drifting afloat, he would say the exact thing to put me in motion; when I was heading the right direction, he would mostly just listen. I have never seen him put less than 100% into anything, including my supervision. Who knows how many late hours he spent patiently reading nigh incomprehensible drafts. At this point, I can probably collate his comments into a second dissertation.

I was no less lucky with my choice of committee members in Barend Beekhuizen and Nathan Sanders. I would like to thank Nathan for his sharp and creative vision. There have been quite a number of things-are-now-snapping-into-focus moments during our conversations. I would like to thank Barend for taking over as my supervisor at the very end. His comments throughout the project helped me stay grounded and keep thinking about the next step.

Additionally, I'd like to thank all other researchers whose work directly or indirectly went into improving the dissertation. First and foremost, these are external members of the committee in Alexei Kochetov, the first linguist I've ever met, and Jessamyn Schertz, the kindest linguist I've ever met. Here, I would also like to thank the external reader, Johann-Mattis List, who helped secure the docking by ensuring that the ü's are dotted and the ø's are crossed. Secondly, I'd like to thank anyone who contributed to the discussion surrounding the project in a less official manner: the participants of the 42nd TABU conference in Groningen, the 2021 AMP meeting in Toronto, 28th Manchester Phonology Meeting, as well as the attendees of the Phonology/Phonetics reading group at the University of Toronto.

Finally, it must be acknowledged that family and friends played a huge part in the dissertation as well. My wife, Eva Plesnik, proof-read what seems like no less than a kilogram of drafts. My childhood friend, Edward Moskovsky (who, in the time it took me to write a dissertation, brought two wonderful children into this world) would chat with me about math, stats, and (unfortunately for him) linguistics, some of which has made it into the dissertation.

Most importantly, I'd like to thank my parents, Andrei and Lilia. Their love, like a ray of sunlight, has shined through the darkest hours of my life, protecting me from harm.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

The comparative method is the primary technique of establishing genetic relationships in historical linguistics. Throughout the comparative reconstruction process, the researcher infers the properties of an ancestral language based on a feature-by-feature comparison of its descendants (Weiss, 2015). Since its origins in the first half of the 19th century, this technique has yielded numerous successes. Most famously, thanks in large part to the comparative method, Proto-Indo-European, an undocumented language spoken at least six thousand years ago, is not only now uncontroversial but explored in seemingly every facet, from segmental inventory and morphology (Ringe, 2017), to syntax and word-order (Kulikov & Lavidas, 2015), to poetic meter (Tuck, 2006). Proto-Indo-European is by no means the only one of its kind. *Glottolog*, an online catalogue of languages and language families, lists 429 top-level language families and language isolates (Hammarström et al., 2022), each with its own implied proto-language.

However, not all products of the comparative method have received the same level of support among scholars. In practice, reconstructions in historical linguistics lie on a spectrum of perceived reliability. On one end of the spectrum, one may find the proto-languages deeply entrenched in the literature, e.g. Proto-Indo-European, Proto-Austronesian, and Proto-Dravidian. On the other end, there exist the proto-languages with little backing outside the original proposal. These range from the most grandiose in Proto-Human (Ruhlen, 1994), to the relatively peripheral in Proto-Ongan-Austronesian, the putative ancestor of the Austronesian languages and the Ongan languages of the Andaman Islands (Blevins, 2007).

Many proto-languages seem to lie somewhere in between the two extremes. For example, Proto-Nostratic, popularized by Illic-Svityc (1963), is the proposed ancestor of Indo-European, Dravidian, Kartvelian, Altaic, and Afro-Asiatic. Although most linguists tend to dismiss the hypothesis (Campbell, 1998; Ringe, 1999), publications arguing or assuming its validity have not subsided (Bomhard, 2008; Dolgopolsky, 2008). In contrast, Proto-Afro-Asiatic, one of the Nostratic branches, is generally without controversy and, unlike its purported ancestor above, is listed in respected language databases, such as *Glottolog* (Hammarström et al., 2022) and *Ethnologue* (Eberhard et al., 2023). However, the membership of Omotic in Afro-Asiatic has provoked some doubt (Theil, 2006). Meanwhile, a comparison of two different Proto-Afro-

Asiatic reconstructions revealed only 6% overlap in suggested cognate pairs (Ratcliffe, 2003). Another example is the Dené-Yeniseian hypothesis (Vajda, 2010; Vajda, 2018), linking the Yeniseian languages of the old world with the Dené languages of the new. The grouping is based on lexical and morphological comparisons and is corroborated by genetic evidence (Rubicz et al., 2002), kinship terminology (Vajda et al., 2010) and comparative mythology (Berezkin, 2019). The hypothesis is endorsed by some linguists (Kiparsky, 2015), contested by others (Campbell, 2011; Starostin, 2012), while others still (Dunn, 2012), including the project's author (Vajda, 2011), recognize that more work is to be done before a conclusion can be reached. Totozoquean (Brown et al., 2011), which links the Totonocan and Mixe-Zoquean language families of Mexico and later Chitimacha (Brown et al., 2014), a language isolate in the US, also has not been deliberated upon to conclusion by the scholarly community.

The perceived reliability of some proto-languages has changed over time. One such example is the Proto-Altaic hypothesis, another branch of Nostratic and the purported ancestor of Proto-Uralic, Proto-Turkic, Proto-Mongolic, Proto-Tungusic, Korean, and later Proto-Japonic (Starostin et al., 2003; Robbeets, 2005). Since its inception, the proposal was championed by some (Ramstedt, 1912; Poppe, 1960) and heavily criticized by others (Doerfer, 1963; Vovin, 2005). At the end of the last millennium, the scales seem to have tipped against the hypothesis, and an overview of the discussion feels the need to reassure the reader that "opinions continue to be divided about the validity of the Altaic theory, but even a cursory look at the relevant literature […] reveals that the issue is far from being settled in the negative" (Georg et al., 1999:91). A few years later, the world would see the publication of the *Etymological Dictionary of the Altaic Languages* (Starostin et al., 2003), which may have swayed some researchers away from the project (Vovin, 2005). The status of the hypothesis in the 21st century appears to be that while "negative criticism has been very influential, leading almost to a consensus that no Altaic language family exists, supporters of the Ramstedt-Poppe theory have by no means disappeared" (Norman, 2009).

Perhaps one of the main causes of the uncertainty surrounding Proto-Altaic, Proto-Nostratic, Proto-Dené-Yeniseian, and other contested proto-language proposals, is the lack of a quantitative measure for the evaluation of individual proposals or individual arguments. Typically, arguments on either side of historical linguistic debates, such as the one surrounding Proto-Altaic, tend to be opaque and language-specific, requiring not only a solid foundation in linguistics but also an intimate knowledge of the languages in question. For example, the *Etymological Dictionary of the Altaic Languages* (Starostin et al., 2003:1269) proposes the Proto-Japonic form *situ ‘damp’, motivated by Middle Japanese [situ] and itself used to support Proto-Altaic *si̯ắŕi ‘earth’. A critique of the dictionary points out that the Middle Japanese form is better explained as an *on-yomi* of 濕 ‘damp’, Modern Japanese [ɕi͡tsu], i.e. a Sino-Japanese loan. The discussion is complicated both by the fact that the presumed Chinese ancestor exhibited a labial stop rather than a coronal one, cf. Cantonese [sap˥], and that there exists a more frequent codaless *on-yomi* for the same character, as in Middle Japanese [sjuu], Modern Japanese [ɕuu] (Vovin, 2005:79).

Even if these language-specific arguments could be reliably assessed by non-specialists, it is still not clear to what extent the particular critique, if justified, detracts from the hypothesis overall. Even the most impassioned detractors of Altaic would likely grant that a single erroneous form does not invalidate the entire language family. Likewise, even the most loyal defenders of the project would concede if every reconstructed Proto-Altaic form is shown to be incorrect. The number of contested forms appears to lie somewhere in between these two extremes. As such, a consensus on the issue is not forthcoming.

The current dissertation addresses this gap in the field by introducing Wordlist Distortion Theory, a probabilistic framework for the evaluation of comparative reconstructions. The primary intent is for this framework to serve as a foundation for objective arguments and discussion in historical linguistics, distilling and expediting the debate surrounding spurious hypotheses. However, the methodology offers a myriad of other applications. Because Wordlist Distortion Theory analyses the reconstruction, rather than the dataset, the framework is perfectly suited as a tiebreaker between competing reconstructions for the same data. Competing proposals can simply be evaluated independently, with the winner chosen by comparing the results. In the same way, Wordlist Distortion Theory can help choose between clashing sound changes or proto-forms during the reconstruction process. Furthermore, the framework may also function as a component of machine learning algorithms in historical linguistics. Finally, the framework can be used to reason about diachronic linguistics theoretically, allowing researchers to quantify the correlation between reconstruction validity and various aspects of language, such as segmental inventory, contrast, and phonotactics.

This dissertation serves as a guidebook for Wordlist Distortion Theory, at the end of which the reader will not only possess an understanding of its underpinnings but will also glean the reasoning behind each decision. Concepts are introduced gradually and are illustrated with simplified examples. Where appropriate, predictions and corollaries of the framework are discussed. Strictly speaking, one need not have read the dissertation to be able to make use of the applications therein. For those wishing to skip ahead, the formulae comprising the framework are listed in Appendix A.

The remainder of this chapter is split into three sections. Section 1.1 serves as an overview of previous quantitative approaches in historical linguistics, surmising that no similar approach has been proposed in the past. Section 1.2 introduces the theoretical backbone of Wordlist Distortion Theory. Section 1.3 summarizes the findings of subsequent chapters.

1.1 Computational Historical Linguistics Background

Since the advent of computers, numerous computational tools have been developed in the aid of historical linguistics. Broadly speaking, these tools fall into two categories: those that automate some component of the comparative method, and those that bypass the comparative method entirely to infer aspects of the languages’ prehistory, e.g. relatedness, time-depth, homeland, etc. However, as this section will show, neither of these avenues of research has yet produced a technique for the quantitative evaluation of the comparative reconstruction process itself.

The task of comparative reconstruction can be loosely described as comprising a number of discrete steps (Arlotto, 1981; Cambell, 2013; Jäger, 2019), e.g. collecting cognates, establishing sound correspondences, reconstructing proto-forms, determining innovations, etc. However, in practice, at least some of these steps are conducted simultaneously. For instance, once a sound change has been identified, it can be used to identify new cognates, which in turn may lead to the discovery of novel sound correspondences and, therefore, a new sound change. As a result, traditional manual reconstruction is often reliant on the intuitive judgment and creativity of the researcher, so much so that a complete automation of the process has remained elusive (Jäger, 2019, Wu et al., 2020).

Nevertheless, most of the individual steps during the reconstruction process have received attention in the computational historical linguistics literature. For example, rather than sifting through potential cognates manually, researchers can make use of one of the available automated cognate detection tools (Kay, 1964; Oakes, 2000; Kondrak, 2002; Rama et al., 2017). Once the cognate lists have been assembled, segment correspondences between them can be also detected automatically (Kondrak, 2009; List, 2019).

A prerequisite for the effective detection of sound correspondences is knowing which segments in a word pair should be compared, also known as the alignment problem. For instance, it may seem obvious that in the comparison of Romanian [pjatrə] ‘stone’ and Italian [pjɛtra] ‘stone’ that [p] corresponds to [p], [j] to [j], etc., concluding with the correspondence between [ə] and [a]. However, it is less obvious what the corresponding segments are between Romanian [pet͡ʃor] ‘foot’ and Italian [pjɛde] ‘foot’ and even less obvious what they are between Romanian [kɨine] ‘dog’ and French [ʃjẽ] ‘dog’. Thankfully, here too computational historical research can come to the rescue in the form of alignment algorithms, tools for finding the most fruitful segmental correspondences for the word pair (Covington, 1996; Kondrak, 2002; List, 2014).

Once a full list of sound changes in a reconstruction is developed, and the proto-forms inferred, there exist tools for checking that the combination of the two produces the correct output (Lowe & Mazaudon, 1994; Marr & Mortensen, 2023). Because a full manual derivation by hand is typically not feasible, it can be useful to feed the proto-forms through software aimed to apply the sound changes in question while also providing diagnostic tools, e.g. the mean phonetic distance. Thus, it was with the help of such software that the voicing of [k] in the history of French, previously thought of as sporadic, was identified as epiphenomenal given a careful reordering of sound changes in the reconstruction (Marr & Mortensen, 2023).

In short, although the researcher is still expected to rely on experience and intuition at many junctures in the comparative reconstruction process, many of the laborious components of the method have been expedited with automation. Nevertheless, there still exists no tool for the evaluation of individual instances of comparative reconstruction. As such, although the researcher is helped at all steps by advances in computation historical linguistics, when they are required to make a judgment call of their own, no formal or quantitative metric to evaluate their decision is available.

A different avenue of research in computational historical linguistics is concerned with accelerating the process of language comparison by bypassing the comparative method altogether and instead focusing on ways to compute proxies of genetic relatedness. The general consensus in the field appears to be that these quantitative alternatives can be useful in cases where a manual comparative reconstruction is not feasible, though the comparative method still constitutes the gold standard when it comes to establishing language relatedness (McMahon & McMahon, 2003; Bostoen, 2007; Downey et al., 2008; Kiparsky, 2015).

The earliest example of a quantitative alternative to the comparative method is lexicostatistics (Swadesh, 1955), where the proportion of apparent shared cognates is used as a stand-in for genetic proximity. The assumption behind lexicostatistics is that *ceteris paribus* closely related languages should share a greater proportion of cognates than distantly related ones. This can be explained by the fact that the ancestor of closely related languages is more recent than that of distantly related ones. So, closely related languages have had less time to undergo independent instances of lexical replacement. Thus, by tallying the proportion of shared cognates between related languages in a table, one can get a sense of the phylogenetic structure of the family. The additional assumption that the proportion of apparent cognates is correlated with the time-depth of the proto-language split is known as *glottochronology*. In practice, the two notions are closely intertwined, and early works on the topic (Swadesh 1955) suggest that the primary purpose of lexicostatistics was not to replace comparative reconstruction, but rather to equip it with a tool for estimating time-depth.

This avenue of research in historical linguistics was substantially bolstered by the development of Bayesian phylogenetic analysis in the field of molecular evolutionary biology (Rannala & Yang, 1996; Mau et al., 2004). In molecular phylogenetics, the technique is used to infer evolutionary trees of species based on the nucleotides of their DNA sequences. Since references to similarities between evolutionary biology and historical linguistics have begun as early as at least Darwin's *Descent of Man* (Pagel, 2017) and continue to the modern age (Atkinson & Gray, 2005; Croft, 2008), it is perhaps of no surprise that the use of statistical techniques imported from molecular biology is actively encouraged in historical linguistics (McMahon & McMahon, 2003). Note, however, that the two fields also exhibit important differences (List et al., 2016).

Bayesian phylogenetic tools aid in the creation of probabilistically justified language trees based on tables of shared elements. Similar to lexicostatistics, these are often cognate tables, most commonly wordlists of 100-200 core vocabulary, a sample which has been shown to sufficiently reflect a language’s phonological patterns (Zhang & Gong, 2016). However, in principle a phylogenetic tree can be inferred based on any kind of data. Core vocabulary, e.g. a Swadesh list, does easily lend itself to this methodology, but so do syntactic features (Longobardi et al., 2013) or typological features (Sicoli & Holton, 2014). For sign languages, comparison of manual alphabets, i.e. gestures for conveying letters, has also been performed (Power et al., 2020). No matter the data used, the central assumption of such methodology is that closely related languages should on average share a greater number of elements than distantly related ones. Nevertheless, because core vocabulary exhibits a slower rate of replacement than abstract phonological or syntactic features (Greenhill et al., 2017), this is the type of data used in most studies to this day.

Naturally, phylogenetic Bayesian methods have been used to argue for the legitimacy of language families and for internal subgrouping. Some research using Bayesian phylogenetics has confirmed language relationships discovered through comparative reconstruction, such as for the Indo-European family (Rexova et al., 2002; Longobardi et al., 2013; Chang et al., 2015, Greenhill et al., 2023) or the Austronesian family (Greenhill et al., 2017). Other research has used the methodology to argue for relationships not widely accepted in the field, such as for the Trans-Eurasian language family, a macro-family with the controversial Altaic family as one of its branches (Robeets et al., 2021).

Even in cases where phylogenetic research and comparative reconstruction are in agreement about the lack or presence of a genetic link, the two may still disagree when it comes to the internal structure of the family in question. For example, a phylogenetic analysis conducted on syntactic features of Indo-European languages finds no close affinity between Bulgarian and South Slavic or between Indic and Iranian (Longobardi et al., 2013), although this view is not shared by any traditional source on Indo-European linguistics. Likewise, a phylogenetic analysis conducted on typological features of Dené-Yeniseian languages shows that typological diversity is greater within the Dené languages than between the Yeniseian languages and certain branches of the Dené languages, which the authors interpret as evidence for migration out of Beringia into Asia (Sicoli & Holton, 2014). Yet, the original Dené-Yeniseian proposal (Vajda, 2010), motivated by the comparative method and supported by genetic studies (Rubicz et al., 2002) as well as comparative mythology (Berezkin, 2019), comes to the opposite conclusion.

Bayesian phylogenetic methods can also be used to answer questions not directly addressed by the comparative method. In the spirit of glottochronology, it is common to couple phylogenetic trees with estimates of proto-language splits, which itself is often used to narrow down the *urheimat*, or homeland. In this way, the dispersal of the Uto-Aztecan language family has been dated to approximately 4.1kya out of Southern California (Greenhill et al., 2023). The dispersal of Trans-Eurasian has been dated to approximately 9.2kya out of Northern China (Robeets et al., 2021). Finally, the dispersal of Indo-European has been dated to around 8.7kya out of Anatolia (Gray & Atkinson, 2003), around 8.1kya out of the southern Caucasus (Heggarty et al., 2023), or around 6.3kya out of the Pontic-Caspian steppe (Chang et al., 2015), depending on the details of the model used.

It should be acknowledged that phylogenetic studies in historical linguistics are not entirely divorced from traditional comparative reconstruction. At long genetic distances, it can be extremely difficult to tell whether two words across wordlists are cognate, c.f. cognates English [sɑŋ] ‘song’ and Ancient Greek ὀμφή [ompʰeː] ‘voice’. As such, on one hand, some studies employing Bayesian phylogenetic gather cognate sets through traditional (i.e. manual) comparative methodology (Bowern & Atkinson, 2012; Chang et al., 2015; Greenhill et al., 2017; Greenhill et al., 2023). However, on the other hand, other studies have argued that cognate sets can be assembled automatically using string similarity (Zhang & Gong, 2016; Rama & List, 2019).

In parallel to methods that estimate genetic relatedness among languages through the proportion of shared elements, e.g. cognates, features, etc., there also exist frameworks that do so by computing the pairwise phonetic similarity between wordlists. The origin of these methods, usually referred to as *mass comparison* or *multilateral comparison*, can be traced to the second half of the 20th century (Greenberg, 1987). The intuition behind multilateral comparison is that a pair of cognates between closely related languages *ceteris paribus* should be more phonetically similar than a pair of cognates between distantly related languages. This can be explained by the fact that, due to more recent shared ancestry, closely related languages have had less time to undergo independent instances of sound change.

In its earliest instantiations, multilateral comparison was typically conducted qualitatively and informally. The approximate proto-forms, segmental inventory of the proto-language, and sound correspondences within the family were presented, but arguments from probability or descriptions of language change were lacking. For example, a multilateral comparison argument for a Proto-World language offers descendants in attested languages for the Proto-World form *BUR ‘ashes, dust’ in Arabic [baraj] ‘dust, soil’, Finnish [poro] ‘hot ashes, course dust’, Kazakh [bor] ‘chalk’, among others (Ruhlen, 1994), though these choices appear to be motivated by perceived phonetic and semantic similarity only.

It should be acknowledged that the original purpose of multilateral comparison was to apply it to possibly related languages to determine if a comparative reconstruction should be conducted and, if so, between which pairs of languages (Greenberg, 1987). As such, the methodology was not intended to replace the comparative method or make strong claims about language descent or relatedness. Rather, it served as a technique to steer more methodologically rigorous approaches in the right direction.

Nevertheless, part of the critique of multilateral comparison and its more informal applications resulted in the formulation of probabilistic arguments in the multilateral comparison framework. Numerous research was devoted to estimating the number of expected phonetic matches or near matches, given the size and number of wordlists and the strictness of the semantic similarity (Nichols & Peterson, 1992; Baxter, 1995; Ringe, 1999). Thereafter, dedicated quantitative multilateral comparison algorithms were developed (Baxter & Ramer, 2000; Kessler & Lehtonen, 2006), though these were already prefigured in earlier work (Oswalt, 1970).

Significance testing in multilateral comparison is typically done through Monte-Carlo simulation. After the phonetic distance is calculated for each pair of (apparent) cognates, a mean phonetic distance score is calculated for each language pair. The wordlists in question are then shuffled and the phonetic distance measure is calculated anew. The likelihood of witnessing the original phonetic distance, or one shorter, in a random wordlist is estimated from the proportion of permutations that yield the same phonetic distance score or shorter. Comparison is, therefore, only possible between pairs of wordlists, but a larger family can be analyzed simply by performing multiple analyses.

At the heart of multilateral comparison lies the particular phonetic similarity metric employed. There is no consensus on what constitutes the best method for phonological string comparison, with some suggesting multivalued articulatory phonetic features (Kondrak, 2003) and others employing basic edit distance (Brown et al., 2009; Holman et al., 2011). So as to make comparisons across typologically disparate languages simpler, some approaches lump segments together by major place of articulation (Kessler & Lehtonen, 2006; Ceolin, 2019), while others address the issue by discarding everything except the first consonant (Kassian et al., 2015).

Nowadays, just as phylogenetic Bayesian analysis, methods of multilateral comparison are used as an argument for language classification and subgrouping. Often, these methods yield results largely compatible with those discovered through traditional methodology. For example, an extensive multilateral comparison study of the world’s languages found agreement between genetic relations found through this methodology and those assumed given traditional analyses (Pompei et al., 2011). In the same vein, given that the debate surrounding the Altaic hypothesis, it is perhaps not surprising that the results of a multilateral comparison analysis of three of its branches – Mongolic, Turkic, and Tungusic – were inconclusive (Ceolin, 2019). Nevertheless, the results of a multilateral comparison can also contradict conclusion reached through traditional means or, indeed, results of other multilateral comparisons. Thus, one multilateral comparison study of the Proto-Indo-European-Uralic hypothesis, not widely accepted in the field, has come away with evidence for the grouping (Kassian et al., 2015), while a previous one found no such evidence (Kessler & Lehtonen, 2006).

Although not developed expressly for this purpose, multilateral comparison, like phylogenetic Bayesian methodology, can also be used to estimate the time depth of a language split (Holman et al., 2011). If geographic data is available, multilateral comparison can even be used to estimate a language family’s *urheimat*, or homeland, modelling migrations as random walks (Wichmann et al., 2010). Finally, it should be noted that, although presented separately here, phylogenetic Bayesian methods and multilateral comparison are not mutually exclusive. The mean pairwise phonetic distance between wordlists, as calculated through multilateral comparison, can serve also as input into a phylogenetic model (Pompei et al., 2011).

Both phylogenetic Bayesian analysis and multilateral comparison are useful techniques in computational historical linguistics as they can offer some of the same insights as the comparative method without necessitating a detailed description of the diachronic changes or a nuanced understanding of the languages in question. Note, however, that these techniques are largely independent of the comparative method.

The comparative method is primarily concerned with assembling sound correspondences in order to uncover the phonological history of the language family. Typically, this is done with the neogrammarian hypothesis in mind, the idea that sound change is exceptionless for all words in the language containing the relevant segments in the relevant environment (Hull, 2014). Meanwhile, unlike the comparative method, lexicostatistics does not engage with sound change in any capacity. Furthermore, because molecular biology does not have an analogue to regular sound change (List et al., 2016), mathematical models developed to address problems in this field, such as phylogenetic Bayesian inference, are not equipped to handle this aspect of language change at all. Although multilateral comparison engages closely with the segments of the languages in question, the fact that the phonetic distance score for each word pair is calculated independently means that no notion of regularity across the wordlist can be enforced. As a result, it is not surprising that statistical techniques for inferring genetic relatedness can give results conflicting with those reached through traditional comparative reconstruction.

Therefore, once a reconstruction is undertaken, although there exists a plethora of tools to expedite individual processes or to estimate relatedness, there is no quantitative technique of evaluating the reconstruction itself. In decisive moments of the comparative method, the researcher is expected to rely on skill and intuition. Furthermore, upon completion of the project, other researchers are expected to rely on their skill and intuition to evaluate the reconstruction in question. The goal of the quantitative evaluation of comparative reconstructions presented in this dissertation is to provide a diagnostic tool for the researcher conducting a comparative reconstruction, as well as for the researchers evaluating it.

1.2 Method

Wordlist Distortion Theory, henceforth WDT, differs greatly from previous quantitative approaches in historical linguistics. In contrast to lexicostatistics or multilateral comparison, WDT is applied in unison with the comparative method. It cannot give a result contrary to the reconstruction in question, as it is not an alternative to comparative reconstruction but merely an evaluation of it.

As input, Wordlist Distortion Theory takes the derivation from one wordlist, the *mother*, into another, the *daughter*, as a series of diachronic processes, or distortions. The framework evaluates the distortions themselves by estimating the likelihood of observing evidence of the same distortions in a random wordlist. In contrast to computational alternatives to the comparative method, such as lexicostatistics or multilateral comparison, WDT is primarily concerned with diachrony. Rather than using similarity as a proxy for genetic relatedness, e.g. proportion of shared cognates or phonetic distance, the framework estimates the reliability of the series of changes posited in the reconstruction by estimating the likelihood that a random wordlist merits the same number and type of changes.

For the purposes of the framework, just as for multilateral comparison, wordlists are ordered pairs of phonological forms. For example, in (1), Latin serves as the mother wordlist and Romanian as the daughter. The first entry in both wordlists corresponds to the word for ‘goat’ in both languages, the second for the word for ‘tomato’, and so on.

(1) *Sample Wordlists*

gloss	*Latin*	*Romanian*
goat	kapra	kaprə
tomato	--	pətləʒikə
spring	veːra	primavarə
summer	ae̯stas	varə

There is no requirement that every position in every wordlist must be filled. Because Latin did not have a word for ‘tomato’, this entry can remain blank in (1). It should be stressed that, to avoid cherry-picking, for this method, or any quantitative metric (Ringe, 1995), the wordlists must not be *ad hoc*. Beyond that, any wordlist selection criteria are acceptable. The most obvious choice is to use a Swadesh list (Swadesh, 1955) or some other preselected list of meanings, or even all of the available words in a language.

In WDT, a reconstruction is defined as a series of transformations, which distort the original mother wordlist into the daughter wordlist. The reconstruction must be exhaustive; if the mother wordlist is taken as input to the transformations, the output must be the daughter wordlist. A possible reconstruction for the data in (1) is given in (2).

(2) *Reconstruction from Latin to Romanian*

(a) *a > ə / —#
(b) *eː > a
(c) [3] > [4]
(d) ∅ > [2]
(e) ∅ > [3]

In WDT, just as in any complete historical account of a language deduced through traditional means, a reconstruction can, and usually must, combine different types of transformations. These transformations can distort the original wordlist in different ways. Some transformations, most notably sound change, are regular and apply to all entries in the wordlist, while others are sporadic and apply to individual entries only. Some transformations alter the phonological content of each entry. For example, change (a) in (2) affects both Latin [kapra] and [veːra] yielding Romanian [kaprə] and [varə] respectively. Other transformations, such as semantic change, are effectively blind to the phonological content. For example, change (c) in (2) maps the 3rd position in the wordlist onto the 4th position, describing the shift in meaning from Latin 'spring' to Romanian 'summer'. Thus, the likelihood of this change and its trajectory do not dependent on the phonological characteristics of Latin [veːra] 'spring'. The same can be said of the two instances of lexical replacement in (d) and (e), introducing Romanian [pətləʒikə] 'tomato' and [primavarə] 'spring'.

One of the strengths of WDT is that it can include any type of distortion on the wordlist, so long as the distortions are well-defined and incorporated into the mathematical backbone of the theory. However, in the interest of presenting the simplest working model, this dissertation focuses only on the diachronic transformations most commonly discussed in the literature: sound change, semantic change, and lexical replacement. Nevertheless, in principle, this methodology is compatible with other types of diachronic transformations, such as morphological change, sporadic change, and analogy.

Unlike in most other quantitative techniques developed in computational historical linguistics, such as Bayesian phylogenetics and multilateral comparison, comparison in WDT is conducted between mother and daughter rather than sister wordlists. While multilateral comparison and phylogenetic analysis strictly speaking do not presuppose a comparative reconstruction, WDT can only be applied after the comparative method or in unison with it. Since the methodology is limited to mother-daughter comparisons, the analysis of a language family requires multiple comparisons, just as in multilateral comparison.

It should be stressed that the aim of WDT is the evaluation of the reconstruction, not the dataset. This is perhaps the largest difference between WDT and frameworks such as phylogenetic Bayesian analysis or multilateral comparison. In WDT, the same dataset reconstructed in different ways through the comparative method yields different results. For the purposes of WDT, the existence of the proto-language, in the exact shape necessitated by the reconstruction, is axiomatic. The only thing that is evaluated is the likelihood of observing the distortions required by the proto-language in random data. This is, in fact, in line with the general attitude in historical linguistics. A comparative reconstruction does not prove the existence of a proto-language; it simply shows how the proto-language would explain attested languages if it had existed (Nichols, 1996).

At its core, Wordlist Distortion Theory operates on standard assumptions in hypothesis testing and probability. No new statistical or mathematical concepts are required by the framework. The null hypothesis is that the daughter wordlists was generated randomly in accordance with the synchronic properties of the language. The alternative hypothesis is that the daughter wordlist was derived from the mother wordlist through the reconstruction posited. WDT weighs the null and alternative hypotheses against each other by estimating the likelihood of observing evidence for the alternative hypothesis in random data. If this likelihood is below an established threshold, the null hypothesis is rejected, and the reconstruction is deemed reliable. Otherwise, the null hypothesis is kept, and the reconstruction is not deemed reliable.

The likelihood that a random wordlist would evidence the same number and type of distortions from the mother wordlist as in the reconstruction is estimated theoretically. An experimental approach to the same problem is conceivable. For example, one could generate wordlists at random and, for each random wordlist, calculate the number and type of distortions required to derive it from the mother. However, such approaches are intractable in practice. The biggest obstacle is that no algorithm exists for determining the number and type of transformations required to derive one wordlist from another. In fact, if such an algorithm were to exist, then it could simply replace the comparative method. WDT does mirror the experimental sampling approach, in that it estimates the likelihood of generating a random wordlist which can be derived from the mother in the same number and type of changes as in the reconstruction. In other words, in theory, the results of experimental sampling should approach the results of a WDT calculation as the size of the sample increases.

To model the likelihood that a random wordlist would evidence the same number and type of distortions from the mother as in the reconstruction, Wordlist Distortion Theory employs two theoretical sets of wordlists. The first, the *synchronically compatible* set, includes all wordlists that share the phonological properties of the daughter wordlist. This set constrains what is considered random. The second, the *diachronically local* set, includes all wordlists that are derived from the mother through the same number and type of distortions as in the reconstruction. This set includes all wordlists exhibiting evidence of the same number and type of distortions as evidenced in the daughter. The following two sections describe synchronically compatible and diachronically local sets in more detail.

1.2.1 Synchronically Compatible Set

A successful probabilistic evaluation of comparative reconstruction has no recourse but to compare the product of language to the product of random chance. This notion has been hinted at as far back as the 18th century in the famous "stronger affinity […] than could possibly have been produced by accident" (Jones, 1786). However, it is well known that the chance resemblance between typologically similar languages is more likely than between dissimilar ones. Therefore, identifying what could have been produced by accident is not a trivial task and requires a keen understanding of the languages analyzed (Ringe, 1999; Kessler, 2015).

In effect, the synchronic properties of the languages in question must be factored out from the analysis. For example, to evaluate the reliability of the reconstruction in (2) on the Latin and Romanian wordlists in (1), one must make a provision for all possible but non-existent versions of Romanian; one must, at least implicitly, assess the link between all unattested but conceivable versions of Romanian wordlists and the Latin wordlist. In this sense, a probabilistic analysis must by necessity engage with the counterfactual.

Recall that in multilateral comparison studies randomness is approximated using a Monte-Carlo simulation (Baxter, 1996; Kessler & Lehtonen, 2006; Croft, 2008; Kessler, 2008; Kessler, 2015;

Hruschka et al., 2015; Kassian et al., 2015; Zhang & Gong, 2016). The simulation generates random wordlists by drawing from the entries of the original at random. For instance, a Monte-Carlo simulation on the Romanian wordlist in (1), could yield one where the entry corresponding to ‘tomato’ is [kaprə] (‘goat’ in the original), or [varə] (‘summer’ in the original), or [primavarə] (‘spring’ in the original), or indeed the faithful [pətləʒikə].

For the purposes of simulating random wordlists, Monte-Carlo methods have the advantage of not yielding undesirable ones, e.g. those that could not occur in the language due to phonotactic constraints. After all, the entries in random wordlist generated by Monte-Carlo simulation are taken from the original and are, therefore, grammatical by definition. Onomatopoeia and recurrent polysemy aside, form and meaning in language are generally considered independent, i.e. Saussarian arbitrariness. Therefore, one would expect each phonological form to be permissable in each position in the wordlist.

Nevertheless, such methods have the disadvantage of not adequately exploring the space of possible wordlists. Wordlists containing grammatical but unattested strings cannot be reached through a Monte-Carlo simulation. In fact, as is discussed throughout this dissertation, the number of possible words is often larger than the number of attested words by at least a few orders of magnitude. As a result, Monte-Carlo simulations explore only a tiny subset of phonologically constrained randomness. Chapter 3 explores why this is an issue for WDT in particular.

Therefore, a different approach to randomness is employed in this dissertation. A reconstruction from a mother wordlist to a daughter wordlist is evaluated with respect to the abstract set of wordlists *synchronically compatible* with the daughter. Given a wordlist *A*, wordlist *B* is defined to be synchronically compatible with *A* if and only if every word in *B* is phonologically permissable in *A*.[1] The synchronically compatible set contains all such wordlists. Stated

[1] Technically speaking, the synchronic compatibility relation is not symmetric. There may exist a wordlist *X* comprising forms permissable in the daughter but also utilizing a smaller phonological inventory or exhibiting phonotactic restrictions not found in the daughter. In such a scenario, all forms in *X* would be permissable in the daughter but not vice versa. In other words, *X* would be synchronically compatible with the daughter, but the daughter not synchronically compatible with *X*. However, note that, for the purposes of Wordlist Distortion Theory, synchronic compatibility only needs to be checked with respect to the daughter wordlist. As such, the question of symmetric synchronic compatibility loses its relevance.

informally, a wordlist synchronically compatible with the daughter must only contain entries that are grammatical in the daughter; thus, the synchronically equivalent set contains all wordlists comprising words grammatical in the daughter.

As an example, imagine a wordlist of English containing the entry [spajdəɹ] ‘spider’. The word contains the segments [p] and [ɹ] and is two syllables long. However, none of these properties need to hold true for the same entry in a synchronically compatible wordlist. In a synchronically compatible wordlist, the entry for ‘spider’ could be occupied with [stajdəɹ], or [spajdəl], or [spajd], or indeed [wɛbstəɹ] since all of these too are phonologically licit strings in the language.

However, the words in a synchronically compatible wordlist cannot vary indefinitely. For example, in a wordlist synchronically compatible with a Modern English wordlist, [spajdəɹ] cannot be replaced by [aɾaɲa] or [χəlməχən], since these contain segments not found in English. Similarly, the word cannot be replaced by [ɹtilə] or [gɛzg] because, although these comprise English segments, their order is illicit in the language, i.e. these words constitute a phonotactic violation. By restricting the synchronically compatible set in this way, one can begin to discuss the probability of observing one phenomenon or another in a language with particular synchronic properties, e.g. English.

The question of which synchronic properties need to be controlled for is independently interesting and receives only a cursory discussion in this dissertation. WDT is compatible with any definition of the synchronically compatible set. However, the two most obvious factors to control for in any analysis are segmental inventory and phonotactics. It makes little sense to evaluate the probability of observing evidence for a particular distortion in a wordlist without considering the segments and syllable structure of the language.

Of course, many other factors can be included in the definition of the synchronically compatible set, such as long-distance harmony, prosody, or neutralization. Definitions of phonological complexity differ depending on which factors are considered (see, for example, Pimentel et al., 2020). We will put this discussion on hold for the remainder of the chapter, with the promise to return to it in Chapters 2 and 3.

1.2.2 Diachronically Local Set

Wordlist Distortion Theory estimates how exceptional the properties of the daughter are with respect to the mother and the reconstruction in question. This estimate comes from the likelihood of observing evidence for the distortion required in the derivation of the daughter from the mother. A wordlist is considered exceptional in this context if it exhibits evidence of the same number and type of distortions as in the reconstruction. As discussed in the previous section, it only makes sense to consider exceptionality with reference to the synchronic properties of the daughter, i.e. the synchronically compatible set. In effect, a probabilistic analysis of comparative reconstruction requires some notion of diachronic distance from the mother wordlist.

In this dissertation, the distance of the daughter wordlist from the mother is defined in the set of wordlists *diachronically local* to the daughter. Given a daughter wordlist *D*, its mother wordlist *M*, and *R*, a reconstruction from *M* to *D*, wordlist *A* is diachronically local to *D* if and only if a different reconstruction *Q* from *M* to *A* exists such that, for every type of distortion in *R*, the number of distortions of that type in *Q* is equal to or small than the number of distortions of that type in *R*. Stated informally, a diachronically local wordlist is one that can be derived from the mother wordlist through the same type and number (or fewer) transformations as required for the daughter wordlist in the reconstruction. If the reconstruction in question is efficient, i.e. posits as few transformations as possible, the set of diachronically local wordlists is the smallest possible such set which also includes the daughter wordlist.

Diachronically local wordlists are constrained by the number and type of transformations posited in the reconstruction but not by the elements in those transformations. The variables in the transformations are deemed *ad hoc* and arbitrary from a probabilistic perspective. For example, in the history of the English language, an intervocalic [d] became [ð], hence OE [fæder] > MnE [fɑðəɹ]. The reconstruction posits the change [d] > [ð] simply because this is what accounts for the data. In principle, the output of the change could have been something else, for example [t], as was the case in German (where the reflex of the same word is *Vater),* or [ɖ], or even [ə]. In this way, we can say that the transformations required for a reconstruction to the daughter imply numerous alternate transformations, which in turn derive diachronically local wordlists.

In practice, some outcomes of diachronic transformations appear more likely than others. For instance, [d] > [t] is phonologically more natural than [d] > [ə]. Therefore, it can be argued that reflexes of Old English [fæder] ‘father’ that exhibit the former change, e.g. [fæter], are less distant from the mother than those that exhibit the latter, e.g. [fæəer]. Much ink has been spilled discussing the mechanics of unnatural sound change (e.g. Picard, 1990; Beguš, 2018), but the discussion is somewhat tangential to this dissertation. WDT is compatible with restrictions on the diachronically local set which go beyond simply the numbers and types of transformations, such as those that treat [d] > [t] and [d] > [ə] differently. For example, Section 3.5.2 discusses a method for incorporating phonological features into the framework. However, for the sake of presenting the simplest working model, it assumed throughout that all outputs of diachronic transformations are equally likely. Thus, the sound changes [d] > [t] and [d] > [ə] are assumed to produce diachronically local wordlists.

A given wordlist always has a well-defined set of synchronically compatible wordlists since, presumably, it always exhibits some type of noteworthy synchronic properties. However, because diachronic locality is evaluated with respect to a reconstruction, a wordlist only has a well-defined set of diachronically local wordlists if a reconstruction is proposed. Given a particular reconstruction, say from Old English to English, the word [fɑðəɹ] may correspond to [fɑtəɹ] in a diachronically local wordlist, derived from OE [fædəɹ] through a different set of regular sound changes. Given a particularly large reconstruction, one that posits many changes, the word [fɑðəɹ] may correspond to something very different, such as [ətk]. Diachronically local wordlists need not be synchronically compatible, as two wordlists may be derived from the same mother wordlists but exhibit different phonological structure.

Whereas the method for estimating the number of synchronically compatible wordlists is the same for all daughter wordlists, the method of counting diachronically local wordlists differs depending on the type of transformation. The bulk of this dissertation is devoted to estimating the number of diachronically local wordlists implied by the basic types of diachronic transformations: lexical replacement, semantic change, and sound change. Each diachronic transformation requires different mathematical machinery, and each will be addressed separately.

1.2.3 General Formula

The greater the number of transformations required by a reconstruction, the larger the set of diachronically local wordlists. As the set of diachronically local wordlists grows, the reconstruction becomes less and less reliable, because a reconstruction of the same magnitude can derive more and more synchronically compatible wordlists. At a certain point, the proposed reconstruction does not bode much better than chance, as the set of diachronically local wordlists contains, along with the daughter, almost any synchronically compatible wordlist. In such a situation, a wordlist generated randomly in accordance with the phonological principles of the daughter is likely to evidence a reconstruction of similar size as the one to the daughter.

To evaluate the hypothesis that the similarities between the mother and daughter wordlists are attributable to chance (the null hypothesis), one needs to evaluate the likelihood that a randomly generated wordlist is as or more easily derivable from the mother wordlist than the daughter. In WDT, this is the probability that a wordlist synchronically compatible with the daughter is also diachronically local to the daughter given the reconstruction, or, in other words, the proportion of diachronically local wordlists in the set of synchronically compatible wordlists. For example, if that proportion were 0.5, i.e. if half of synchronically compatible wordlists are also diachronically local, a wordlist that shares the phonological properties of the daughter would merit a reconstruction of equal or lesser magnitude 50% of the time.

The formula for calculating the probability of a randomly substantiated reconstruction is given in (5). We define S as the set of wordlists synchronically compatible with the daughter and D as the set of wordlists diachronically local to the daughter given the proposed reconstruction. By the formula for conditional probability, the likelihood that a member of S is also a member of D is equal to the cardinality (size) of the intersection of the two sets divided by the cardinality of S. The resulting measure, $P(D|S)$, is used in this dissertation as a proxy for reconstruction reliability. Less reliable reconstructions are more likely to be substantiated by chance and exhibit high $P(D|S)$ values. More reliable reconstructions are less likely to be substantiated by chance and exhibit low $P(D|S)$ values.

(3)

$$P(D|S) = \frac{|D \cap S|}{|S|}$$

General Formula: Formula for calculating the probability that a wordlist generated at random in accordance with the phonological properties of the daughter wordlist evidences the same number and type of transformations as those in the reconstruction. *S* is the set of wordlists synchronically compatible with the daughter. *D* is the set of wordlists diachronically local to the daughter given the proposed reconstruction.

The General Formula shares its motivation with many methods of statistical hypothesis testing. The intention is to determine the likelihood that a random measurement yields an equally or more extreme result than the observed measurement. In the case of a t-test, this probability is known as the *p*-value. In the current framework, we refer to the probability as $P(D|S)$, which more closely reflects the underlying calculation, while also avoiding association with random sampling from a normal distribution. Nevertheless, the motivation behind $P(D|S)$ is very similar to the motivation behind *p*-values in statistics. In fact, in some cases, the two can be shown to be equivalent, as is demonstrated in Appendix B.

The General Formula forms the backbone of Wordlist Distortion Theory. The remaining chapters are largely concerned with ways of estimating the size of *S* and its intersection with *D*, given particular kinds of transformations commonly posited in comparative reconstruction. These estimates require numerous assumptions about language structure and language change, which are addressed throughout the dissertation. However, none of the assumptions are intrinsic to the theory itself. Different instantiations of WDT may involve other ways of calculating *S* and its intersection with *D*, arriving at a different final answer. It is only the General Formula that is indispensable to the project.

It is perhaps a good time to point out that the purpose of WDT is not to faithfully model diachronic change. Rather, the purpose is to estimate the likelihood of observing evidence of particular types of diachronic change in the data. For instance, given a reconstruction of Modern English from Old English, WDT cannot be used to estimate the likelihood that Modern English would have arisen from Old English or the probability of each individual change.[2] Instead, WDT can be used to estimate the likelihood that evidence of the same type and number (or fewer) as required in the reconstruction could be observed in phonologically constrained randomness, i.e. the likelihood that the reconstruction is spurious.

One may conceptualize the General Formula abstractly. Imagine a high-dimensional space of wordlists, approximated two-dimensionally in Figure 1. The position of each wordlist is determined by the values (segments) in each entry. Thus, distance between similar wordlists is shorter, and distance between dissimilar wordlists longer. A reconstruction starts with the point corresponding to the mother wordlist. Each transformation in the reconstruction alters the mother wordlist in some way, translating the starting point to some new point in the space. Alternative transformations of the same type may alter the wordlist to the same extent but in a different way, which can be thought of as a translation of the same distance but in a different direction. In this space, a reconstruction is a (not necessarily straight) path from the mother to the daughter. The set of diachronically local wordlists is a high-dimensional sphere with the mother wordlist at its center. With each new transformation the sphere grows to include the newest intermediate wordlist. The reconstruction stops when the latest layer of the sphere contains the daughter wordlist. At that point, the sphere corresponds to the entire set D, and the portion of the sphere containing synchronically compatible wordlists corresponds to $|D \cap S|$.

[2] If this value were to be estimated, it is not clear how one should interpret it. Presumably, the likelihood of events conspiring to yield any attested language in its exact shape is quite small, but this has little significance in hypothesis testing.

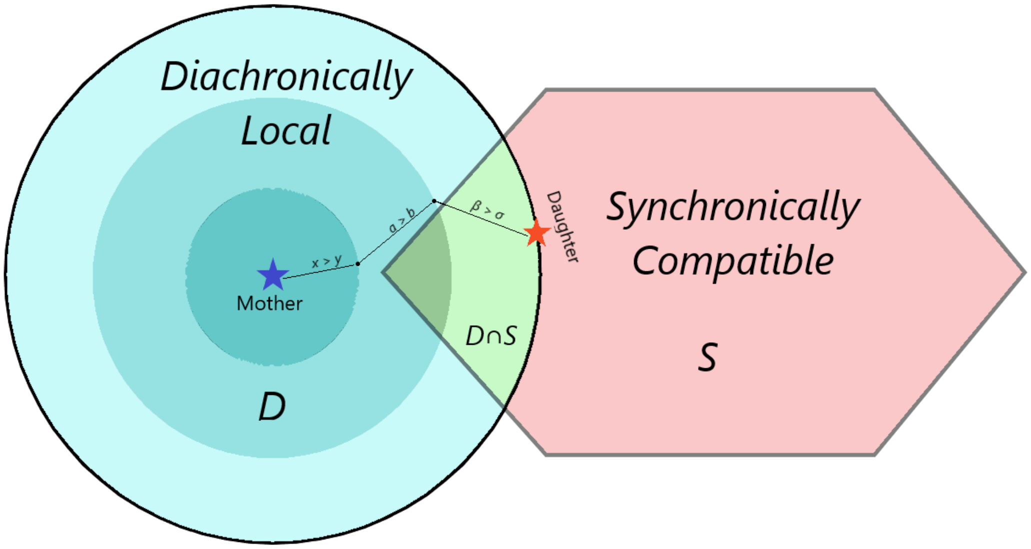
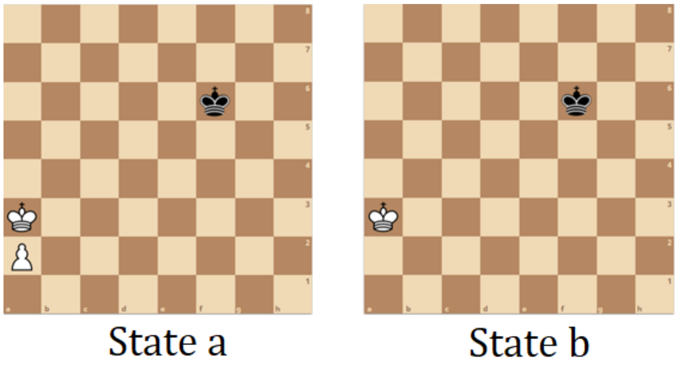


Figure 1: *Abstract representation of the General Formula*. The reconstruction is marked as a path of transformations from the mother (red star) to the daughter (blue star) wordlist. Each transformation introduces an additional layer of diachronically local wordlists, until the newest layer envelops the daughter wordlist.

1.3 Non-linguistic Applications

Despite the name, the backbone of Wordlist Distortion Theory is not exclusive to linguistics. Put simply, the framework estimates the likelihood that a state generated randomly in accordance with some restrictions can be derived through an equal or lesser number of distortions as some other state. This type of reasoning can perhaps be useful in other domains as well, so long as there exists some notion of states (e.g., wordlists in historical linguistics), which can be distorted into other states. If a corresponding definition of the synchronically compatible set and its intersection with the diachronically local set can be found in these other domains, then a WDT-like approach can be used to calculate the likelihood that a randomly generated state evidences the same number and type of distortions as those posited in the ‘reconstruction’. This section briefly surveys some other fields where reasoning similar to WDT may be fruitful. While tangential to the topic at hand, this discussion is still be useful to linguists, as it serves to illustrate the logic behind the General Formula in familiar environments.

The game of chess is another domain where the logic of WDT can be meaningfully applied. For an illustration, observe two chess board-states in Figure 2. Just as in historical linguistics, one might pose the question about the likelihood that state (b) is derived from state (a) through some number of distortions, i.e. chess moves. This is the same as asking the likelihood that (b) occurred after (a) in the same chess game.


State a

State b

Figure 2: *Two chess board states*. White to move in both.

As in historical linguistics there are two approaches to solving this problem. A surface-oriented approach, loosely corresponding to lexicostatistics or multilateral comparison, might calculate a heuristic similarity measure, under the assumption that distortions in chess are more likely to yield similar board states rather than dissimilar ones. For instance, one might tally the number of cells between the two states containing the same pieces. In this approach, state (a) and state (b) come out as very similar, as they are identical in 63/64 cells (98.4%). Therefore, a surface-oriented approach may determine the likelihood of (b) coming from (a) to be very high.

The other approach, loosely corresponding comparative reconstruction and WDT, is to identify a sequence moves to derive (b) from (a) and approximate the proportion of board-states with the properties of (b) that can be derived from (a) through the same number of moves or fewer. In this carefully selected example, deriving (b) from (a) is rather difficult and requires many moves. In fact, the minimum number of distortions, or *ply* in chess jargon, that derive (b) from (a) is 17.[3]

[3] White must move the king out of the way of the pawn (1 move), promote to a queen on a8 (5 moves), sacrifice the queen on f6 (2 moves), and place the king back on a3 (1 move). The black king can do anything in the meantime, so long as it can capture the promoted queen on f6. A seasoned chess player might notice that some of the moves required by such a reconstruction are quite unlikely in practice. As such, a more realistic ‘reconstruction’ may require more moves still.

Just as with language, for the chess example in Figure 2, it only makes sense to estimate randomness with reference the synchronic properties displayed in the data. For instance, one might start with assumption that board states are synchronically compatible with (b) if they contain a white and a black king and no other pieces, and the two kings are positioned legally on the board (not adjacent), as is the case in (b). There are 3612 synchronically compatible board states of this kind.[4] It is tricky to calculate what proportion of these is diachronically local, i.e. can be reached from (a) in 17 ply. After some trial and error, I estimate $P(D|S)$ for a reconstruction from (a) to (b) to be approximately 0.65. In other words, approximately 65% of the time, setting up two kings at random results in a board state that can be reached from (a) as easily as (b) can. As such, a ‘comparative reconstruction’ and subsequent analysis in the spirit of WDT suggests that there is no reason to reject the null hypothesis that the two boards states in Figure 2 are simply random. Because Wordlist Distortion Theory is not surface-oriented, the fact that (a) and (b) appear so similar has no bearing on the result. Instead, the framework simply quantifies how unlikely it is to observe a random state that could have been reached from another through some number of distortions, i.e. chess moves.

Games such as chess, with discrete transformations that distort states in non-obvious ways, are particularly conducive to an analysis using Wordlist Distortion Theory. Naturally, chess variants, such as Japanese *shogi* and Chinese *xiangqi* can be analyzed in similar ways. Many other board games, such as *checkers* and *Othello* are also a good fit for the framework. The presence of luck, as found in board games such as such as *Monopoly* or *Settlers of Catan* or in most card games, does not preclude a Wordlist Distortion Theory analysis.

Photo editing is yet another domain where a Wordlist Distortion Theory analysis is meaningful. Figure 3 presents two images, the original image on the left and a version enhanced by automated software on the right (Oh, 2002). Assuming the history of the two images is unknown,

[4] The first king can be placed anywhere. If it is placed on one of the 36 squares not on the edge of the board, the second king can be placed on one of 55 squares, as the first king takes away 9 possible squares from the second. By the same logic, if the first king is placed on one of the edge squares (but not the corners), the second king can be placed on one of 58 squares. Finally, if the first king is placed on a corner, the second king can be placed on one of 60 squares. $36 * 55 + 24 * 58 + 4 * 60 = 3612$.

it is reasonable to ask about the likelihood that the picture in (b) is the product of photo manipulation conducted on (a).

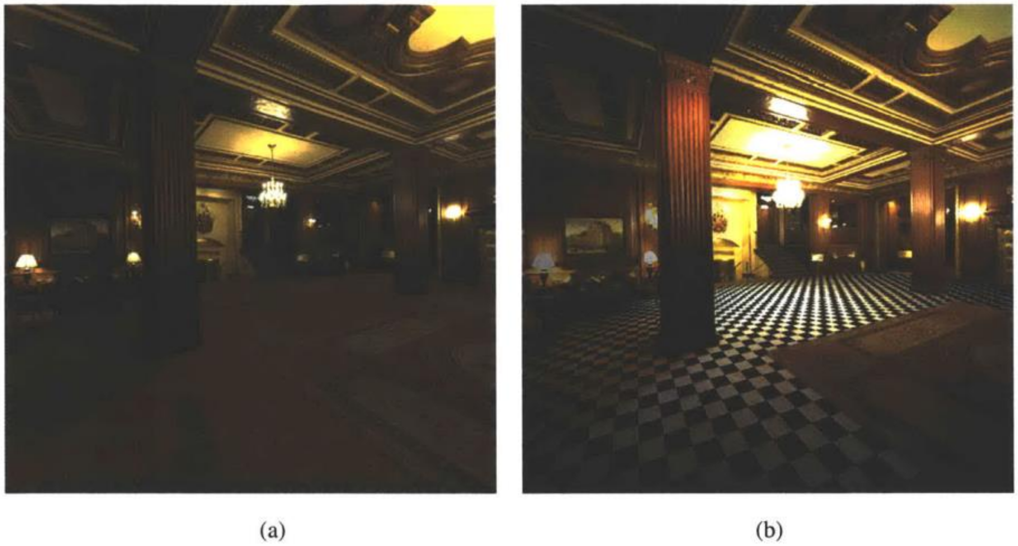


Figure 3: *Two images*. Original image is on the left; the edited image is on the right (Oh, 2002).

Once again, there are two approaches to this problem. A surface-oriented approach, loosely corresponding to lexicostatistics or multilateral comparison, might utilize a heuristic similarity measure, under the assumption that distortions on photos are more likely to yield similar photos rather than dissimilar ones. For example, one might measure the absolute distance between RGB values for every pixel in (a) and its corresponding pixel in (b) and average the results. Because (a) and (b) exhibit different brightness, shading, and background detail, it is very likely that a simple similarity measure such this one would find that the two images are not very similar. This in turn may lead the analysis to incorrectly keep the null hypothesis that these two images are simply random. However, it is possible that more advanced surface-oriented measures, e.g. those that track object edges, would successfully conclude that (b) is derived from (a).

The other approach, loosely corresponding to comparative reconstruction and WDT, is to find a ‘reconstruction’ from (a) to (b). In this case, a reconstruction corresponds to an exhaustive series of edits, using some photo-editing software, which transform (a) into (b). Thereafter, WDT would need to estimate the size of the synchronically compatible set, e.g. the number of all photos with the same color palette as (b), and the proportion of diachronically local pictures therein, those that can be derived from (a) using the same number and type of edits. Giving an estimate of $P(D|S)$ for the example in Figure 3 is impossible without a detailed review of picture composition and manipulation. However, assessing the two images visually suggests that the number of edits that occurred between the two is relatively low. In other words, it appears unlikely that a randomly generated image could be derived from (a) through as few edits as (b). This should in principle result in a relatively low $P(D|S)$ value.

In essence, Wordlist Distortion Theory can be used as a plagiarism detection tool for examples like in Figure 3. It is unclear if the domain of photo-editing can support well-defined notions of synchronically compatible and diachronically local sets. However, if it can, then Wordlist Distortion Theory would provide a different perspective from surface-oriented plagiarism detection algorithms, one that takes into account the photo-editing process itself. In the same vein, Wordlist Distortion Theory can be extended to other domains where plagiarism detection is a meaningful prospect, such as text and music.

It should be acknowledged that there are domains where an analysis in the style of Wordlist Distortion Theory is possible but not useful. Most notably, there is little sense in analyzing domains where each distortion replaces the state in its entirety. For example, in the game of craps, each round players bet on the roll of two dice. In principle, one could inquire about the probability that a given craps state or dice roll, e.g. ⚀ and ⚄, preceded another state, e.g. ⚅ and ⚃. However, because a single distortion, i.e. re-roll, in the game of craps can derive any state from any other state, the proportion of diachronically local states in the synchronically compatible states is always 1. In other words, all craps states are equally derivable from any other state. As a result, while a WDT analysis is possible but not illuminating.

Although there exist applications outside of linguistics, comparative reconstruction likely constitutes the most lucrative domain for Wordlist Distortion Theory. Even in fields where WDT could be used to estimate the likelihood that one state is derived from another, it is usually not the case that detecting and listing the number of transformations between prior states and later states is a common endeavor. For example, chess professionals are not trained to look at a chess board and figure out what moves have been played. Likewise, experienced graphic designers may not be able to reproduce the exact series of edits required to transform one picture into another. In contrast, it is one of the chief objectives of historical linguistics to identify and document all sound changes that have occurred in a language’s past. As such, historical linguistics is particularly suited for a derivation-oriented evaluation metric such as Wordlist Distortion Theory.

1.4 Overview

The remaining chapters are primarily focused on calculating the number of diachronically local wordlists introduced by various types of distortions, namely lexical replacement, semantic change, and sound change. The final chapter of the dissertation synthesizes the theoretical underpinnings of Wordlist Distortion Theory and introduces a machine learning algorithm for generating and evaluating reconstructions from a mother wordlist to a daughter wordlist. The algorithm is used to evaluate the genetic links within the Austronesian language family (Blust & Trussel, 2010) and its possible relation to the Ongan language family (Blevins, 2007; Blust, 2014). The results of the case study are in line with what is know about Austronesian and closely mirror the results of a multilateral comparison conducted on the same dataset.

Chapter 2 focuses on word-level transformations, those that operate on entire entries in the wordlist rather than the phonological forms within, namely lexical replacement and semantic change. Section 2.1 introduces a method of estimating the number of synchronically compatible wordlists. Section 2.2 presents two methods of estimating the number of diachronically local and synchronically compatible wordlists implied by lexical replacement. The first method is accurate but computationally unwieldy. The second method introduces some error but lends itself to simplification.

Section 2.3 discusses the role of semantic change in the framework. It is argued that, in spite of impressionistic critiques of comparative reconstructions in the literature, the effect of semantic change on $P(D|S)$ is relatively minor. In fact, it can be shown that that reliable reconstructions are possible even with across-the-board semantic change, i.e. those where comparisons in the mother and daughter wordlists are made exclusively between words of different meaning.

Chapter 3 focuses on segment-level transformations, those that manipulate the phonological form of an entry, specifically sound change. The chapter begins by introducing the *homotope* relation between segments in Section 3.1. Homotopes, defined as segments which can occur in the same phonological position in a language, are indispensable to the project, as they serve to both estimate the number of synchronically compatible wordlists as well as the number of diachronically local wordlists introduced by sound change.

Throughout the chapter, it is argued that that the effect of a sound change on $P(D|S)$ is dependent on its effect on the phonological system. Sound changes which erode contrast, discussed in Section 3.2, can drastically affect $P(D|S)$. Sound changes which preserve contrast in the same position, discussed in Section 3.3, have a negligible effect on $P(D|S)$. The effect of sound changes which alter the position of contrast within a word, discussed in Section 3.4, falls somewhere in between. The low impact of contrast-preserving sound change ensures that reconstructions between wordlists with non-overlapping inventories remains worthwhile, something that is not true for many other techniques in computational historical linguistics, such as multilateral comparison.

An unexpected corollary of the framework, discussed in Section 3.2.3 is that the effect of a proposed sound change on $P(D|S)$ can be estimated based on the number of words the change applies to. The section presents a rule-of-thumb formula which can be used to vet potential changes or word pairs. This formula is also useful in machine-learning implementations of the framework, as demonstrated in Chapter 4.

Chapter 4 presents a machine learning algorithm which suggests sound changes stochastically with a bias towards those that decrease $P(D|S)$. The algorithm was tested in a case study, where it was tasked with finding reconstructions from 5 Austronesian proto-languages to a genetically diverse set of 74 Austronesian languages. The results are in line with the general understanding in the field as well as the results of multilateral comparison. In comparisons of Austronesian proto-languages and direct descendants, the algorithm found reconstructions with $P(D|S) < 10^{-14}$ in all instances. Furthermore, the $P(D|S)$ of a reconstruction appears to be correlated with time-depth. Reconstructions from distant ancestors to attested languages yielded higher $P(D|S)$ than reconstructions from recent ancestors.

The case study also presents several insights about the Austronesian language family. The results confirm that the Batanic languages can be reliably reconstructed from proto-Philippine, a currently accepted but previously contentious classification. The result also show that Chamorro can be reliably reconstructed from Proto-Malayo-Polynesian, something which is still debated in the field. Finally, unusually high $P(D|S)$ values in reconstructions from proto-Philippine to

Philippine languages spoken in Indonesia suggest extensive contact between these and local non-Philippine languages.

The case study was extended to evaluate a possible genetic connection between Austronesian languages and Ongan languages, not widely accepted in the field. The algorithm found reliable reconstructions from the proto-Ongan-Austronesian wordlist to 28 of the 74 Austronesian languages tested. Reconstructions to the Ongan languages yielded mixed results, with a few seemingly reliable but some not reliable at all. In summary, the results with respect to the Ongan-Austronesian hypothesis are not conclusive. However, a multilateral comparison on the same dataset finds statistically significant phonetic similarity in most cases.

The simulated annealing machine learning algorithm, while not equipped with any notion of phonological naturalness, suggested sound changes which exhibited common phonological patterns in general and attested patterns in Austronesian in particular. A *Python* script implementation of the algorithm is made publicly available online.

Chapter 2
Word-level Operations

The current chapter as well as the following one are primarily concerned with estimating the number of diachronically local wordlists implied by transformations common in historical linguistics. In other words, the purpose of these two chapters is to estimate the effect that individual diachronic transformations have on the reliability of the reconstruction. The different types of diachronic transformations operate in different ways and imply different mathematical machinery. A method of combining the different formulae for estimating the number of diachronically local wordlists is presented in Section 3.6.

For convenience, this dissertation separates *word-level* operations from *segment-level* operations. Segment-level operations, such as sound change, are defined as those that alter the phonological content of an entry. In contrast, word-level operations, such as lexical replacement, are blind to the phonological content and operate on the entry in its entirety. Because word-level operations are mathematically simpler, they are addressed first in Chapter 2. The discussion of segment-level operations is reserved for Chapter 3.

This chapter also introduces a method for estimating the number of synchronically compatible wordlists, as in the denominator of the General Formula. The formula for estimating the set of synchronically compatible wordlists is expanded upon in Section 3.1 to account for segment-level operations.

Section 2.1 presents the method for estimating the number of synchronic wordlists used in this dissertation. Section 2.2 discusses the lexical replacement operation and introduces two methods for calculating diachronically local wordlists resulting from its application, one which is accurate but computationally inconvenient and one which is convenient but introduces error. Section 2.3 discusses the semantic change operation and its associated formula.

2.1 Synchronic Formula

This section presents a method of estimating the number of synchronically compatible wordlists based on word complexity. We define *word complexity* to be the number of phonologically possible forms in a language. This number is assumed to be finite though potentially extremely large. Word length in languages, while technically unbounded, is limited in practice, as word frequency tends to decrease with word length, following a roughly Zipfian distribution (Sigurd et al., 2004). Combined with a limited phonological inventory, this implies a finite number of phonologically possible strings.

Word complexity must differ from language to language, as both the size of the phonological inventory and the average word length varies. Languages also exhibit different restrictions on the arrangement of segments within a string, e.g. phonotactics. Accurately estimating word complexity from limited data is a challenge in and of itself and merits independent research. Nevertheless, an algorithm for doing so is presented in Section 3.1.2. For the current chapter, we treat word complexity as a given.

Recall that wordlists synchronically compatible with the daughter comprise phonological forms grammatical in the daughter. By definition, the number of forms that can appear in each entry of a synchronically compatible wordlist is equal to word complexity. The synchronically compatible set can be thought of as the set of all wordlists where each wordlist is drawn from the set of possible words, i.e. a set with cardinality equal to word complexity. It follows that a finite word complexity entails a finite synchronically compatible set *S*.

Although all words in a wordlist are selected from the same set of possible words, it is assumed here that they are otherwise phonologically independent. In other words, beyond the shared phonological inventory and phonotactics, the shape of one word cannot be predicted from the shape of another. There are a few reasons why this assumption may not be correct. Firstly, there is some evidence of between-word homophony avoidance in historical linguistics (Silverman, 2009; Ogura & Wang, 2018), supported both by computational modelling (Blevins & Wedel, 2009; Winter & Wedel, 2016) and by language learning experiments (Yin & White, 2018). Similarly, taboo avoidance is a well-known driver of language change (Burridge & Benczes, 2018). Thus, a natural language wordlist is less likely to exhibit homophony than a randomly generated one. We assume here that the dependence between phonological forms in a language is minor. However, WDT is compatible with formulae for calculating |*S*| that take homophony avoidance into account.

Under the assumptions that word shape is finite and independent, the way to calculate |*S*| is given in (1), where *c* is *word complexity* and *t* is *total wordlist length*. Stated intuitively, each phonological form for a wordlist in *S* is selected at random from the set of possible forms in the daughter language.

(1)

$$|S| = c^t$$

Synchronic Formula: Formula for calculating the size of the synchronically compatible wordlist set |*S*|, given the number of possible words *c* and wordlist size *t*.

As is discussed in the following section, the Synchronic Formula implies that |*S*| grows exponentially as a factor of wordlist length, meaning that *P*(*D*|*S*) decays exponentially as a factor of it. The growth of |*S*| from the increase of word complexity is polynomial.

2.2 Lexical Replacement

Lexical replacement is a diachronic process in historical linguistics whereby old lexical material is displaced by new lexical material. Often lexical replacement occurs as a result of language contact, i.e. borrowing. For example, due to contact between speakers of Old English and speakers of Norman French, Old English [sije] ‘victory’ (Mitchell & Robinson, 2012:397) was replaced by a French term for the same concept, whence modern English [vɪktəri]. The original Old English root was lost entirely and is evidenced only in written records and related languages, c.f. Dutch [zeɣə] ‘victory’. Lexical replacement can also take place between forms in the same language. For example, Old English [æːrdæj] ‘daybreak’, from [æːr] ‘before’ (whence *ere*) and [dæj] ‘day’ (Mitchell & Robinson, 2012:340), was replaced by a competing native term, a compound of *day* and *break*.

Lexical replacement has an obvious and substantial effect on reconstruction reliability; the more words have undergone lexical replacement, the less reliable the reconstruction as a result. The robust effect of lexical replacement on the history of a language is the reason that lexicostatistics and phylogenetic Bayesian methods are effective methods of approximating language history. Tallying the presence and absence of apparent cognates between wordlists is an indirect method for modelling lexical replacement. Although lexicostatistics ignores other types of diachronic transformations, the correlation between lexical replacement and language change in general is evidently strong enough that the phylogenies inferred from such approaches are mostly in line with what is discerned through traditional techniques (e.g. Greenhill et al., 2023).

However, it should be stressed that the approach to lexical replacement in WDT and in lexicostatistics is different. In lexicostatistics lexical replacement is, in a sense, epiphenomenal since it manifests only as the presence of a cognate in one wordlist and its absence in another. In contrast, in WDT lexical replacement is a type of diachronic transformation like any other (sound change, semantic change, etc.). Crucial also is that WDT compares mother and daughter wordlists, whereas lexicostatistics compares sister wordlists. As a result, the number of lexical replacements evidenced in a lexicostatistic comparison is far greater than in a WDT comparison, as the former is a result of independent changes in two languages. Most important, however, is that, like with any other transformation, WDT uses lexical replacement to estimate the reliability of the reconstruction rather than to discover aspects of the language's history. In contrast, lexicostatistics, as well as phylogenetic Bayesian methods that use cognate tables as input, compute the effects of lexical replacement in order to infer relatedness and phylogeny, as well as proto-language time depth or homeland.

Unfortunately, in comparative reconstructions, lexical replacement is often implied rather than stated. Because so much attention in the comparative method is granted to sound change, forms that do not undergo sound change, i.e. those that have been replaced, are sometimes omitted from reconstruction proposals. Nevertheless, any word in the daughter consulted during the reconstruction but not derived from the mother wordlist through a combination of other diachronic transformations must have undergone lexical replacement, whether or not this is made explicit. For instance, a reconstruction deriving 4 words through sound change after consulting 100 words is implicitly also positing 96 instances of replacement, one for every unexplained entry. It is worth stressing that an accurate estimate of $P(D|S)$ requires an exhaustive list of transformations, including lexical replacement. Failure to report lexical replacement is equivalent to cherry-picking data and greatly increases the chance of a false positive, a fact that has been pointed out in the literature previously (Ringe, 1999).

For the purposes of WDT, lexical replacement is an insertion operation, where a phonological form is inserted into an entry in the wordlist. The operation is blind to the phonological content of the original entry. In other words, the phonological shape of the input does not affect the phonological shape of the output.

(2) *Modern English and Old English*

gloss	*Old English*	*Modern English*
heart	xeorte	hɑɹt
eagle	eɑrn	igəl †
time	tide	tajm †
aluminum	--	æluminiəm †

Compare Old English and Modern English in (2), where replacement is marked with a dagger (†). The Old English [xeorte] is the direct ancestor of Modern English [hɑɹt]. Assume that the differences between the two forms can be explained using sound change and that the word-pair does not require replacement. In contrast, the original Germanic root for ‘eagle’, found in Old English [eɑrn] (modern dialectal *erne*), was replaced with the Anglo-Norman loan *egle*, whence Modern English [igəl], requiring an instance of lexical replacement in the reconstruction. The word pair [tide] ~ [tajm] ‘time’ also calls for an instance of lexical replacement. Although both roots existed in Old English, the Old English [tima] ‘season’ replaced [tide] in the entry corresponding to ‘time’. Therefore, for the purposes of a reconstruction on (2), [tajm] functions as a loan.[1]

An analogue of a word in the mother language may be absent from the mother language altogether. For example, largely because the element had not been discovered until the 19th century, Old English did not have a word unambiguously corresponding to ‘aluminum’. As a result, Old English wordlists, including the one in (2), do not give a phonological form for the entry. For the purposes of WDT, this too counts as replacement. We assume that Old English had a way to refer to aluminum (if such a need arose), e.g. perhaps with a term for a similar substance, but that this term is simply inaccessible to the reconstruction. In other words, no amount of sound change can derive the Modern English [æluminiəm] from Old English. However, due to the additional instance of lexical replacement required by the word-pair [] ~

[1] As is discussed in Section 2.3, cases where the replacing material already existed in the language are better analyzed as semantic change. However, because Old English [tima] is not part of the dataset in (2), at least from the perspective of the reconstruction, there is no evidence that a semantic change has taken place.

[ælumɪniəm], $P(D|S)$ of the reconstruction is still penalized for consulting the item during the reconstruction, i.e. checking for an Old English analogue.

In practice, entries such as ‘aluminum’ may be filtered out of the wordlist before the reconstruction processes, as the history of the substance strongly implies the absence of an Old English form. The question of which entries have the highest probability of substantiating a reconstruction is interesting but not strictly related to WDT.

The fact that different concepts exhibit different rates of lexical replacement (Embleton, 1986) is one of the intuitions behind the Swadesh list, a template for wordlists with low replacement rates and high attestation rates in the languages of the world (Swadesh, 1955). More generally, the rate of lexical replacement has been shown to be correlated with a multitude of factors. For instance, infrequent words are more likely to be replaced than frequent ones, while open class words are more likely to be replaced than closed class words (Pagel et al. 2007; Vajdemo & Hörberg, 2016). Numbers exhibit a particularly slow rate of replacement (Pagel & Meade, 2018). Words with concrete meanings are less likely to be replaced than words with abstract meanings and words with more polysemous meanings tend to be replaced less often in general (Vajdemo & Hörberg, 2016).

For the sake of presenting the simplest working model, this dissertation does not take differences in rate of lexical replacement into account. All entries in the wordlist are treated as equally likely to undergo replacement. For example, in (2), it is assumed that ‘heart’ is just as likely to be replaced as ‘eagle’ and ‘aluminum’. I leave the modelling of different rates of replacement to future implementations of WDT.

2.2.1 Lexical Replacement Formula

Understanding that replacement is a word level operation is crucial for calculating the number of diachronically local and synchronically compatible wordlists implied by the operation. The phonological output of lexical replacement is *ad hoc*. In other words, the phonological form resulting from lexical replacement could just as easily be any other licit string in the language and still evidence a reconstruction of the same size. As such, daughter wordlists with alternate outputs of lexical replacement are equally removed from the mother wordlist and considered diachronically local. For example, a descendant of Old English where [eɑrn] ‘eagle’ was replaced with [igəl] is no less arbitrary than one where the form was replaced with [kəkɑ] or [ɑbvəɹs].

The only phonological form that cannot be the output of lexical replacement is the input. In other words, a reconstruction cannot posit that an entry underwent replacement and was replaced with itself. For example, it is not possible for Old English [xeorte] ‘heart’ to be replaced with [xeorte]. In practice, there would be no way of knowing if such a replacement had taken place. Furthermore, because its presence or absence does not affect the wordlist, positing lexical replacement with the input as output is simply unparsimonious from the perspective of the researcher. A mathematical reason for the restriction on replacing a form with itself is provided in Section 2.2.2.

It is not only the output of replacement that is probabilistically arbitrary; the entries in the wordlist undergoing lexical replacement are also *ad hoc*. Lexical replacement could have just as easily applied to other entries and still merit a similar reconstruction. Therefore, daughter wordlists requiring the same number of replacements applied to different entries are considered equally as removed from the mother wordlist and are diachronically local. For example, a descendant of Old English which replaced ‘eagle’ with French [igel] is no less arbitrary than one which replaced ‘heart’ with, for example, [kəɹ].

Finally, recall that the set of diachronically local wordlists contains all wordlists derivable from the mother wordlist through as many or fewer changes as the daughter. Thus, if a reconstruction posits 10 instances of lexical replacement, diachronically local wordlists that are derived from the mother using 9 instances of lexical replacement should also be considered, as should be those that use 8, or 7, etc. By the same logic, wordlists derived from the mother through no lexical replacement is always a part of the diachronically local set.

With these provisions in mind, the number of wordlists both synchronically compatible with and diachronically local to the daughter given some number of lexical replacements is the number of wordlists that exhibit the same phonological restrictions as the daughter and can be derived from the mother with the same number or fewer replacements. This is expressed mathematically in (3).

(3)

$$|D \cap S|_l = \sum_{i=0}^{l} (c-1)^i \binom{t}{i}$$

Replacement Formula: Formula for calculating the number of wordlists that are both synchronically compatible with and diachronically local to the daughter in a reconstruction that posits a number of lost lexemes *l*, in a wordlist of length *t*, with word complexity *c*.

Let *l* be the number of *lexical replacements* in a wordlist. In practice, *l* ≤ *t*, since, due to parsimony, a reconstruction should not contain more replacements than entries. A summation is needed for all numbers of replacements from 0 up to *l* to account for diachronically local and synchronically compatible wordlists derived from the mother through fewer replacements than the daughter. The first term in the summation $(c-1)^i$ captures the number of possible replacements for the lost lexemes, i.e. the number of possible words but one. The second term in the summation, the choose function $\binom{t}{i}$, captures the way in which the replaced words can be distributed in the wordlist.

As a sanity check, inserting an *l* of 0 into the Replacement Formula always yields $|D \cap S|_l = \binom{t}{0} = 1$. In other words, a reconstruction that posits no instances of lexical replacement (or other transformations), i.e. where the mother and daughter wordlist are identical, has only one wordlist that is both synchronically compatible with and diachronically local to the daughter: the daughter itself. That is to say, the likelihood that a reconstruction positing no transformations can be substantiated by random chance is $\frac{1}{|S|}$, which is the likelihood of choosing the daughter out of the set of all synchronically compatible wordlists.

Less obvious is the fact that, when $t = l$, $|D \cap S| = |S|$. Put simply, this means that a reconstruction that posits lexical replacement for every word in the wordlist is compatible with every possible wordlist, i.e. every wordlist in *S* is also in *D*∩*S*. For this to be the case, the numerator and denominator in the General Formula must be equal, implying the equality in (4), where the left-hand side corresponds to |*S*| (Synchronic Formula), and the right-hand side to $|D \cap S|_l$ (Replacement Formula). This equality can be rearranged from a special instance of the binomial formula, though a complete derivation is out of place here.

(4)

$$c^t = \sum_{i=0}^{t} (c-1)^i \binom{t}{i}$$

As the Replacement Formula implies, the exact effect of the amount of lexical replacement *l* on *P*(*D*|*S*) depends on other variables, such as the number of possible words *c* and the size of the wordlist *t*. Due to parsimony, a large number of replacements precludes the appearance of other transformations in practice. For instance, there is no sense in a reconstruction that posits 90 instances of lexical replacement and 50 instances of semantic change in a wordlist of 100 words. As such, the effect of *l* on *P*(*D*|*S*) is also indirectly tied to the number and type of other transformations.

To better understand the correlation of lexical replacement and reconstruction likelihood, refer to Figure 4. *P*(*D*|*S*) increases roughly exponentially as a factor of *l*. A greater number of possible words results in a lower *P*(*D*|*S*). For all word complexity values, *P*(*D*|*S*) converges on 1 when the number of replaced entries equals the total number of words. The effect of word complexity *c* on the final value decreases roughly exponentially as a function of *c* itself.

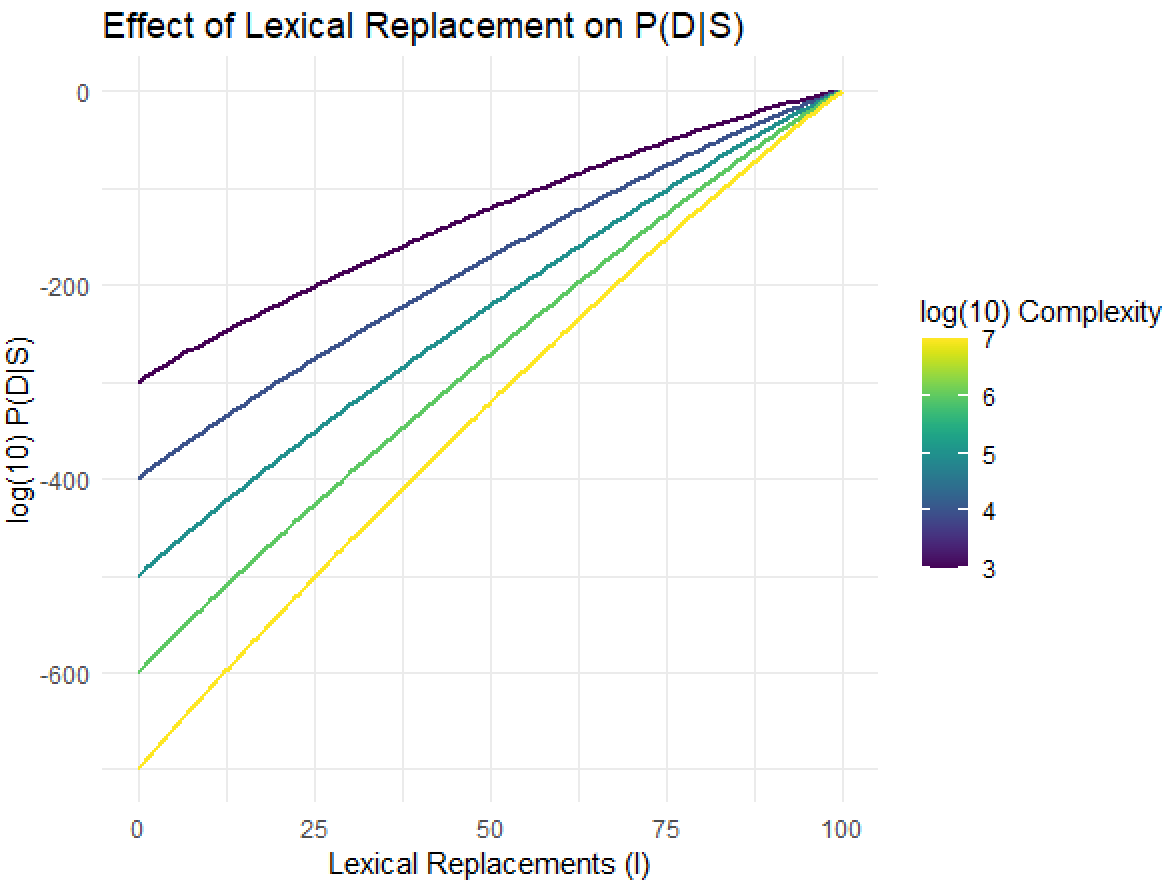


Figure 4: *Effect of lexical replacement on P(D|S)*. Log base 10 of *P*(*D*|*S*) is on the y-axis; number of replaced words *l* out of a total wordlist length *t* of 100 is on the x-axis. Exponential increase to word complexity *c* is represented by lines of different color.

Although the effect of replacement on *P*(*D*|*S*) is substantial, it is nevertheless possible for a reconstruction which posits lexical replacement for most words to be convincing, i.e. unlikely to be substantiated through a chance resemblance. For example, 4 perfect matches between mother and daughter forms out of a Swadesh list of 100, i.e. 96 instances of lexical replacement, still results in a 2 in 1 000 000 probability of being evidenced by random chance, as shown in (5).

(5)

$$\frac{\sum_{i=1}^{96}(1999)^{i}\binom{100}{i}}{2000^{100}} = 0.0000000236$$

As will become apparent in Chapters 3 and 4, the effect of lexical replacement on *P*(*D*|*S*) is larger than any other type of change. In a way, this is vindication for approaches in computational historical linguistics focused on tallying the presence and absence of cognates

(Swadesh, 1955; Greenhill, et al., 2017; Power et al., 2023). Therefore, for comparative reconstruction too, it is imperative that all instances of lexical replacement be listed. Otherwise, any quantitative evaluation of the reconstruction is likely to give a false positive.

The Replacement Formula is simple enough that it can be used to model other data types and domains. In fact, Appendix B serves to illustrate that, when $c = 2$, using the Replacement Formula to calculate $P(D|S)$ is equivalent to conducting a one-sided binomial test.

2.2.2 Lexical Replacement Approximation

Unfortunately, the calculations required by the Replacement Formula are often not feasible when dealing with real language data, as in (5), where incredibly modest values for word complexity and wordlist length still yield the astronomical $|S| = 2000^{100}$ or approximately 10^{330}. With a larger wordlist and richer phonotactics, it would not be surprising for $|S|$ to reach 200000^{500}, or approximately 10^{2651}. Because working with numbers of this magnitude poses a challenge to modern machinery, the case study in Chapter 4 of this dissertation utilizes a different method for calculating the number of synchronically compatible and diachronically local wordlists implied by lexical replacement. This alternative method is introduced in this section.

The main weakness of the Replacement Formula is that its components do not simplify with the denominator of the General Formula, i.e. the Synchronic Formula. This section will introduce the Replacement Approximation which, although yielding values of comparable size, lends itself to simplification and can be handled by even the humblest hardware. The drawback of this approximation is the introduction of error into the calculation.

The alternative method for evaluating lexical replacement lacks the stipulation that a phonological form cannot be replaced with itself but is otherwise similar to the method for calculating the Replacement Formula. Without this stipulation, the number of replacements for a lost word is simply the number of licit strings in the language, i.e. word complexity c. There is no longer any reason to consider diachronically local wordlists that are derived through fewer replacements than the number required by the daughter. To understand why this is the case, revisit the Old English to Modern English reconstruction in (6), where lexical replacement is marked with a dagger (†).

(6) *Modern English and Old English II*

gloss	*Old English*	*Modern English*	*Modern English′*
heart	xeorte	hɑɹt	hɑɹt
eagle	eɑrn	igəl †	əɹn (†)
time	tide	tajm †	tajm †
alunminum	--	æluminiəm †	æluminiəm †

Modern English requires three instances of lexical replacement. Modern English′ technically requires only two instances, if we assume that [əɹn] is derived from Old English [eɑrn] through regular sound change. Therefore, Modern English′ is both synchronically compatible with and diachronically local to Modern English. The Replacement Formula correctly counts Modern English′ in *D*∩*S* as a wordlist that can be derived from the mother through fewer transformations than the daughter. However, if lexical replacement has the ability to replace a phonological form with itself, then Modern English′ can also be derived from Old English through three instances of lexical replacement: two for ‘time’ and ‘aluminum’, and an additional replacement for ‘eagle’, where [əɹn] (or its ancestor) was replaced with itself. In this way, a framework that allows for lexical replacement to replace a word with itself automatically counts all synchronically compatible and diachronically local wordlists derived with fewer transformations than the daughter.

The Replacement Approximation is expressed mathematically in (7). Note the similarities with the Replacement Formula in (3). Both employ the choose function to calculate the number of ways to distribute *l* replacements between *t* entries. Both estimate the number of alternative outputs to lexical replacement that yield synchronically compatible and diachronically local wordlists; the Replacement Approximation sets this value at *c*, the number of possible words, while the Replacement Formula sets it at *c*-1. The Replacement Formula, but not the Replacement Approximation, has a separate count for the diachronically local wordlists resulting from the application of fewer replacements than required by the daughter.

(7)

$$|D \cap S| = c^l \binom{t}{l}$$

> Replacement Approximation: Approximation of the number of wordlists that are both synchronically compatible with and diachronically local to the daughter in a reconstruction that posits a number of lost lexemes *l*, in a wordlist of length *t*.

The primary advantage of this formula over the Replacement Formula is its simplification with the Synchronic Formula. When combining the two to calculate *P*(*D*|*S*), we can see in (8) that the exponents on *c* in the numerator and the denominator reduce.

(8)

$$P(D|S) = \frac{c^l\binom{t}{l}}{c^t} = \frac{\binom{t}{l}}{c^{t-l}}$$

Now, the calculation in (5) can be repeated in (9) without any difficulty.

(9)

$$\frac{\binom{100}{96}}{2000^{100-96}} = \frac{3921225}{2000^4} = .0000000245$$

If the values in the numerator and denominator are still too large to process computationally, both the Synchronic Formula and the Replacement Approximation are easily convertible into log units, as shown in (10) and (11) respectively. A log$_c$ is used here, as this conveniently removes the base in the synchronic formula entirely, though other bases can be used as well.

(10)

$$\log_c(\,|S|) = \log_c(c^t) = t$$

(11)

$$\log_c(\,|D \cap S|) = \log_c(c^l\binom{t}{l}) = l + \sum_{l+1}^{t} \log_c(i) - \log_c(i-l)$$

Although the Replacement Approximation is extremely convenient for Wordlist Distortion Theory, it does introduce some error. The error can be observed by comparing (5) and (9), where, although the inputs to the two formulas are identical, the outputs are different. The Replacement Approximation gives a slight overestimate of |*D*∩*S*| because it counts some diachronically local wordlists multiple times. To illustrate why this is the case, refer to the already seen Old English to Modern English reconstruction (12), where lexical replacement is marked with a dagger (†).

(12) *Modern English and Old English III*

gloss	*Old English*	*Modern English*	*Modern English′*	*Modern English″*
heart	xeorte	hɑɹt	hɑɹt	hɑɹt †
eagle	eɑrn	igəl †	əɹn †	əɹn
time	tide	tajm †	tajm †	tajm †
aluminum	--	æluminiəm †	æluminiəm †	æluminiəm †

Modern English again requires three instances of lexical replacement to be derived from Old English. Although Modern English′ technically requires only two, the faithful [əɹn] can once again be listed as replaced with itself. However, the same can be done for the faithful [hɑɹt] in Modern English″. Observe that, although the wordlists for Modern English′ and Modern English″ are identical, i.e. their entries match in phonological form, they assume different reconstructions. Because the Replacement Approximation counts reconstructions as proxies for wordlists, Modern English′ and Modern English″ are counted separately.

It should be stressed that, although the Replacement Formula also uses reconstructions as proxies for wordlists, the same problem does not arise. The reconstructions counted by the Replacement Formula are guaranteed to yield different wordlists. In the Replacement Formula, Modern English′ and Modern English″ merit only two instances of lexical replacement, one for ‘time’ and one for ‘aluminum’, and the two wordlists are counted as one.

To give an extreme example of the error introduced by the Replacement Approximation, imagine a situation where the size of the wordlist is greater than the number of possible words, which, although somewhat fantastical, is conceivable if the daughter wordlist exhibits homophony. In (13), the calculation is carried out for a reconstruction that requires lexical replacement in 999 out of 1000 words, with a total of 900 possible words in the daughter. The resulting $P(D|S)$ is 1.11. Although the example is somewhat far-fetched, it should be obvious that a probability greater than 1 is impossible. This value is particularly strange given that the daughter wordlist actually faithfully preserves the phonological form in one of its entries.

(13)

$$P(D|S) = \frac{900^{999}\binom{1000}{999}}{900^{1000}} = \frac{\binom{1000}{999}}{900^{1}} = \frac{1000}{900} = 1.11$$

The amount of error introduced by the Replacement Approximation is unbounded and varies by the number of words in the wordlist, the number of possible words, and the number of replacements. The effect of word complexity on the amount of error introduced by the Replacement Approximation is depicted in Figure 5, which compares $P(D|S)$ for the two formulas for lexical replacement.

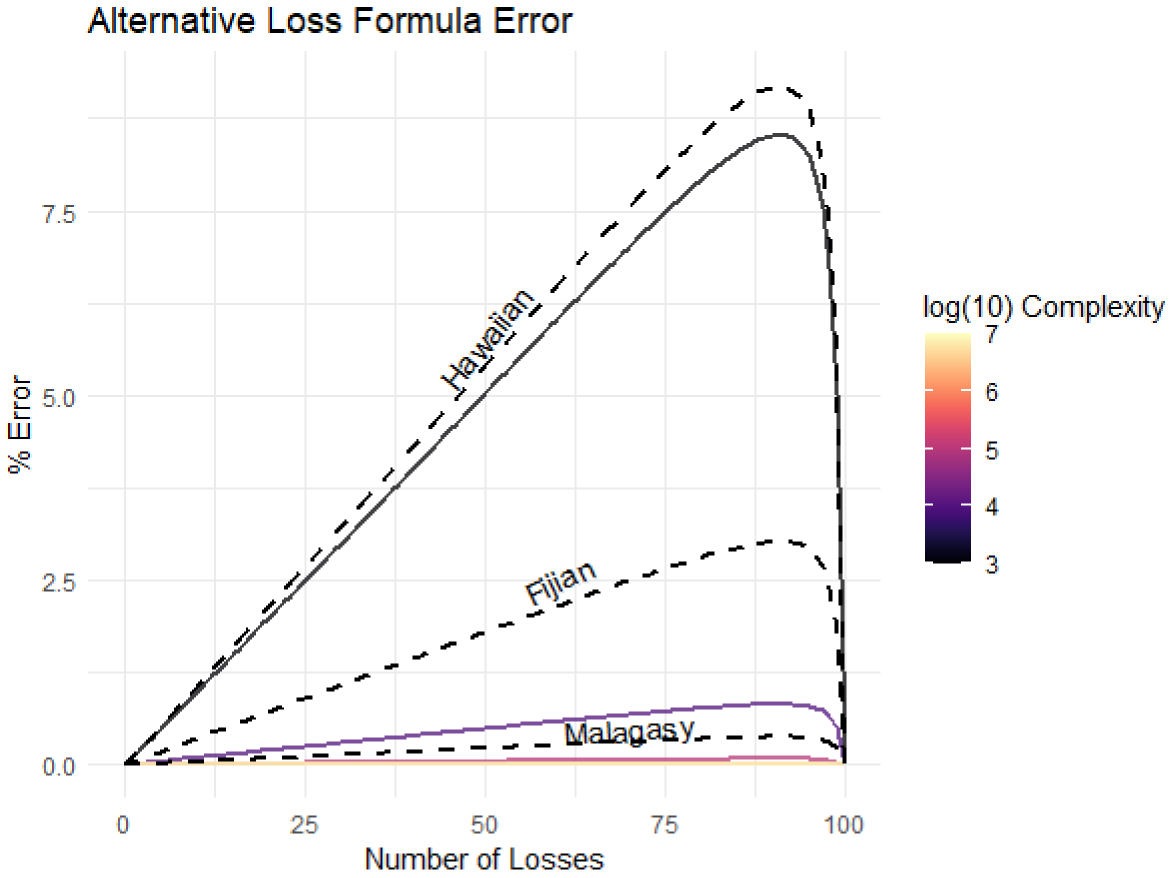


Figure 5: *Replacement approximation error*. Percent error in $P(D|S)$ from using the Replacement Approximation is on the y-axis; number of replaced words l out of a total wordlist length t of 100 is on the x-axis. Exponential increase to word complexity c is represented by lines of different color. Dashed lines correspond to word complexity values for sample languages found in the case study: Hawaiian (932), Fijian (2746), Malagasy (21627).

Figure 5 also presents error rate for complexity values found in sample languages in the case study. Of the three languages presented Hawaiian exhibits the lowest word complexity at 932, followed by Fijian at 2746, and Malagasy at 21627. All three languages exhibit a relatively low word complexity and, therefore, a high degree of error. Previewing the case study in Chapter 4, the average word complexity was approximately 50669, though this ranged from 699 in Ngadha

(Indonesia) to 523691 in Bikol (Philippines). Nevertheless, one would expect the degree of error introduced by the Replacement Approximation in real-life data to be relatively low, on the order of a few percent or less.

The two formulas yield the same result when $l = 0$ (no words are lost) and when $l = t$ (all words are lost). The Replacement Approximation gives the greatest error when l is just below t, as in (13). As word complexity c increases, the error decreases, though it never goes to zero.

In summary, the Replacement Approximation is computationally simpler than its regular counterpart and it readily lends itself to calculations in logarithmic space, at the cost of introducing a certain degree of error. When working with a language that has a large number of possible words or one exhibiting few instances of lexical replacement, the error is likely negligible. Additionally, because the error is always an overestimate of $|D \cap S|$, which in turn results in an overestimate of $P(D|S)$, the Replacement Approximation may be tried in cases where type II error is not a concern. In all other situations, the theoretically sound Replacement Formula is preferable.

2.3 Semantic Change

Semantic change is a diachronic process in historical linguistics whereby the meaning invoked by some phonological material changes over time. Descriptively, this change can fall into several broad categories. For example, Old English [dogga], whence Modern English [dɑg] ‘dog’, originally only referred to specific kinds of dogs. As such, the word *dog* can be said to have undergone semantic broadening (also known as generalization). In contrast, Old English [hund], whence Modern English [hawnd] ‘hound’, referred to all dogs, not just hounds. As such, the word *hound* can be said to have undergone semantic narrowing. There are many other types of semantic changes, such as pejoration and metonymization (Fortson, 2003; Traugott, 2017).

Although ubiquitous in languages of the world, semantic change does not play a large role in most computational approaches in historical linguistics. Techniques such as lexicostatistics and multilateral comparison presuppose the existence of cognate sets (Swadesh, 1955; Greenhill, et al., 2017; Power et al., 2023). Before computing mean pairwise distance or proportion of shared cognate between wordlists, the pairs of cognates must already be established. In effect, therefore, genetic relatedness as estimated through lexicostatistics or multilateral comparison is conducted without explicit consideration of semantic change.

However, a significant portion of computational linguistics research is concerned detecting and tracking the evolution of word meaning across time from real-world data (Tahmasebi at al., 2021). Such studies usually employ mathematical measures of word cooccurrence as a proxy for semantic proximity, e.g. pointwise mutual information (Mahanty et al., 2021). Generally speaking, regardless of the measure employed, words that are likely to cooccur within some predetermined window size in a corpus (e.g. five words) are treated as semantically related. For example, if the corpus exhibits many instances of the word *night* near the word *day* (e.g. "…all day and all night…") but not many instances of *night* near *castle*, then the semantic distance between *day* and *night* is assumed to be smaller than between *day* and *castle*.

Given a diachronic corpus, the cooccurrence measures can be used to track changes in word distribution over time, which, in turn, imply changes in word meaning. For example, this type of methodology has been used to show that words related to psychological harm (e.g. *trauma*, *bullying*, *harassment*) have undergone semantic broadening since the 1970s (Vylomova & Haslam, 2021). The same techniques can also reveal more general properties of semantic change. For instance, it has been shown that words of related meanings, such synonyms and antonyms, tend to undergo related changes in meaning (Xu & Kemp, 2015) and that words with polysemous meanings tend to change less overall (Hamilton et al., 2016).

However, because there is no technique for estimating reconstruction reliability, there is no way to evaluate the effect of semantic change in a comparative reconstruction setting. In contrast, estimating reconstruction reliability and the effect of diachronic transformations on reconstruction reliability is the purview of Wordlist Distortion Theory. As such, the approach taken in this dissertation with respect to semantic change bears little overlap to previous approaches in the literature.

For the purposes of WDT, semantic change refers to any instance where the phonological form in one entry in the daughter wordlist is derived from the phonological form in a different entry in the mother wordlist. In other words, semantic change is a copy operation on the phonological content of a word. For example, refer to the Latin and Romanian wordlists in (14), where the etymological origin of the Romanian forms is indicated with an arrow.

(14) *Romanian and Latin*

gloss	*Latin*	*Romanian*
woman	feːmina	femeje
household	familia	kasə
cottage	kasa	vilə
mansion	wilːa	vilə

Recall that entries in WDT are determined by the semantic properties of the word rather than its phonological (or etymological) properties. Thus, in (14), Romanian [vilə] meaning ‘cottage’ and Romanian [vilə] meaning ‘mansion’ are separate entries. The fact that the two are etymologically related is not indicated in the dataset itself. Instead, the relation can be signaled in the reconstruction, by positing a semantic change from one entry into the other. This perspective is particularly useful for WDT as it makes semantic change discrete and easy to combine with other types of diachronic transformation.

Nevertheless, the reverse perspective is more common in research that tracks changes in word meaning over time (Xu & Kemp, 2015; Hamilton et al., 2016; Vylomova & Haslam, 2021). In this view, words are delineated by their etymological origin rather than by their semantics. Thus, Romanian [vilə] is treated as a single entry and any change in its usage over time indicates some change in the semantics. This perspective allows for a nuanced and gradient analysis of word meaning, something that is useful when discussing the minutia of semantic change.

As such, for the purposes of WDT, the Romanian entry for ‘woman’ in (14) does not come from the Latin word for ‘woman’, but rather the Latin entry for ‘household’. The Romanian for ‘household’ itself comes from the Latin for ‘cottage’. Finally, the Latin for ‘mansion’ is the source for both the Romanian for ‘mansion’ and the Romanian for ‘cottage’. Therefore, Romanian underwent three instances of semantic change; the phonological content in one entry was copied into another entry three times: one for Lat. *casa*, one for Lat. *familia*, and one for Lat. *femina*. Note that the words in (14) have also undergone segment-level operations.

Semantic change is conceptually very similar to lexical replacement. In fact, if the slot with the meaning ‘household’, for example, were not part of the dataset in (14), there would be no recourse but to count the output of Latin *femina* as a replacement in Romanian, since the source of Romanian *femeie* could not be identified. Therefore, the only difference between lexical replacement and semantic change is that the output of the former is limited to all possible words, whereas the output of the latter is limited to the words in the wordlist, a subset of possible words.

In fact, for this reason, it is not uncommon for the two processes to be grouped together in the literature (Tahmasebi et al., 2021).

Just as with lexical replacement, it is worth pointing out that the semantic change operation is indispensable to any reconstruction. Although this is not always stated explicitly, any phonological form not occupying the same slot in the wordlist as its purported ancestor must have undergone semantic change. Therefore, it is imperative for an accurate estimate in Wordlist Distortion Theory that instances of semantic mismatch are reported as semantic shift. Failure to do so, would result in an underestimation of $P(D|S)$, i.e. an increase in type I error.

Note that semantic granularity may differ across different analyses. As such, the degree of semantic shift required to merit a semantic change may vary. For example, while ‘cottage’ and ‘mansion’ are kept separate in (14), that need not be the case in every reconstruction. A different analysis may choose to conflate the two meanings under the same entry in the wordlist. In such an analysis, no semantic change occurs between Lat. [wilːa] and Rom. [vilə].

There is no downside to conflating meanings into a single entry in this way. However, decisions with respect to semantic granularity in the wordlist should not be *ad hoc*. Ideally, a predetermined list of meanings to be used in the reconstruction should be assembled before consulting the data, e.g. a Swadesh list. Also before consulting the data, a decision should be made about potential forms competing for the same entry in the wordlist, i.e. synonyms. For example, only the most common synonym should be used, or the first synonym in the dictionary, or the synonym with a narrower usage, etc.

2.3.1 Semantic Change Formula

As in the case of lexical replacement, or even more so, it is well established that the trajectory of semantic change is not random (Traugott & Dasher, 2001; Hansen & Visconti, 2009). The main variable in the likelihood of semantic change is the semantic proximity of the two concepts in question. Semantic change between closely related concepts is more likely than semantic change between distantly related or unrelated concepts.

However, once again in the interest of presenting a simple but functional model, this dissertation assumes that each word in the wordlist is equally likely to undergo semantic change, i.e. have its phonological content replaced, with every other word in the wordlist. Wordlist Distortion Theory is compatible with approaches to semantic change that further restrict the output of the process. As such, future applications of the theory may want to take a more nuanced approach to semantic change.

Due to the similarities between semantic change and lexical replacement, the formula for estimating the number of synchronically compatible and diachronically local wordlists introduced by the two processes is almost identical. Let *s* be the number of *semantic changes* required by the reconstruction. The Semantic Change Formula for calculating the number of diachronically local and synchronically compatible wordlists introduced by semantic change is given in (15).

(15)

$$|D \cap S|_s = \sum_{i=0}^{s} (\min\{c, t\} - 1)^i \binom{t}{i}$$

Semantic Change Formula: Formula for calculating the number of wordlists that are both synchronically compatible with and diachronically local to the daughter in a reconstruction that posits a number of semantic changes *s*, in a wordlist of length t¸ and possible words *c*.

Both word-level operations select a subset of the wordlist to undergo the transformation. Both operations replace the phonological form of the input in its entirety. The only difference between lexical replacement and semantic change is the set of possible outputs. In the case of replacement, the output can be any phonologically licit string. In the case of semantic change, the output can be either any string already in the wordlist or any phonologically licit string, whichever is smaller. This caveat is required for situations where the number of words in the wordlist is greater than the number of possible words, which occurs exclusively in wordlists with (perhaps unrealistically high) degrees of homophony. Semantic change targeting different homophonous entries as inputs can result in overcounting of $|D \cap S|$, since the same output can be derived from different inputs. Without the caveat, this can lead to $P(D|S) > 1$. It highly unlikely that a wordlist in a natural language is found exhibiting fewer possible words than the wordlist length. As such, the outputs of semantic change are in practice limited to the phonological forms in the daughter wordlist. Therefore, the Semantic Change Formula is generally expected to choose replacements out of a set of size *t*, the number of words in the wordlist, and not out of one of size *c*, the number of possible words.

Since the only difference between the Semantic Change Formula and the Replacement Formula is the set of possible replacements, so long as the number of words in the wordlist *t* is less than the number of possible words *c*, the number of diachronically local wordlists resulting from the application of semantic change must be smaller than the number resulting from the application of lexical replacement. Recall that when the number of replacements is equal to the number of words in the wordlist, the Replacement Formula yields $P(D|S) = 1$. Therefore, so long as $t < c$, the Semantic Change Formula yields $P(D|S) < 1$ for any number of semantic changes. In other words, even if a reconstruction posits across-the-board semantic change, i.e. if comparison is made exclusively between forms of different meaning, the likelihood that the reconstruction can be substantiated by chance resemblance is less than 1, so long as the number of possible words is greater than the number of words in the wordlist and so long as no other changes are posited in the reconstruction.

To assess the relationship between the number of semantic changes *s* and $P(D|S)$ refer to Figure 6. The growth of $P(D|S)$ relative to *s* is almost exponential. As expected, word complexity is inversely correlated with $P(D|S)$. The degree to which word complexity affects $P(D|S)$ is a function of word complexity itself. Unlike in the case of lexical replacement, $P(D|S)$ neither reaches 1 nor converges on a given value, and its maximum depends on word complexity and wordlist length.

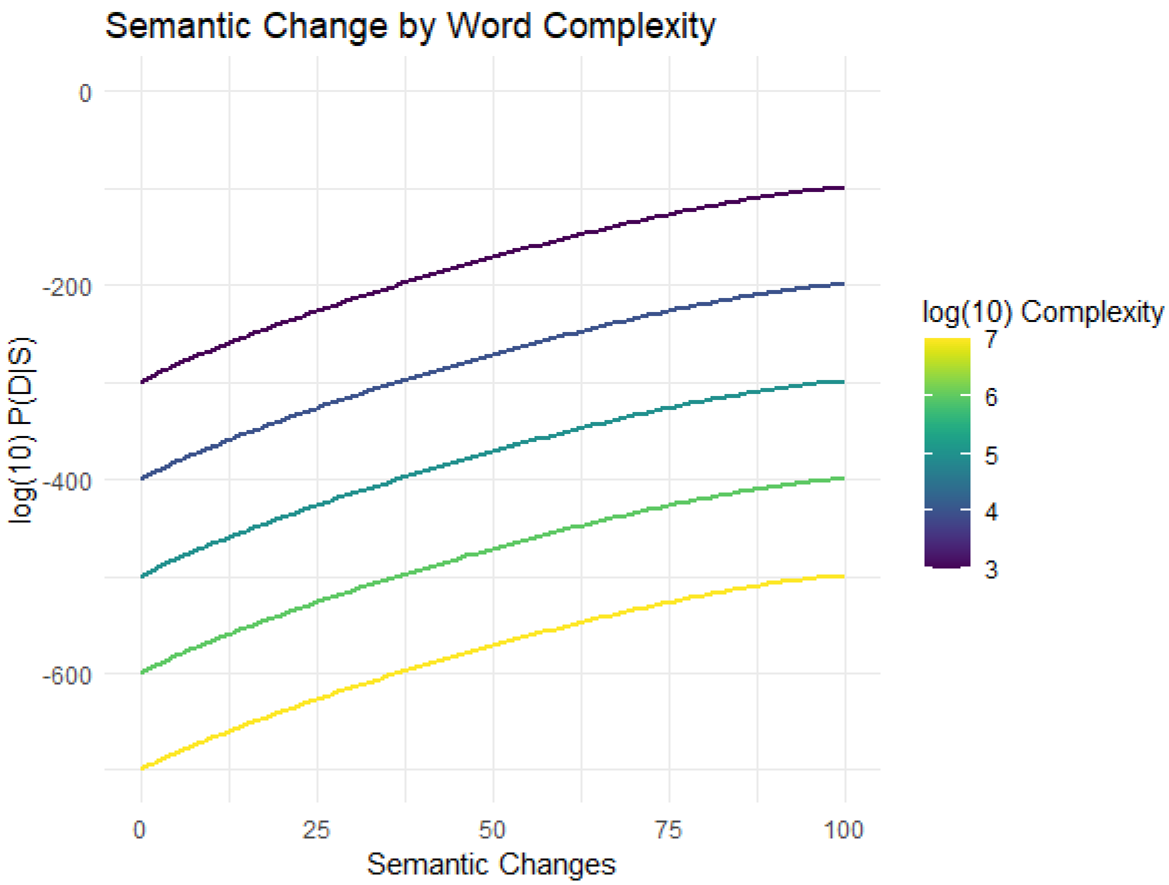


Figure 6: *Effect of semantic change on* P(D|S). Log base 10 of *P*(*D*|*S*) is on the y-axis; proportion of semantic changes *s* is on the x-axis. Different values for possible number of words *c* are represented by color in intervals of 10000 from 1000 to 91000.

The likelihood that a reconstruction can be substantiated by chance can be considered minor in comparison to the corresponding effect of lexical replacement and, arguably, in comparison to the corresponding effect of sound change, as is discussed in Chapter 3. Although across the board semantic change in Figure 6 can increase *P*(*D*|*S*) by 200 orders of magnitude, unless $t \geq c$, *P*(*D*|*S*) will never reach 1 due to semantic change alone. The synchronically compatible set, i.e. the denominator of the General Formula, simply has more degrees of freedom to vary than the intersection of the diachronically local and synchronically compatible set resulting from the application of semantic change, as the former fills each entry with one of *c* forms and the latter with one of *t* forms. So long as word complexity *c* is greater than wordlist length *t*, the effect of semantic change alone on reconstruction reliability should likely not be decisive.

This conclusion is in stark contrast with the attitude towards semantic change in the comparative reconstruction literature. Perhaps the most common locus of criticism for controversial reconstructions is the fact that word pairs not matching in meaning are used, i.e. word pairs that exhibit semantic change (Baxter, 1995; Ringe, 1999; Vovin, 2005). Some sources (Swadesh, 1955; Ringe, 1999) even surmise that comparison should exclusively be made between words that are direct translations of each other, eschewing the role of semantic change in historical linguistics altogether.

This avenue of criticism appears weaker when a probabilistic evaluation of comparative reconstruction is taken into account. It is unlikely that a spurious connection between two random wordlists can be fabricated simply by positing too many instances of semantic change in the reconstruction, in other words, by comparing semantically non-congruent forms. Nevertheless, it is the case that semantic change in cooperation with other diachronic transformations can result in a probabilistically unreliable reconstruction. Therefore, as in the case of lexical replacement, semantic change should not be omitted or ignored. Any comparison between forms with incongruous meanings implies an instance of semantic shift in the analysis.

Recall that this dissertation takes a simplified approach to semantic change; we assume that semantic change is equally likely between all entries in the wordlist. However, the relatively small effect of semantic change on $P(D|S)$ persists even if a more nuanced approach is employed. In fact, in a more nuanced approach to semantic change, one where the output of semantic change is taken from a subset of the wordlist, the effect of semantic change on $P(D|S)$ would be smaller still. As it stands, a wordlist requiring a semantic change, say from ‘mother’ > ‘sister’ (as occurred in the history of Albanian, cf. Alb. *motër* ‘sister’ and Latin *mater* ‘mother’), populates $|D \cap S|$ with diachronically local wordlists that exhibit semantic changes from the same input (‘mother’) to all possible outputs in the daughter wordlist, as well as semantic changes from other inputs (e.g. ‘father’) to all possible outputs in the daughter wordlist. Some of these changes, such as ‘mother’ > ‘aunt’ and ‘father’ > ‘brother’, appear reasonable and yield wordlists that would likely be considered diachronically local in any analysis. However, the vast majority of semantic changes between two random entries in the wordlist are entirely unrealistic, e.g. ‘mother’ > ‘anarchy’ or ‘father’ > ‘port-wine’. As a result, a more involved semantic analysis would remove wordlists exhibiting such semantic changes from the diachronically local set, reducing $P(D|S)$ beyond what is given here.

2.4 Chapter Summary

The fact that even across-the-board semantic change does not invalidate a reconstruction on its own demonstrates the utility of Wordlist Distortion Theory, as it likely cannot be reached on purely intuitive grounds. This chapter presented other incidental conclusions about the connection of word-level operations on reconstruction reliability. Wordlist Distortion Theory demonstrates that the effect of each additional replacement or semantic change on $P(D|S)$ is roughly exponential, that $P(D|S)$ converges on 1 as the number of lexical replacements approaches the number of entries in the wordlist, and that $P(D|S)$ does not converge as the number of semantic changes increases.

Perhaps the biggest take-away from this chapter is methodological. Word-level operations such as lexical replacement and semantic change are often forgotten and omitted from the reconstruction in exchange for a greater focus on segment-level operations. While the inclusion of these changes into the reconstruction can be argued for on intuitive grounds, the probabilistic argument is more concrete. Failure to derive words through other changes implies lexical replacement; comparison of forms that are not direct translations of one another implies semantic change. While almost all reconstructions posit these processes implicitly, it is impossible to accurately estimate the reliability of a reconstruction without an exhaustive list of both segment-level and word-level operations required to derive the wordlists in questions. As such, in addition to changes in phonology and morphology, reconstructions should devote some explanation to word-level transformations that must have taken place in the history of the language.

Chapter 3
Segment-Level Operations

This chapter is concerned with using Wordlist Distortion Theory to calculate $P(D|S)$ of a reconstruction based on the number of regular sound changes proposed. As in the case of word-level operations in the previous chapter, the logic of the calculations follows from the General Formula introduced in Section 1.2. For each transformation required to derive the daughter wordlist from the mother wordlist, WDT is used to estimate the number of implied synchronically compatible and diachronically local wordlists. The greater the proportion of diachronically local wordlists in the set of synchronically compatible wordlists, the greater the chance that a randomly generated wordlists could have evidenced a reconstruction of the same magnitude as the one required by the daughter wordlist.

It should be immediately apparent that segment-level operations require more mathematical machinery to be integrated into WDT than do word-level operations. To simplify the discussion, the dissertation treats sound change as segment-to-segment mapping, either unconditioned or conditioned by exactly one adjacent environment. While powerful enough to encode all segmental changes in historical linguistics, this model omits much of what is known about sound change naturalness. A more nuanced model, one that utilizes phonological features, is discussed at the end of the chapter.

The most important finding of this chapter is the effect of sound change on $P(D|S)$. It follows from the General Formula that all operations with multiple possible outputs, be they word-level or segment-level, introduce diachronically local and synchronically compatible wordlists. As a result, all diachronic transformations increase the arbitrariness of a reconstruction, sound change included. This finding appears contrary to the general attitude toward sound change in the historical linguistic literature, where reconstructions are usually critiqued based on the presence of other diachronic transformations (e.g. replacement or semantic change) but rarely based on the number of sound changes.

This chapter also reveals several non-intuitive corollaries of Wordlist Distortion Theory. It is shown that the number of sound changes that may be posited in a reconstruction without increasing *P*(*D*|*S*) is linearly correlated with average word length in the language. It is also shown that the effect of contrast-preserving sound changes on *P*(*D*|*S*) is minimal. In fact, the first contrast-preserving change posited has no effect on *P*(*D*|*S*). Moreover, for wordlists of reasonable length, even across-the-board contrast-preserving sound change results in a negligible increase.

Finally, this chapter also introduces the *homotope* relation between segments. Segments are defined as homotopes if they can occur in the same phonological environment in a language. This concept is essential to Wordlist Distortion Theory for two reasons. Firstly, homotopes define the number of sound change outputs and environments that yield synchronically compatible wordlists in a language. Secondly, homotopes are useful in the estimation of word complexity in a language.

Section 3.1 discusses the homotope relation and connects it to the Synchronic Formula. Section 3.2 focuses on the effects of mergers on *P*(*D*|*S*), starting with unconditioned mergers and moving on to conditioned mergers. Section 3.3 discusses contrast-preserving shifts. Section 3.4 is focused on chain shifts and sound change interaction. Section 3.5 outlines a method of incorporating phonological features into the framework. Section 3.6 concludes with a brief discussion of how the various formulas introduced in this dissertation are to be combined in practice.

3.1 Homotopes

Recall that the General Formula calculates the proportion of synchronically compatible wordlists that are diachronically local given the daughter and a reconstruction from the mother. Because word-level operations replace an entry in its entirety and because each word-level operation affects only a single entry, it is trivial to check whether the application of lexical replacement or semantic change results in a synchronically compatible wordlist. The resulting wordlist is synchronically compatible if the output is a possible word in the daughter.

In contrast, it is not immediately clear which segment-level operations (or how many) yield synchronically compatible wordlists. For an example, refer to the data in (1), where modern Japanese is derived from a hypothesized earlier stage of the language.

(1) *Japanese Palatalization*

gloss	*Middle Japanese*	*Japanese*	*Japanese'*	*Japanese"*
leg	asi	aʃi	aɾi	azi
stone	isi	iʃi	iɾi	izi

A reconstruction for the data in (1) may need to posit the sound change *s > ʃ / _i (palatalization).[1] In the analysis, the elements of the sound change – the input, output, and environment – are motivated by nothing except for the data at hand and are, therefore, *ad hoc*. Any alternative sound changes that can derive a wordlist synchronically compatible with modern Japanese populates the set of diachronically local and synchronically compatible wordlists.

Reasonable alternative to palatalization could be *s > ɾ /_i (rhotacization) or *s > z /_i (voicing), corresponding to Japanese′ and Japanese″ respectively. As modern Japanese itself, these two wordlists can be derived from the mother through one sound change and are, therefore, diachronically local to it given the reconstruction. One can confirm that both [ɾ] and [z] are possible segments of Japanese in examples such as [kazaɾi] ‘decorations’ and [iɾezumi] ‘tattoo’. However, to be certain that Japanese′ and Japanese″ are synchronically compatible with Japanese, one must additionally confirm that [ɾ] and [z] are grammatical before [i] in particular, since this is where the changes take place in the dataset. As it turns out, while [ri] is a possible sequence in Japanese, *[zi] is not. As a result, while Japanese′ is both synchronically compatible with and diachronically local to Japanese, Japanese″ is diachronically local but not synchronically compatible.

Performing a phonological analysis of such depth on every potential member of the diachronically local set is not feasible, as the number of diachronically local wordlists given reconstruction may routinely surpass 10^{1000}. Therefore, rather than test whether each potential alternative change when applied to each word in the wordlist yields a grammatical result, it is

[1] This change is evidenced primarily by morphophonological alterations in Modern Japanese, such as [kas-u] ‘lend’ vs [kaʃ-imasu] ‘lend (polit.)’. It is unclear at what point this change took place, though it could not have been after the 16th century, since transcriptions of the language by Portuguese missionaries of the time evidence palatalization (Miyake, 2003). Thus, the Middle Japanese forms in (1) do not necessarily represent the pronunciation at a particular point in time.

more advantageous to estimate the average number of sound changes which yield wordlists synchronically compatible with the daughter. Recall that only the cardinality of the intersection of the synchronically compatible and diachronically local sets is required in the General Formula. The exact membership of the two sets or their intersection is not relevant to the calculation.

To estimate the number of diachronically local and synchronically compatible wordlists, this dissertation introduces the *homotope* relation. A segment *x* is a defined to be a homotope of segment *y* in environment *z* if and only if both *x* and *y* can occur in environment *z*. Looking back at (1), one can say that [ɾ] is a homotope of [ʃ] before [i] in Japanese, while [z] is not. The fact that Japanese″ is not synchronically compatible with Japanese follows directly from the fact that the outputs of the two sound changes used to derive these wordlists are not homotopes of one another in the given environment. In other environments, such as before [oː]. all three sounds may still be homotopes of one another, cf. [ɾoːsaku] ‘labor’, [ʃoːsaku] ‘ruse’, and [zoːsaku] ‘building (process)’.

By definition, if two segments are homotopes in a given position, then replacing one with the other still results in a phonologically possible word. Therefore, to estimate the number of synchronically compatible and diachronically local wordlists implied by a given sound change in a reconstruction to the daughter, one need only know the number of homotopes of the output in the sound change.

3.1.1 Counting Homotopes

Rather than searching for homotopes of individual segments in individual environments, it is more advantageous to calculate a single mean homotopy value per language, one that describes how many segments on average can appear in a phonological position. Let us call this mean number of homotopes $\hbar$, by analogy with the sample mean $\overline{x}$ in statistics. A larger $\hbar$ implies that a random segment in a random word can be replaced with many other segments, either due to a large phoneme inventory or few phonotactic restrictions, while a small $\hbar$ implies the reverse, either a small phoneme inventory or many phonotactic restrictions.

To get an intuitive sense for homotopes, it is simplest to imagine a language with obligatory onsets and no codas, i.e. a language with CV syllables only, where each C and V cooccur freely. In such a language, words comprise alternating vowels and consonants; all vowels are homotopes of each other but not of consonants; all consonants are homotopes of each other but not of vowels. Stated differently, any vowel can be replaced with any other vowel but not a consonant, and every consonant can be replaced by any other consonant but not a vowel.

To estimate the mean number of homotopes $\hbar$ we will turn to the concept of entropy in information theory. This is not strictly speaking necessary, as it is possible to simply compute the average number of adjacent environments in language independently. However, because the resulting calculation is almost identical to the one used to compute conditional entropy in information theory, we will use information theory to ground our discussion.

Entropy, written as $H(x)$, where x is the variable of interest, is an estimate of a system's average amount of information per symbol, also known as *surprise* or loosely speaking *randomness* (Lubbe, 2018). A variable that is entirely random is maximally informative and its elements are maximally surprising. For example, the entropy of a series of fair coin tosses should approach 1 bit, which is $\log_2 2$. If the coin is not fair, or if results of the coin tosses are correlated, the entropy per coin toss is lower than 1 bit. The units of measurement of entropy are multiplicative and are defined by the base. Thus, an additional bit of information corresponds to a doubling of the number of possible outcomes.

The concept of entropy has proven useful in many disciplines. Entropy has been used to schedule factory production (Yang et al., 2020), identify geothermally active locations (Li et al., 2022) and groundwater quality (Hasan & Rai, 2020). It has also been used to quantify the visual complexity of landscape scenery (Kuper, 2020) and the effective cast-size for Hollywood blockbusters (Roughan et al., 2020). In linguistics, entropy is commonly used to determine the amount of information in the linguistic signal, both on the level of the sentence (Montemurro & Zanette, 2011; Shi & Lei, 2020) and on the level of the word (Pimentaal et al., 2020; Vera et al., 2021). However, it can also be used to identify periods of change in word usage (Degaetano-Ortlieb & Teich, 2018).

The *conditional entropy*, written *H*(*x*|*y*), where *x* is the variable of interest and *y* is the conditioning variable, measures the average amount of information in a variable given that the value of another variable is known. For the purposes of this dissertation, it is useful to think of the conditional entropy of a segment given the knowledge of an adjacent segment.[2] As with homotopy, conditional entropy for adjacent segments is lower for languages with fewer segments or with languages with more phonotactic restrictions, i.e. languages where segments are strongly predictable based on adjacent segments. The formula for conditional entropy is given in (2); it is the negative sum of the joint probabilities of each instance of *x* occurring with each instance of *y* multiplied by the log of the conditional probability of each instance of *y* given each instance of *x*. For our purposes, *X* and *Y* comprise the phoneme inventory of the daughter. Values of *x* are segments, and values of *y* are the segments adjacent to *x*.

(2)

$$H(x|y) = -\sum_{x \in X, y \in \mathrm{Y}} p(x, y) \log p(y|x)$$

Conditional Entropy Formula: Formula for calculating the average amount of surprise in a variable X conditioned on a variable Y, for x∈X and y∈Y.

Entropy measures are traditionally given in log units. Thus, when the space of possibilities increases multiplicatively, entropy increases linearly. While this property has practical and theoretical advantages in information theory, it is detrimental to WDT, which must handle both additive and multiplicative relationships, e.g. Semantic Formula. Because addition does not have a clear analogue in logarithmic space, there is no way of transforming these formulas into log units.

Therefore, for the purposes of Wordlist Distortion Theory, conditional entropy needs to be converted into linear space through exponentiation with the appropriate base in (2). The resulting value conveniently corresponds to the average number of possible values for variable X, given knowledge about variable Y. If X and Y comprise the phoneme inventory of a language, where elements of Y are the segments adjacent to elements of X, then the number of possible values for

[2] This could be either the preceding or the following environment. I do not know of a theoretical reason to prefer one over the other. However, it should be noted that the preceding environment consistently yields higher entropy values for the Austronesian data in Chapter 4 (Trussel & Blust, 2010) as well as for the CELEX corpus of English (Baayen et al., 1995). The *Python* script used in the Chapter 4 case study uses an average of the two.

X given Y is identical to the definition of mean homotopy $\hbar$. Thus, for entropy measured in bits, $\hbar = 2^{H(x|y)}$. As such, information theory already comes equipped with tools for calculating the average number of possible adjacent segments in a language. Measures derived from entropy through exponentiation are also commonly found in other fields, such as physics, biology, and economics (Jost, 2006).

The discussion so far has been focused on adjacent dependencies between segments, such as neutralization and rudimentary phonotactics. However, there are many non-random patterns in the distribution of non-adjacent segments as well, such as prosody and harmony. Non-adjacent segment dependencies will be largely ignored in this dissertation. This is a simplification motivated by the observation that most phonological processes in languages are local (Finley, 2011). In fact, phonological patterns that appear non-local on the surface are often argued to be local on some tier (Odden, 1994; Heinz, 2010; Gafos, 2013). Nevertheless, measures of entropy that take non-local dependencies into account have been applied to phonological data (Pimentel et al., 2020). The framework presented in this dissertation is compatible with any kind of entropy calculation, or indeed with estimates of $\hbar$ not based on information theory at all.

The only requirement for estimating $\hbar$ is a list of phonological entries for the language in question. Table 1 previews the case study in Chapter 4 by presenting $\hbar$ estimates and bits/phoneme of an etymologically and phonologically diverse subset of the Austronesian language family. Bits/phoneme were calculated using the formula in (2), where the conditional entropy for both the preceding environment and the following environment were computed and averaged. The word boundary (#) was treated as a possible environment; but its entropy was not computed. As discussed, the mean number of homotopes $\hbar$ was calculated simply by exponentiating 2 to the bits/phoneme value (e.g. for Ilokano, $2^{3.2} = 9.17$).

Table 1: Select homotopy estimates for Austronesian

language	*branch*	$\hbar$	*bits/phoneme*
Ilokano	Philippine	9.17	3.20
Amis	Formosan	7.82	2.97
Tboli	Western	7.00	2.88
Hawaiian	Oceanic	6.11	2.61
Wetan	Central	5.43	2.44

For comparison, applying the conditional entropy formula to the CELEX corpus of English (Baayen et al., 1995), gives an *ℏ* of 10.41 or about 3.38 bits of information per segment. To interpret this homotopy value, one can say that a randomly chosen phonological position in English can on average be occupied by just over 10 segments with the adjacent context in mind. To interpret the bit count, one can say that, armed with knowledge about the adjacent phonological dependencies of English, one should be able to guess the following segment in a word after asking 3 to 4 yes-or-know questions, since each yes-or-no question decreases the number of possibilities by a factor of 2.

The method presented here gives entropy estimates roughly in-line with more nuanced methods found in the literature, i.e. those that consider non-local dependencies as well, as these tend to fall between 2 bits/phoneme and 4 bits/phoneme (Pimentel et al., 2020). A *Python* script for estimating *ℏ* from a wordlist based on adjacent dependencies using the conditional entropy formula in (2) is submitted along with this dissertation. The script forms one of the components of the computational implementation of this framework described in Chapter 4.

3.1.2 Word Complexity

Recall that the Synchronic Formula is used in Wordlist Distortion Theory to calculate the size of the synchronically compatible wordlist set *S* from word complexity *c*, i.e. the number or phonologically possible words in the language. In Section 2.1, where the Synchronic Formula is introduced, word complexity *c* is treated as a given and its calculation not discussed. This gap is remedied in the current section.

Just as it can be assumed that each slot in a wordlist can be filled with a possible word, it can be assumed that each slot in a possible word can be filled with a homotopic segment. This use of homotopic segments follows from the definition of homotopy in the previous section; the average number of homotopes is the average number of segments that can appear in a given position. As such, the number of possible words can be expressed through a combination of word-length and the mean number of homotopes in the language, as in (3), where mean word-length is used.

(3)
$$c = \hbar^{w}$$

Word Complexity Formula: Formula for estimating the number of possible words c, using the mean word-length w and the mean number of homotopes $\hbar$.

Using the formula in (3) to calculate word complexity for the CELEX corpus of English (Baayen et al., 1995) yields $10.4^{7.1} = 16{,}413{,}813$. As could be expected, this number is far larger than the 45,400 unique strings actually found in the corpus. In other words, most (99.7% in this example) of the possible phonological strings predicted by the homotopy calculation are unattested. In the same way, permuting a wordlist explores only a tiny subset of the synchronically compatible set. For many studies in computational historical linguistics, particularly those employing multilateral comparison, this does not pose an issue, as phonologically possible but unattested strings (and phonologically possible but unattested wordlists) are not part of the general framework (Baxter, 1995; Kessler & Lehtonen, 2006; Kassian et al., 2015; Ceolin, 2019). However, for WDT, where the size of the synchronically compatible set plays a role in the calculation, a theoretical approach to approximating randomness is required instead.

In all likelihood, the Complexity Formula still gives an underestimation of the number of possible words because using the mean word-length to estimate the number of possible words *c* carries with it the assumption that only words of mean length are grammatical. Therefore, strings that are shorter or longer than the mean are effectively ignored. How many actual possible strings of different length are possible in a language remains an open question. In practice, even though languages exhibit words of different length, word-length and frequency are inversely correlated following roughly a Zipfian distribution (Sigurd et al., 2004). In other words, there always exists some number *n* such that no word of length *n* or greater is found in the language. Moreover, strings of length *n* - 1 are expected to be marginal, i.e. occurring a handful of times or perhaps even once. It is not clear how to incorporate such rare words into the calculation or, by extension, word length in general. For instance, it is likewise not clear how to account for the fact that different word lengths are attested at different rates. For the current dissertation, I choose to employ the simplified approximation in (3), where only words of mean word length are considered.

It should be highlighted that an atheoretical approach to quantifying $\hbar$ and c, such as the one employed in this dissertation, cannot distinguish between accidental and systematic gaps. The absence of an adjacent pair of segments in the data is treated as a phonotactic restriction. It follows that for smaller data sets, word complexity c and total wordlist length t are strongly correlated, as additional data is likely to provide additional adjacent contexts. Therefore, it is imperative that the dataset used for training is large enough to accurately depict the phonology of the daughter language. It has been argued that a sample of 100 or 200 words, as is convention in historical linguistics, is sufficient for a reasonable approximation of a language's segmental distribution for the purposes of identifying recurrent sound correspondences or performing a lexicostatistic analysis (Zhang & Gong, 2016). Thus, we can infer that the use of smaller datasets may result in an underestimation of $\hbar$ (and therefore c), i.e. an increase in type II error or a false negative.

3.2 Mergers

Sound change is the most studied type of diachronic transformation in the literature. In fact, traditional comparative reconstruction is often concerned primarily with assembling segment correspondences between wordlists as evidence for individual sound changes and subgrouping (Fox, 1995; Ross & Durie, 1996).

In contrast, computational alternatives to the comparative method, e.g. Bayesian phylogenetic analysis and multilateral comparison, do not engage with sound change. Frameworks that tally the proportion of shared cognates (Swadesh, 1955; Miller, 1984; Greenhill et al., 2017), or alternatively the proportion of typological features (Sicoli & Holton, 2013) or syntactic features (Longobardi et al., 2013), do not incorporate segmental information into the model altogether. As such, these approaches are entirely divorced from sound change.

Approaches similar to multilateral comparison do incorporate segmental information into the model (Baxter, 1995; Kessler & Lehtonen, 2006; Holman et al., 2011) but the connection to sound change in such models is implicit. All things being equal phonetic distance between an ancestor and a descendant is expected to correlate with the degree of sound change in the history of the language. In fact, this is the motivation behind such frameworks. Although there is a strong correlation between phylogenies inferred through multilateral comparison and those deduced through the comparative method (Pompei et al., 2011), there is, to my knowledge, no research connecting phonetic distance as measured using multilateral comparison and the number of sound changes in the history of the language. In fact, the case study in Chapter 4 of this dissertation also finds that these two are largely independent.

For the purposes of this dissertation, sound change is assumed to come in one of three forms, as listed in (4). Sound changes can be unconditioned (4a), or conditioned in one of two ways, by the preceding environment (4b), or by the following environment (4c). No other conditioning environments are possible in this model.

(4) *Sound Change Templates*
(a) $x > y$ (b) $x > y\ /\ \text{–}z$ (c) $x > y\ /\ z\text{–}$

In this dissertation, the inputs to these templates are individual segments, not segment sets, i.e. phonological features. WDT is compatible with sound changes that operate on phonological features. However, in the interest of presenting the simplest working model, only segment-to-segment mappings are considered here. This simplification has the added benefit of being less theory-specific, shirking concerns about whether features are universal or emergent, binary or privative, hierarchical or linear, etc. Nevertheless, some ideas about incorporating phonological features into the framework are discussed in Section 3.5.

It may seem at first that the templates in (4) are insufficient for describing segmental change in languages. After all, the templates do not directly include changes conditioned by both the preceding and following environments, long-distance effects, prosody, and phonotactics. However, it can be shown that more complex changes can always be restated as a series of simpler changes. Thus, at least in terms of empirical coverage, the templates in (4) are sufficient. Let us take some time here to demonstrate that this is indeed the case with some examples.

Changes conditioned both by the preceding and following environment can be restated as a sequence of three changes. The first is conditioned by the preceding environment; the second is conditioned by the following environment. The last change removes any output of the first change that is not also an output of the second. The example in (5) shows intervocalic voicing converted into three changes in this way. For this example, assume that [z] is not a segment of the language prior to the changes in question.

(5) *Bi-directional Changes Simplified*

One Complex Change	*Multiple Simple Changes*
s > r / a–a	s > z / a–
	z > r / –a
	z > s

Long distance changes can be restated as a sequence of multiple local changes. Some of these transfer contrast onto an adjacent position; others erode contrasts created by earlier changes. The example (6) shows regressive backness harmony converted into three changes in this way. For this example, assume that the language prohibits consonant clusters. Additionally, recall that the current dissertation deals with segment-to-segment mappings only. As such, a complete description of backness harmony would require multiple changes for different inputs and intervening consonants.

(6) *Long-distance Changes Simplified*

One Complex Change	*Multiple Simple Changes*
u > y / –li	l > l^j / –i
	u > y / –l^j
	l^j > l

The conversion in (6) demonstrates that a contrast between two words can be relocated to any position in the word. For example, after the changes in (6) apply, strings originally differing only at the right word-boundary, e.g. [buli] and [bulu], also differ elsewhere, e.g. [byli] and [bulu]. It follows that sound change can independently manipulate pairs of words so long as the inputs differ in at least one segment. The effects of all long-distance transformations can be relayed with the compounding of one or more strictly local ones. Any word can be converted into any shape and any wordlist into any other wordlist through local sound change alone. The only caveat is that homophonous forms in the mother wordlist cannot be disambiguated in the daughter using sound change.[3] In all other situations, sound change, even the limited version presented here, is an all-powerful transformation. Given enough sound changes, it is possible to derive even unrelated wordlists from a common ancestor.

[3] This is not an issue in practice, since it is almost always the case that the researcher posits the mother wordlist in accordance with existing daughter wordlists. Therefore, there is no reason to posit homophonous forms in the mother for non-homophonous forms in the daughter.

This conclusion is in contrast to the implicit attitude toward sound change in the literature. Perhaps because sound change is regular and grounded in physiology (articulation), it rarely sparks doubt in the evaluation of comparative reconstructions. Controversial language groupings and reconstructions are most often criticized for cherry-picking of data (i.e. lexical replacement in Wordlist Distortion Theory), semantic-overpermissiveness (i.e. semantic change In Wordlist Distortion Theory), arbitrary morphological boundaries, faulty data, and general flaws in the methodology (Baxter, 1995; Vovin, 2005; Blust, 2014). However, to my knowledge, never is a reconstruction considered unconvincing due to the number of sound changes proposed. In fact, sound change, so long as it is properly evidenced and justified, tends to be viewed as ‘free’ in historical linguistics literature, in that any amount is permissable, at least implicitly (Ringe, 1999; Harrison, 2003; Hock, 2009; Hoenigswald, 2019). A probabilistic evaluation of reconstruction reveals that sound change is no different from lexical replacement or semantic change, in the sense that it increases the arbitrariness of the reconstruction. In fact, for every mother-daughter pair, there must exist an upper bound to the number of sound changes beyond which a reconstruction between the two becomes statistically unreliable.

3.2.1 Mergers and Sound Change Outputs

Mergers are contrast-eroding sound changes; after a merger, two originally contrastive phonological forms may become homophonous. In theory, every merger yields multiple homophonous pairs. In practice, such a pair may not be present in the data. For the purposes of WDT, a sound change is treated as a merger regardless of whether the homophonous pairs of words it yields are actually attested in the wordlist. In other words, a sound change is treated as a merger if it is potentially neutralizing.

Alternatively, a merger can be thought of as a sound change whose output is already a member of the segmental inventory and is grammatical in the phonological position in question. The input of a merger is either absent from the daughter language, i.e. *mother-exclusive*, or at least absent from the daughter language in the phonological position in question. Therefore, a merger can be thought of as converting a wordlist that is not synchronically compatible with the daughter to one that is synchronically compatible by turning an illicit segment or sequence of segments into an already existing segment or sequence of segments respectively.

An example of a merger may be observed in the history of Russian in (7).[4] Note that vowel reduction is omitted from the transcription.

(7) *East Slavic Wordlists*

gloss	*Proto-East Slavic*	*Russian*	*Russian'*
cripple	*kalʲæka	kalʲeka	kalʲika
log	*polʲæno	polʲeno	polʲino
spring	*vʲesna	vʲesna	vʲesna
village	*sʲelo	sʲelo	sʲelo

One can derive the Russian wordlist from the Proto-East-Slavic wordlist through a single merger [æ] > [e]. The vowel [æ] is mother-exclusive (i.e. does not exist in Russian), and any wordlist containing it cannot be synchronically compatible with modern Russian. Therefore, a reconstruction from Proto-East Slavic to Russian must posit at least one sound change in order to remove [æ]. All wordlists derivable from Proto-East Slavic through one merger are diachronically local to the Russian wordlist. In wordlists both synchronically compatible with and diachronically local to Russian, the output of the merger is any segment that does not yield phonologically illicit forms in Russian, not necessarily [e]. The merger [æ] > [i] produces Russian′, a wordlist synchronically compatible with and diachronically local to Russian, since [i] is an attested segment in the environment in question. The total number of possible merger outputs for [æ] that introduce both synchronically compatible and diachronically local wordlists is equal to the number of segments grammatical in the position occupied by [æ] in Proto-East-Slavic. By definition, this is also approximately the average number of homotopes in the language.

Outputs of mergers are independent of one another. To illustrate that this is the case, observe the Pali and Sanskrit wordlists in (8) (Campbell, 2013:46, adapted). Both the post-alveolar [ʃ] and

[4] The exact pronunciation of the Proto-East-Slavic vowel here transcribed as [æ], also known as yat’ after the Cyrillic ѣ or ě, is unknown. It is known that this was likely [+low] and [-back] vowel which was contrastive with other vowels up until sometime in the medieval period, whereupon it merged with existing [e] in Russian (Lindgren, 1989).

the retroflex [ʂ] are mother-exclusive, i.e. present in Sanskrit but not Pali, and have merged with the alveolar [s].

(8) *Pali and Sanskrit*

gloss	*Sanskrit*	*Pali*	*Pali'*
hare	ʃaʃa	sasa	sasa
hair	ʃaraɳa	saraɳa	saraɳa
fault	doʂa	dosa	dora
slave	daʂa	dasa	dara
mind	manas	manas	manas
son	suta	suta	suta

A wordlist diachronically local to Pali with respect to the reconstruction must also be derived from Sanskrit using two mergers. A wordlist synchronically compatible with Pali may only exhibit merger outputs which are grammatical in the given environment, i.e. homotopes. The outputs of these mergers can either be the same segment, as in actual Pali [s], or two different segments, as in Pali′, which underwent [ʃ] > [s] and [ʂ] > [r]. Crucially, the choice of output of one merger does not depend on the choice of output for the other.

Let φ be the number of mergers and $\hbar$ the average number of homotopic segments in the language. Because the merger outputs are independent, the number wordlists diachronically local to and synchronically compatible with the daughter introduced by φ mergers can be calculated using (9).

(9)

$$|D \cap S\,|_{\varphi} = \hbar^{\varphi}$$

Merger Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by φ mergers with mean number of homotopes $\hbar$.

For each unconditioned merger posited by the reconstruction, there are on average $\hbar$ sound change outputs that yield synchronically compatible and diachronically local wordlists. Assuming the daughter wordlist is large enough to evidence the phonological contrasts of the language, all of these wordlists are distinct from one another.

The Merger Formula in (9) implies that the number of synchronically compatible and diachronically local wordlists as well as $P(D|S)$ grows exponentially with an increase to the number of mergers posited. In contrast to lexical replacement semantic change, whose number is bounded by the size of the wordlist, there is no limit on methodological grounds alone to how many mergers can be posited in a reconstruction. However, there does exist a limit to the number of mergers which can be posited in a reconstruction before $P(D|S) = 1$. At this point, the number of changes is estimated to be sufficient to convert any wordlist into any other wordlist, and positing more changes would not be productive. I leave the calculation of this limit to future research.

3.2.2 Mergers and Sound Change Environments

The discussion so far has been limited to unconditioned sound change, i.e. sound change that affects segments regardless of their phonological environment. However, in real data, sound changes are sometimes conditioned. Moreover, some types of sound change, such as palatalization, assimilation, and vowel harmony, are always conditioned. This section presents a method of incorporating sound change environments into Wordlist Distortion Theory. The Merger Formula in (9) does not need to be updated, because, as is demonstrated in this section, variation in sound change outputs is independent of the variation in sound change environments.

Just as in the case of sound change outputs, identifying which environments of sound change yield synchronically compatible wordlists is not a simple matter. For example, to derive Korean from a hypothesized earlier version of the language in (10), a reconstruction may posit the sound change n > l / l_ (progressive assimilation).[5] This is a merger, as its output was already a segment in the relevant environment in the language. Stated differently, the change eroded the contrast between the sequences [ln] and [ll].

(10) *Korean Assimilation*

gloss	*Middle Korean*	*Korean*	*Korean'*	*Korean"*
fire moth	pulnabi	pullabi	puldabi	punnabi
New Year’s	sʌlnal	sʌllal	sʌldal	sʌnnal
wine song	sulnore	sullore	suldore	sunnore
dwelling	sallim	sallim	sallim	sallim

[5] This change is evidenced primarily by morphophonological alterations in Modern Korean, as in [k^hallal] ‘knife edge’ from [k^hal] ‘knife’ and [nal] ‘blade’. Although it is not clear when exactly the change took place, it appears to be a part of a general trend of contrast loss in coda position beginning as early as the Middle Korean period (Yi & Ramsey, 2011). Thus, the Middle Korean forms in (10) do not necessarily represent the pronunciation at a given point in time.

Recall that replacing the output of the sound change with a homotopic element produces a synchronically compatible and diachronically local wordlist. This remains true for conditioned sound changes. For example, Korean′ is derived from Middle Korean through a sound change that targets the same segment in the same environment as the original change but yields a (different) homotopic segment as output, n > d / l_ (progressive denasalization). One can demonstrate that the resulting wordlist is synchronically compatible with Korean by confirming that the sequence [ld] is attested in Korean, as in [t͡ɕʌlde] ‘never’.

Just as unconditioned mergers remove mother-exclusive segments from the wordlist, conditioned mergers remove mother-exclusive sequences of segments from the wordlist. This can be achieved by altering either segment in the sequence. While Korean and Korean′ alter the original target, Korean″ alters the original environment through the sound change l > n / n_ (regressive assimilation).[6] The resulting wordlist is synchronically compatible with Korean because [nn], the sequence it produces, is attested in the language, as in [ʌnni] ‘older sister’.[7]

To generalize, synchronically compatible and diachronically local wordlists can be derived from the mother through mergers in two different ways. The two ways are presented schematically in (11), where *h* stands for any homotopic segment.

(11) *Conditioned Sound Change Alternatives*

Original	*Output Target*	*Context Target*
x > *y* / _z	*x* > h / _z	z > h / *x*_
x > *y* / z_	*x* > h /z_	z > h / _*x*

A diachronically local and synchronically compatible wordlist can be reached through a merger that targets the original input, where the environment is the same and the direction is the same as the original (preceding or following), listed as *output target* in (11). Korean′ in (10) is an example of such a wordlist, cf. n > d / l_ (denasalization) and the original n > l / l_ (progressive assimilation).

[6] Incidentally, this change is also attested in Korean, though it is mostly restricted to Sino-Korean loans, such as [ɨmunnon] ‘phonology’ from Chinese 音韻論, c.f. modern Mandarin pronunciation [jin^{1}jun^{4}lun^4] where the [nl] is faithfully preserved (Um, 2003).

[7] It should be noted that the sequence [nn] is extremely rare in native Korean vocabulary. However, it is well attested in loans from Chinese and other languages (Um, 2003).

Alternatively, a diachronically local and synchronically compatible wordlist can be reached through a merger that targets the original environment, where the environment is the original output and the direction is reversed, listed as *context target* in (11). Korean″ in (10) is an example of such a wordlist, cf. l > n / n_ (regressive assimilation) and the original n > l / l_ (progressive assimilation). No other methods exist for removing a mother-exclusive sequence of segments to produce a wordlist synchronically compatible with the daughter.

Counting the number of synchronically compatible and diachronically local wordlist for one of the options in (11) is the same as counting the number of wordlists introduced by unconditioned mergers using the formula in (9). For both output target mergers and context target mergers, the number of outputs yielding synchronically compatible wordlists is on average equal to the number of homotopic segments. The only difference between conditioned and unconditioned mergers is the choice between output target and context target for conditioned changes. In other words, a conditioned sound change posits double the number of diachronically local and synchronically compatible wordlists as an unconditioned one, since for each conditioned sound change the choice must be made between an output target and a context target. Therefore, for each sound change environment proposed in the reconstruction, the number of synchronically compatible and diachronically local wordlists increases by a factor of 2.

Let ζ be the number of sound change environments for sound change posited in a reconstruction. The estimated number of diachronically local wordlists that are also synchronically compatible introduced by ζ environments is formalized in (12).

(12)

$$|D \cap S\,|_{\zeta} = 2^{\zeta}$$

Environment Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by ζ conditioning environments of mergers.

Note that the Environment Formula is independent of the Merger Formula in (9). In other words, each additional merger increases $|D \cap S|$ by a factor of $\hbar$ and an additional factor of 2 if it is conditioned. The number of synchronically compatible and diachronically local wordlists introduced by conditioned mergers is simply the product of (9) and (12), i.e. $\hbar^{\varphi}2^{\zeta}$.

The Environment Formula in (12) bears some similarity to the Merger Formula in (9). The Merger Formula implies that the number of diachronically local and synchronically compatible wordlists increases exponentially with the number of mergers. Likewise, the Environment Formula implies that the number of diachronically local and synchronically compatible wordlists increases exponentially with the number of sound change environments, i.e. the number of conditioned changes. Because the number of diachronically local and synchronically compatible wordlists is the numerator in the General Formula, $P(D|S)$ also increases exponentially as a factor of both.

The only difference between the Environment Formula and the Merger Formula is the base in the equation. The number of diachronically local and synchronically compatible wordlists, as well as $P(D|S)$, increases by a factor of $\hbar$ for every merger and an additional factor of 2 for every conditioned change. In real languages, $\hbar$ is expected to be substantially higher than 2 (refer Table 1 for some examples). In other words, the increase to $P(D|S)$ from positing mergers is expected to be higher than the increase from positing sound change conditions. As such, given the choice between two unconditioned mergers and one conditioned one, reconstructions should opt for the latter *ceteris paribus*.

3.2.3 Sound Change and Lexical replacement Trade-Off

Although the comparative method offers some guidance during the reconstruction process, many decisions are still left entirely to the researcher. For instance, the comparative method says nothing about whether one conditioned sound change is better than two unconditioned ones, or if a merger and a semantic shift are better than two mergers, or if three mergers are better than a one instance of lexical replacement. Fortunately, Wordlist Distortion Theory is perfectly suited to address such questions.

Perhaps most often, the researcher is tasked with weighing lexical replacement against sound change. For example, observe the data from Svan and Proto-Kartvelian in (13) (Fahnrich, 2007).

The first three entries (‘carrion’, ‘wooden plate’, and ‘fresh grass’) in Svan are relatively easy to derive from their Proto-Kartvelian analogues.[8]

(13) *Svan and Proto-Kartvelian*

gloss	*Proto-Kartvelian*	*Svan*
carrion	*ʒor	ʒwer
wooden plate	*gob	gweb
fresh grass	*mol	mwel
flea	*grt͡s’q’il	zesq’

However, the Svan form for ‘flea’ [zesq’] appears radically different from the Proto-Kartvelian analogue [grt͡s’q’il], as the two differ in length and share only a single segment in [q’]. The researcher may choose to ignore the entry ‘flea’ altogether, i.e. conclude that the two are unrelated and posit one instance of lexical replacement. Alternatively, the researcher may choose to derive the Svan form from the Proto-Kartvelian analogue through some number of additional sound changes, those that are not required elsewhere in the wordlist.

Traditionally, these two options are weighed against each other intuitively. However, Wordlist Distortion Theory is perfectly suited to handle the comparison probabilistically. A simple solution is to calculate *P*(*D*|*S*) for both reconstructions and choose the reconstruction with the lower *P*(*D*|*S*) value. After all, *P*(*D*|*S*) calculates the likelihood that a wordlist evidences the same number and type of sound changes. Because the transformations are orthogonal to the data, the framework is equipped to compare competing reconstructions for the same dataset.

However, calculating *P*(*D*|*S*) at every step in this way may become cumbersome. It would be more convenient if the effects of lexical replacement can be compared to the effects of sound change directly. WDT is perfectly suited for this task as well. As it happens, the similarities between the Merger Formula and the Replacement Approximation make a direct comparison between the two transformations relatively trivial.

When weighed against replacement, one or more mergers are justified probabilistically if their increase to *P*(*D*|*S*) is smaller than the increase caused by lexical replacement applied to the words that the mergers serve to explain, i.e. the words that have undergone only the mergers in question. In effect, sound change and lexical replacement are competing to derive some number

[8] Employing the simplified sound change template introduced in at the beginning of Section 3.2, two changes are required: *o > w and ∅ > e / w_

of word-pairs. Whenever the number of synchronically compatible and diachronically local wordlists is smaller when the words in question undergo lexical replacement as opposed to sound change, the proposed sound changes are not justified probabilistically. Contrariwise, whenever the number of synchronically compatible and diachronically local wordlists is smaller when the words in question undergo sound change as opposed to lexical replacement, the sound changes are justified probabilistically.

As per (9) and (12), the number of diachronically local and synchronically compatible wordlists introduced by conditioned mergers is $\hbar^{\varphi}2^{\zeta}$. As per the Replacement Approximation in Section 2.2.2, the number of diachronically local and synchronically compatible wordlists introduced by lexical replacement is the product of two parts. The larger part, c^{l}, is the number of alternative outputs to the replacement; the smaller part, $\binom{t}{l}$, is the number of possible positions of replacement in the wordlist. If $\hbar^{\varphi}2^{\zeta}$ is smaller than the change to the product of these two parts after the subset of words explained by sound change ceases to undergo lexical replacement, then the proposed sound changes serve to decrease *P*(*D*|*S*). *A fortiori*, this is also true if $\hbar^{\varphi}2^{\zeta}$ is smaller than the change to either of these parts. As such, it is most convenient to ignore $\binom{t}{l}$ altogether, since it is typically much smaller and does not lend itself to simplification easily.

We can use the term Δl to refer to the change in number of replacements *l*. Thus, the effect on *P*(*D*|*S*) after the introduction of φ sound changes, ζ of which are conditioned, is negative if $\hbar^{\varphi}2^{\zeta} < c^{-\Delta l}$, where the left side of the inequality refers to the change to *P*(*D*|*S*) induced by the derivation using mergers, while the right side refers to the change induced by replacement. If needed, the inequality can be simplified further with a few safe assumptions. The mean number of homotopes in natural language must be higher than 2, as there is no language that simultaneously has fewer than two consonants and fewer than two vowels. Therefore, the inequality remains true if 2 is replaced by $\hbar$. Furthermore, because conditioned sound changes incur a greater penalty on *P*(*D*|*S*) than unconditioned sound changes, a conservative estimate can hold all sound changes to be conditioned. The updated inequality states that the effect on *P*(*D*|*S*) after the introduction of φ sound changes is also negative if $\hbar^{2\varphi} < c^{-\Delta l}$.

Conveniently, the Word Complexity Formula in (3) can be used to replace *c* on the right side of the inequality with $\hbar^{w}$, a combination of the mean number of homotopes and average word-length. This transformation is useful because ensures that both sides of the inequality are expressed with $\hbar$ raised to some power, $\hbar^{2\varphi}$ for the left side of the inequality $(\hbar^{w})^{-\Delta l} = \hbar^{-\Delta lw}$ for the right. Taking the log base $\hbar$ of either side and dividing by 2 yields the statement in (14).

(14)

$$\varphi < \frac{-\Delta lw}{2}$$

Sound change vs lexical replacement comparison: Rule-of-thumb formula which, if satisfied, guarantees that the suggested sound change(s) decrease *P*(*D*|*S*).

Stated informally, if the number of a proposed series of mergers is lower than half the average word-length multiplied by the number of words in the wordlist that the sound changes serve to explain, then the introduction of the sound changes into the reconstruction decreases *P*(*D*|*S*). Because, as will be shown later in the chapter, the effect of mergers on *P*(*D*|*S*) is larger than that of other types of sound changes, this rule-of-thumb can be applied to sound change in general.

Returning to the Svan example in (13), *P*(*D*|*S*) is guaranteed to decrease if the number of (additional) sound changes required to derive [zesq’] ‘flea’ is less than or equal to $\frac{-\Delta lw}{2} = \frac{4}{2} = 2$. This does not appear possible, as roughly 7 sound changes are needed to derive the Svan form from the Proto-Kartvelian analogue.[9] Therefore, at least for the data in (13), the rule-of-thumb does not indicate that a reconstruction deriving Svan [zesq’] form from Proto-Kartvelian [grt͡s’q’il] is less arbitrary than one that treats the entry as a loan.

It should be stressed that, for a different dataset of the same languages, one where the changes used to explain ‘flea’ are also employed elsewhere in the wordlist, the conclusion could be different. For instance, imagine that the Proto-Kartvelian form *mart͡s’q’w ‘strawberry’ and its Svan analogue [basq’] were included in (13). Depending on the analysis, the word-pair may require two additional sound changes not evidenced elsewhere, *m > b /_a and w > Ø. The remaining sound changes exhibited by Svan [basq’] ‘strawberry’ are also required for [zesq’] ‘flea’. As such, the two forms now require approximately 9 sound changes between them but reduce the number of replacements by two rather than one. As it stands, $\frac{-\Delta lw}{2} = \frac{8}{2} = 4$, and the

[9] The history of this form in Svan has been described as follows: *grt͡s’q’il > *git͡s’q’il > *ʒisq’ > zisq’ (Fahnrich, 2007:113). The number of sound changes this progression implies depends on the analysis, though it is difficult to imagine an analysis requiring fewer than 7.

rule-of-thumb still does not indicate that the sound changes are justified, though this may not be the case with more words exhibiting the sound changes in question.

The rule-of-thumb formula in (14) can be applied to the entire set of mergers, or to smaller groups or even individual changes. In the most basic case, the formula implies that positing one sound change per word in the wordlist is always justified, since the average word length in natural language data should always be greater than 2, as is positing two changes per two words, three per three, etc. In fact, the formula implies that positing one sound change (or fewer) for every two segments in the words that undergo the sound changes is generally expected to decrease $P(D|S)$.

The inequality in (14) is not an accurate estimate of the expected change to $P(D|S)$; it is a convenient and conservative shorthand that is guaranteed to decrease $P(D|S)$ if true but not necessarily increase it if false. In other words, because the inequality overestimates the effect of sound change on $P(D|S)$ while underestimating the effect of lexical replacement, it is still possible for reconstructions to decrease $P(D|S)$ while positing more sound changes than half the average word length multiplied by the number of words evidencing the changes. This is also true of the analysis of Svan [zesq’] in (13), where a complete calculation of $P(D|S)$ may yield a different result. The rule-of-thumb does not confirm that the Svan word is best analyzed as being derived from Proto-Kartvelian through sound change, but it does not deny this either.

The conclusion that the upper bound for the number of sound changes in a statistically reliable reconstruction is linearly correlated with word length is not an obvious one. The rule-of-thumb formula is useful in many situations during the reconstruction process. More generally, the formula serves to illustrate of how approaching the task of comparative reconstruction using Wordlist Distortion Theory can reveal insights about the reconstruction process.

3.3 Shifts

This section presents methods for incorporating contrast-preserving sound changes into the current framework. Additionally, this section also reveals some insights about sound change and comparative reconstruction that are inaccessible without a probabilistic evaluation metric. In particular, WDT implies that the effect of contrast-preserving changes is incredibly minor. The first contrast-preserving change does not increase $P(D|S)$ at all, while the increase from even across-the-board contrast-preserving sound change can be thought of as negligible.

Contrast-preserving sound changes are called *shifts* in this dissertation. Specifically, after a shift, a contrast in two phonological forms is always preserved in the same position in the word, although the shift alters the segments in question. Like in the case of a merger, the input of a shift is a mother-exclusive segment (or sequence of segments). Unlike in the case of a merger, the output of a shift is a member of the daughter's segmental inventory but not the mother's, i.e. *daughter-exclusive*. An example of a shift occurred in the history of Czech, as can be seen in (15). The Proto-West Slavic voiced velar stop [g] became the fricative in Old Czech [ɦ], a segment absent from the language at the time.

(15) *West-Slavic Wordlists*

gloss	*Proto-West Slavic*	*Old Czech*
mountain	*gora	ɦora
castle	*grad	ɦrad
foot	*noga	noɦa
God	*bog	boɦ

The reconstruction for the data (15) requires a single shift [g] > [ɦ]. At a minimum, any reconstruction on the same data must remove [g], which is mother-exclusive,[10] and introduce [ɦ], which is daughter-exclusive. There is no way to do both in a sound change other than [g] > [ɦ]. Thus, there is only one reconstruction that yields a synchronically compatible and diachronically local wordlists to the Old Czech data in (15); it is the one deriving the Old Czech wordlist itself. In other words, the likelihood of a randomly generated wordlist substantiating a reconstruction of equal size is the likelihood of picking the Old Czech wordlist from the set of all possible Old Czech wordlists.

The Old Czech example demonstrates that the outputs of shifts which yield synchronically compatible wordlists are limited to daughter-exclusive segments. In other words, a reconstruction on the data in (15) yielding a synchronically compatible wordlist using a single shift must use [ɦ] as the shift output. This is in contrast to mergers, where any homotopic segment in the output produces a synchronically compatible wordlist. A peculiar consequence of

[10] Note that *g* was reintroduced into Czech through loans at a later point, e.g. [garaːʒ] 'garage' and [orgaːn] 'organ'.

this fact is that a reconstruction requiring a single shift, as in (15), is mathematically equivalent to one that requires no shifts (one where the mother and daughter are identical). This is because in both cases, the set of synchronically compatible and diachronically local wordlists contains a single member, the daughter language itself. As such, the first shift posited in a reconstruction is ‘free’, as its introduction does not alter $P(D|S)$. Stated differently, the Old Czech wordlist in (15) is as close to Proto-West Slavic wordlist as it can be, as no other wordlist that shares its phonological properties can be equally close or closer.

To understand what happens when more than one shift is required, let us turn to data from Hawaiian and Proto-Oceanic in (16) (Trussel & Blust, 2010). In the history of Hawaiian, [p] became [h] and [t] became [k].[11] Thus, there are two mother-exclusive segments and two daughter-exclusive segments, meriting two shifts. There are two ways of mapping the mother-exclusive segments onto the daughter-exclusive segments. Therefore, there exist two reconstructions deriving wordlists synchronically compatible with and diachronically local to Hawaiian. In the first reconstruction, *p > h and *t > k derive Hawaiian. In the second reconstruction *t > h and *p > k derive Hawaiian′.

(16) *Hawaiian and Proto-Oceanic*

gloss	*Proto-Oceanic*	*Hawaiian*	*Hawaiian'*
fire	*api	ahi	aki
which	*pea	hea	kea
hence	*atu	aku	ahu
die	*mate	make	mahe

The set of diachronically local and synchronically compatible wordlists contains two wordlists equidistant from the Proto-Oceanic wordlist, both requiring two shifts. The two shifts remove two mother-exclusive segments in [p] and [t] and introduce two daughter-exclusive segments in [h] and [k]. There are only two ways of doing so in two shifts. As a result, $P(D|S)$ in this case is equal to 2 divided by the number of wordlists in the synchronically compatible set.

The pattern is similar for reconstructions requiring more shifts. Let χ stand for the number of shifts posited in the reconstruction. For every shift, there is one mother-exclusive segment and one daughter-exclusive segment by definition. For all diachronically local and synchronically

[11] The change *p > h, which was actually a series of changes, could not have been a shift since [h] was a sound already in the language. The change *t > k was, in fact, a shift, but a part of a longer chain (Lynch et al., 2002). The changes are presented as shifts for illustrative purposes.

compatible wordlists, the number of shifts must equal to χ. Any fewer and the reconstruction fails to remove a mother-exclusive segment, and the resulting wordlist is not synchronically compatible with the daughter; any more and the reconstruction is too large and the resulting wordlist is not diachronically local to the daughter.[12]

The number of diachronically local and synchronically compatible wordlists is equal to the number of ways to map mother-exclusive segments to daughter-exclusive segments. Mathematically, this is equivalent to computing the number of *bijective functions*, which is also the number of permutations to the input, as illustrated in Figure 7.

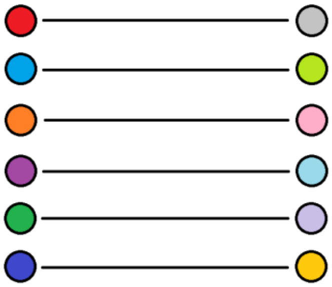


Figure 7: *Illustration of a bijective functions.* Every element in the domain (left) corresponds to exactly one element in the codomain (right) and *vice versa*. For our purposes, elements in the domain correspond to mother-exclusive segments, while elements in the codomain correspond to daughter-exclusive segments.

The number of diachronically local and synchronically wordlists given some number of shifts can be calculated using the formula in (17).

(17)

$$|D \cap S\,|_{\chi} = \chi!$$

Shift Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by χ shifts.

The number of diachronically local wordlists introduced by shifts can be counterintuitively small. As a demonstration, imagine that one is comparing a hypothetical mother and daughter wordlist word by word. For simplicity, imagine that all words in both wordlists are monosyllables; imagine also that every monosyllable in both wordlists is of the shape *Ca*, where

[12] Positing too many shifts also yields a reconstruction that is not synchronically compatible with the daughter, as the additional shifts must introduce segments that are not part of the daughter's segmental inventory.

C is any consonant.[13] Finally, assume that every *C* is unique in both lists, as the example in (18). In other words, no onset appears more than once, either within a wordlist or across wordlists. It is reasonable to ask how many lines of the data it would take to establish a convincing resemblance with a given *P*(*D*|*S*), e.g. .05.

(18) *A Statistically Significant Resemblance*

gloss	*Mother*	*Daughter*
temple	fa	da
fire	ga	xa
horse	ta	ra
tiger	wa	na
dream	ma	ka
…	…	…

The data in (18) serves to highlight the difference between WDT and previous quantitative approaches in historical linguistics, namely multilateral comparison, since lexicostatistics does not deal with segments altogether. Computing pairwise similarity, as in multilateral comparison (Baxter, 1995; Kessler & Lehtonen, 2006; Kassian et al., 2015), will always reveal that no amount of data is enough to establish a relation between the two wordlists. If the consonants are taken to be random, then the mean pairwise distance of this example is simply the average distance between two random consonants calculated by the phonological distance metric. As such, multilateral comparison will always conclude that the two wordlists are as distant as an average pair of random wordlists, i.e. no evidence of genetic relation.

Rather surprisingly, a WDT analysis conducted on a reconstruction for the data in (18) suggests that the two wordlists in (18) are similar beyond coincidence. Each segment correspondence necessitates a single shift in the reconstruction, as the input is mother-exclusive and the output daughter-exclusive. Therefore, the first line implies f > h, the second g > x, etc. Using the Shift Formula in (17), for *n* lines of the dataset, $|D \cap S|_\chi = n!$. In other words, each line introduces the need for a new shift and also a new possible output for previous shifts.

Notice that the consonants are all homotopes of one another, while the vowels are fully predictable; so, for *n* lines of the dataset, $|S\ | = n^n$. In other words, each line introduces a novel

[13] The [a] nucleus is included to make the example at least somewhat reminiscent of real language data. Because the appearance of the segment is completely predictable from the point of view of either language, it should not be taken as evidence of genetic relatedness and does not affect any of the measures discussed in this dissertation. All of the arguments presented in this section apply to the onsets only.

possible word and a novel position for the word to appear in. Therefore, as per the General Formula, $P(D|S) = \frac{n!}{n^n}$, where *n* is the number of lines given. Not only does this mean that $\lim_{n \to \infty} P(D|S) = 0$, but it also means that *P*(*D*|*S*) decreases incredibly quickly. In fact, already for the five lines shown in (18) *P*(*D*|*S*) = .38, while for ten lines *P*(*D*|*S*) < .001. In other words, for the wordlists in (18) it is unlikely that a randomly generated daughter wordlist would be as or more easily derivable from the mother wordlist than the daughter itself.

For an intuition as to why this is the case, consider an unusual property of the data in (18): even though the phonological structure of the language is highly conducive to homophony, no homophones are attested (recall that the lack of homophones comes directly from the definition of the daughter set). In terms of the reconstruction, homophonous pairs in the daughter, which are not homophonous in the mother, require either mergers of lexical replacement, i.e. a greater increase to *P*(*D*|*S*). As is, while not as telling as matching instances of homophony, the matching instances of *non*-homophony in (18) are also unlikely. In fact, one can compute exactly how unlikely they are by using the Wordlist Distortion Theory.

It is not reasonable to expect researchers performing comparative reconstruction to develop an intuition about what proportion of matching non-homophony between mother and daughter signals relatedness beyond a reasonable doubt. This is especially true since the patterns of the type in (18) may still be probabilistically significant even if the wordlists exhibit some degree of homophony, so long as homophonous positions are correlated between the two wordlists or if the incidence of homophony is lower than expected. Additionally, the arguments hold for wordlists with entries that are longer and more varied in shape than the ones in (18). A heuristic similarity metric, be it pairwise similarity or proportion of cognacy, cannot tackle such examples. However, the example poses no issues for a derivation-focused probabilistic evaluation of comparative reconstruction, such as WDT.

3.3.1 Mergers and Shifts

Although the outputs of mergers do not interact with the outputs of other mergers, their outputs do interact with the outputs of shifts. Because both shifts and mergers take mother exclusive segments as input, it is not enough to compute the number of synchronically compatible wordlists derived from mergers and shifts separately and then multiply the two. Which mother-exclusive segments undergo mergers and which undergo shifts can vary. Let us refer to (19) for an example from Proto-Polynesian and Hawaiian to demonstrate that this is the case (Trussel & Blust, 2010).

(19) *Hawaiian and Proto-Polynesian*

gloss	*Proto-Polynesian*	*Hawaiian*	*Hawaiian'*
line	*afo	aho	aʔo
nose	*ihu	ihu	ihu
root	*aka	aʔa	aha
two	*rua	lua	lua
leaf	*lau	lau	lau

The simplest reconstruction deriving Hawaiian from Proto-Polynesian in (19) requires three transformations, two mergers and one shift, as listed in (20). Applying the Shift Formula and the Merger Formula separately, assumes that, to produce a synchronically compatible wordlist, the correspondent of proto-Polynesian *f and *r, the inputs of the original mergers, can be any homotopic segment in Hawaiian, while the correspondent of Proto-Polynesian *k, the input of the one original shift, must be the daughter-exclusive Hawaiian [ʔ]. However, because it is not predetermined which input undergoes which type of change, a synchronically compatible wordlist may exhibit a mapping from Proto-Polynesian *k to any homotopic segment, so long as the either *f or *r map to [ʔ], as in Hawaiian′ in (19).

(20) *Hawaiian reconstructions*

Hawaiian				*Hawaiian'*			
input		*output*	*type*	*input*		*output*	*type*
f	>	h	merger	f	>	ʔ	shift
r	>	l	merger	r	>	l	merger
k	>	ʔ	shift	k	>	h	merger

Therefore, to calculate the number of synchronically compatible and diachronically local wordlists, one must figure out the number of ways to map the mother-exclusive segments onto the daughter-exclusive segments, such that every mother-exclusive segment maps onto something and every daughter-exclusive segment is mapped onto at least once. In other words, the combination of mergers and shift must remove every mother-exclusive segment and introduce every daughter-exclusive segment to ensure that the resulting wordlist is synchronically compatible with Hawaiian. In mathematics this is known as counting the number of *surjective functions*, also known as ‘onto functions’, where each element of the codomain is paired with at least one element of the domain. This relationship is illustrated in Figure 8.

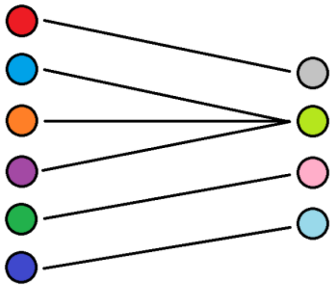


Figure 8: *Illustration of a surjective function.* Every element in the domain (left) corresponds to some element in the codomain (right). Every element in the codomain corresponds to at least one element in the domain. For our purposes, elements in the domain correspond to mother-exclusive segments, while elements in the codomain correspond to daughter-exclusive segments. Mother-exclusive segments that map onto the same daughter-exclusive segment, can be said to have undergone a merger. Daughter-exclusive segments that are mapped to only once, can be said to have undergone a shift.

The formula for calculating the number of surjective functions given a domain and codomain of particular sizes is well known in mathematics. However, a complete derivation is somewhat more involved than would be appropriate here. Inserting the appropriate linguistic variables yields (21). Stated informally, the formula calculates the total number of mappings (surjective or not) such that every mother-exclusive segment participates in a mapping, then subtracts all mappings where at least one of the daughter-exclusive segments is not mapped onto.

(21)

$$|D \cap S\,|_{\chi,\varphi} = \sum_{i=0}^{\chi} (-1)^i \binom{\chi}{i} (\hbar - i)^{\chi+\varphi}$$

Sound Change Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by χ shifts and φ mergers, with mean number of homotopes $\hbar$.

As a sanity check, if $\chi = 0$ the Sound Change Formula is the same as the Merger Formula in (9), as there is only one summand, and everything other than $\hbar$ and φ cancels out. Similarly, if $\varphi = 0$, the Sound Change Formula is the same as the Shift Formula in (17). This latter equality is more difficult to show algebraically, as it has to do with the *Stirling numbers of the second kind,* a way of counting the number of non-zero subsets of a set and its relation to formulas of the type in (21) (Boyadziev, 2018). Stated informally, it is possible to factor out the term $\chi!$ in the Sound Change Formula, and when $\varphi = 0$ the remainder of the equation is equal to 1, rendering the entire formula equal to $\chi!$, just as the Shift Formula in (17).

If a reconstruction posits either shifts or mergers but not both, the Shift Formula or the Merger Formula on their own are easier to manipulate then the combined formula (21). However, if a reconstruction posits both, then this is no longer an option. In this case, there are two reasonable courses of action: mergers and shift could be tallied separately, and their effect estimated using the Sound Change Formula in (21); or all sound changes could be as mergers, and their effect estimate using the Merger Formula in (9). Although, the latter solution introduces error into the calculation, treating all sound changes as mergers leads to a conservative rather than liberal estimate, i.e. an increase in type II but not type I error. Recall that mergers impart a greater increase on $P(D|S)$ than shifts and, as will be shown in Section 3.4, greater than the remaining type of sound change. The benefit to such a shortcut is the use of the much less unwieldy Merger Formula and the option to ignore sound change type altogether.

3.4 Links

The same set of sound changes can produce different outputs depending on the order of application. On the surface sound change interactions appear to pose an issue for Wordlist Distortion Theory. This section shows how sound change interactions can be accounted for in the framework while also introducing, *links*, the final type of sound change to be discussed in this dissertation.

Before talking about links, we must first talk about chain shifts or chains. In this dissertation, chain shifts are henceforth referred to as ‘chains’ to avoid confusion with shifts introduced in Section 3.3. In (22), two instances of vowel raising are applied in different order. In (22a) mid vowels raise and low vowels take their place, preserving contrast. In (22b) low vowels raise to become mid vowels and then raise once more along with the original mid vowels, eroding the contrast between the two vowels. The only difference between the two reconstructions is the order in which the two sound changes are applied.

(22) *Sound Change Ordering I*

(a)	*Contrast Maintained*			(b)	*Contrast Lost*		
	input	ta	te		*input*	ta	te
	e > i	--	i		a > e	e	--
	a > e	e	--		e > i	i	i
	output	te	ti		*output*	ti	ti

The ordering in (22a) is known as a chain. Chains are usually analyzed as a series of sound changes that occur in unison historically and are presumed to have a causal relationship with one another (Gordon, 2011). In this dissertation, a chain is defined more broadly to include any series of changes that, in its entirety, performs the same role as a merger or a shift, i.e. removes a mother-exclusive segment and optionally introduces a daughter-exclusive segment. A chain led by a merger removes a mother-segment, while a chain led by a shift both removes a mother exclusive-segment and introduces a daughter-exclusive segment. For instance, the chain in (22a) removes the mother-exclusive [a] and introduces the daughter-exclusive [i]. The order of the changes in a diachronic derivation of a potential chain depends solely on the resulting phonological contrast not on the causal relationship or the historical context. This means that, for the purposes of WDT, the changes within the chain need not be contemporaneous or causally linked, since neither contemporaneousness nor causality are expressed in the reconstruction.

Chains can also be defined structurally. In a chain, the output of a change is the input of a preceding change, the output of which may itself be the input of an earlier change. These intermediate changes within a chain are neither mergers nor shifts, as they do not take mother-exclusive segments as input. We will refer to this final type of sound change as *links*, since they are the building blocks of chains. As is discussed below, chains consist of one or more links as well as exactly one merger or shift.

Observe the schematic version of Grimm's Law in (23), a well-known chain in historical linguistics. Only the dental series of consonants are given, though the change also applied to labials and dorsals. Grimm's Law itself can be described as a chain performing the function of the shift *$d^{ɦ}$ > θ, with two links going through [d] and [t]. Grimm's Law ′ has the same effect on the segmental inventory as Grimm's Law, the removal of [dʱ] and the introduction of [θ], while using only a single link going through [t]. Finally, Grimm's Law″ is the simple shift dʱ > θ.

(23) *Grimm's Law and alternatives I*

Grimm's Law	*Grimm's Law′*	*Grimm's Law″*
*t > θ	*t > θ	*dʱ > θ
*d > t	*dʱ > t	
*dʱ > d		

All three reconstructions in (23) alter the phoneme inventory in the same way; all three remove the mother-exclusive [dʱ] and introduce the daughter exclusive [θ]. As a result, the wordlists produced by the three reconstructions in (23) are synchronically compatible with one another. However, the processes yield distinct wordlists, as can be confirmed in (24), where each version of Proto-Germanic is derived from the corresponding reconstruction in (23) (original data from Ringe, 2017).

(24) *Proto-Germanic I*

gloss	*pre-G*	*Proto-G*	*Proto-G′*	*Proto-G″*
three	*trinz	θrinz	θrinz	trinz
intestine	*tarmaz	θarmaz	θarmaz	tarmaz
sweet	*swoːduz	swoːtuz	swoːduz	swoːduz
he knows	*waid	wait	waid	waid
judgment	*dʱoːmaz	doːmaz	toːmaz	θoːmaz
middle	*midʱjaz	midjaz	mitjaz	miθjaz

It is a general property of links, and hence of chains, that additional changes do not alter the phoneme inventory. As such, unlike mergers and shifts, links have no bearing on whether the result is synchronically compatible with the daughter and are, in this sense, optional. Stated mathematically, if a chain of *n* links with input *x* and output *y* derives a synchronically compatible wordlist, a chain of *n*-1 links with input *x* and output *y* also derives a synchronically compatible wordlist, as can be shown by comparing Proto-Germanic′ and Proto-Germanic″ in (24). Both reconstructions convert a wordlist bearing pre-Proto-Germanic phonology into one bearing Proto-Germanic phonology in fewer steps than the reconstruction to actual Proto-Germanic. As a result, Proto-Germanic′ and Proto-Germanic″ are closer to pre-Proto-Germanic than actual Proto-Germanic, ensuring that they are diachronically local to Proto-Germanic.

Unlike for mergers and shifts, different combinations of links are not guaranteed to yield distinct outputs. Regardless of the size or composition of the wordlist, there are combinations of links within a chain which always produce the same result. For instance, the wordlists in (24) can be reached through the efficient minimal chains in (23) but also the corresponding redundant chains in (25). The outputs of Grimm’s Law in (23) and (25) are identical, as are the outputs of the two instances of Grimm’s Law′ and of the two instances of Grimm’s Law″.

(25) *Grimm’s Law and alternatives II*

Grimm’s Law	*Grimm’s Law′*	*Grimm’s Law″*
*t > θ	*t > θ	*dʱ > θ
*d > t	*t > t	*t > dʱ
*dʱ > d	*dʱ > d	*dʱ > t

As the examples in (25) demonstrate, the outputs of links in a chain are not independent of one another. Because WDT uses the number of transformations to estimate the number of diachronically local wordlists, in the case of links, it is important to establish not only that a resulting wordlist is synchronically compatible and diachronically local, but also that it is unique. The question of which link outputs yield distinct outputs is addressed in more detail in Section 3.4.1.

It is always possible to introduce redundant links or redundant groups of links. It follows that, for every chain *x* of length *n*, there exists a chain *y* of length *n*+1 such that the output of *x* is also the output of *y*. Therefore, counting the shorter chains in (23) and the longer chains in (25) separately inflates $|D \cap S\ |$ and increases the likelihood of type II error, i.e. a false negative. A simple way to address the issue is to not count chains of lengths shorter than the one required by the reconstruction to the daughter. In other words, while chains shorter than the one required by the daughter can in fact yield synchronically compatible wordlists, as in (23), there always exist longer chains that do so in the same way.

As in the case with the outputs of mergers, the outputs of links, which themselves may be the inputs of subsequent links, are arbitrary from a probabilistic perspective. For example, the two reconstructions in (26) yield synchronically compatible wordlists, even though one operates through links [t] and [d] while the other through [z] and [s].

(26) *Grimm’s Law and alternatives III*

Grimm’s Law	*Grimm’s Law′*
*t > θ	*s > θ
*d > t	*z > s
*dʱ > d	*dʱ > z

Both reconstructions in (26) remove the mother-exclusive [dʱ] and introduce the daughter-exclusive [θ] and alter the phonological inventory in no other way. Therefore, provided that the daughter wordlist exhibits [s] and [z], the product of Grimm's Law′ is synchronically compatible with and diachronically local to the product of Grimm's Law. However, the wordlists resulting from the two reconstructions are distinct, as can be confirmed in their respective outputs to pre-Proto-Germanic in (27).

(27) *Grimm's Law III*

gloss	*pre-G*	*Proto-G*	*Proto-G′*
three	*trinz	θrinz	trins
intestine	*tarmaz	θarmaz	tarmas
sweet	*swoːduz	swoːtuz	θwoːdus
he knows	*waid	wait	waid
judgment	*dʱoːmaz	doːmaz	zoːmas
middle	*midʱjaz	midjaz	mizjas

The link outputs in (26) were chosen arbitrarily. The links can also go through other homotopic segments, e.g. [l] and [n] or [j] and [h], and still yield synchronically compatible wordlists. As in the case of mergers, applying a chain to the mother wordlist results in a wordlist synchronically compatible with the daughter if and only if the outputs of the individual links are homotopes.

Because chains are headed by mergers and shifts, the number of diachronically local and synchronically compatible wordlists introduced by links is also dependent on the number of mergers and shifts. Let us explore the behavior of links when two or more chains are posited by focusing on data from the Great English Vowel Shift in (28).

(28) *Great English Vowel Shift I*

gloss	*Middle English*	*Modern English*
bite	biːt(ə)	baɪt
beet	beːt(ə)	bi(ː)t
bake	baːk(ə)	be(ː)k
bow	buː(w)	baʊ
boot	boːt	bu(ː)t
boat	bɔːt	bo(ː)t

The Great Vowel Shift can be described as two chains of equal length, each headed by a shift. The front vowel chain contains two links going through [eː] and [iː]; the back vowel chain contains two links going through [oː] and [uː]. Figure 9 illustrates this relationship in the vowel space. The two chains remove the mother-exclusive [aː] and [ɔː] and introduce the daughter-exclusive [a͜ɪ] and [a͜ʊ], respectively. As discussed thus far, the effect of each chain on the segment inventory is the same as the effect of an individual shift.

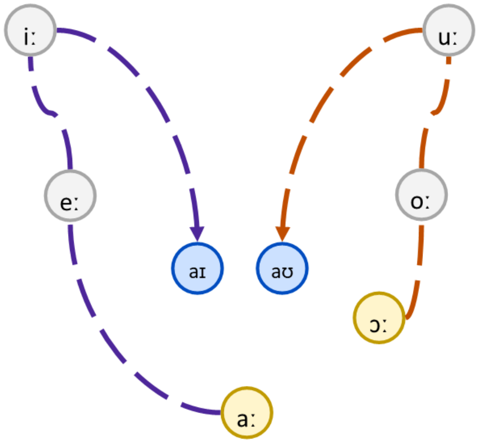


Figure 9: *Great English Vowel Shift*. The process is illustrated as two chains each headed by a shift. Mother-exclusive segments are in yellow. Daughter-exclusive segments are in blue. Link outputs are in grey.

In the case of the Great Vowel Shift, the links are distributed evenly between the front vowel chain and the back vowel chain. However, there is no inherent reason for this to be the case. Working with a reconstruction, there is no way of knowing if the two chains operated during the same time period and whether they were motivated by similar linguistic and extra-linguistic causes. As such, it should not be assumed that chain length is uniformly distributed within a reconstruction. For the Great Vowel Shift, a different partition of four links into two groups is equally plausible. For example, the front vowel chain may contain three links while the back vowel chain only one, as illustrated in Figure 10. Modern English′ in (29), derived through this alternative chain, is diachronically local to and synchronically compatible with Modern English.

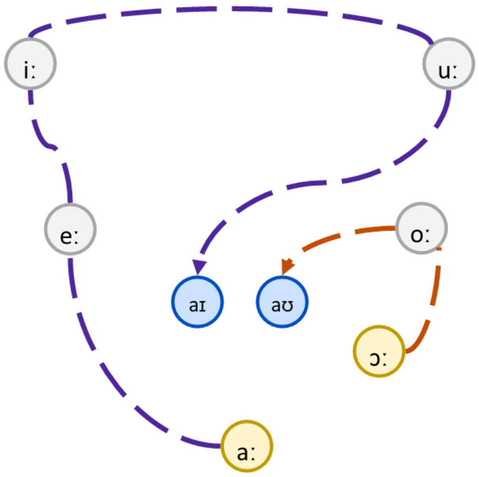


Figure 10: *Alternative Great English Vowel Shift*. The process is illustrated as two chains of unequal length each headed by a shift. Mother-exclusive segments are in yellow. Daughter-exclusive segments are in blue. Link outputs are in grey.

(29) *Great English Vowel Shift II*

gloss	*Middle English*	*Modern English*	*Modern English'*
bite	biːt(ə)	baɪt	bu(ː)t
beet	beːt(ə)	bi(ː)t	bi(ː)t
bake	baːk(ə)	be(ː)k	be(ː)k
bow	buː(w)	baʊ	baɪ
boot	boːt	bu(ː)t	baʊt
boat	bɔːt	bo(ː)t	bo(ː)t

Note that the arbitrariness in the distribution of links between chains is distinct from the arbitrariness in the outputs of the links themselves. The example in Figure 4 is meant to illustrate the former. Thus, the vowel outputs in the chain are of no consequence here; what matters is that the chain introducing [a͜ɪ] contains two more links than the one introducing [a͜ʊ].

Every shift or merger is a potential landing site for a one or more links. When dealing with more than one link, all partitions of links between shifts and mergers yield synchronically compatible and diachronically local wordlists, since the links themselves do not alter the phonological inventory of the language. For the English Vowel Shift data, this means that wordlists derived from Middle English through the simple shifts *aː > a͜ɪ or *ɔː > a͜ʊ also yield synchronically compatible and diachronically local wordlists.

The number of ways some number of links can be distributed between shifts and mergers is a well understood problem in combinatorics known as distributing identical objects to distinct recipients (Bogart, 1983). The formula for finding the number of ways to distribute k identical objects to *n* distinct recipients is $\frac{(k+n-1)!}{(n-1)!k!}$, or simply $\binom{k+n-1}{k}$. A full derivation is out of place in this dissertation. However, stated informally, the formula finds the number of permutations of k objects along with n-1 dividers which demarcate the objects' recipient. This product is then divided by the number of ways to permute the objects and the number of ways to permute the dividers, since the objects and the dividers are not distinct.

Adapting the formula to estimate the number of synchronically compatible and diachronically local wordlists introduced by some number of links λ yields (30), where φ is the number of mergers, χ the number of shifts, and $\hbar$ the mean number of homotopes in the daughter. The formula combines the two sources of arbitrariness introduced by links. First, the links can be distributed between shifts and mergers in one of $\binom{\lambda+\varphi+\chi-1}{\lambda}$ ways, where λ acts as the number of identical objects and $\varphi + \chi$ as the number of recipients. Second, the output of each link can be one of $\hbar r$ segments. Recall that the formula does not capture the arbitrariness resulting from shorter chains because their output is included in the redundancy of the longer chains.

(30)

$$|D \cap S\,|_\lambda = \binom{\lambda+\varphi+\chi-1}{\lambda}(\hbar r)^\lambda$$

Link Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by λ links distributed between φ mergers and χ shifts, with mean number of homotopic segments $\hbar$. The variable $\hbar$ is scaled by *r* to account for non-unique outputs of chains.

As already discussed, not all combinations of link outputs produce unique wordlists. Therefore, although the product of a chain when applied to a wordlist is synchronically compatible wordlist if the output of each link is a homotope of the original output, not all of these are unique. To avoid the overcounting of non-unique outputs, the formula in (30) scales the mean number of homotopes $\hbar$ by a factor *r*.

The scaling factor *r* corresponds to the average proportion of homotopic segments $\hbar$ that, when a part of a chain, produce unique wordlists. The scaling factor *r* can also be thought of as the proportion of unique wordlists in the set of diachronically local and synchronically compatible wordlists resulting from the application of links. Because the number of unique synchronically compatible wordlists is smaller than or equal to the number of synchronically compatible wordlists, *r* must be smaller than or equal to 1. In principle, the formula in (30) could have been modified in other ways, or a separate formula for estimating redundancy could be used. I choose to scale $\hbar$ alone as this captures the intuition that the set of unique link outputs that yield synchronically compatible wordlists is a subset of the set of outputs that yield synchronically compatible wordlists.

3.4.1 Unique Link Outputs

As it stands, the exact value of the scaling factor *r* in the Link Formula is unknown. This section is concerned with estimating *r* using computer simulation. The simulations were conducted in *Python*. Given a mother wordlist, daughter wordlists were derived through chains stochastically. The proportion of unique wordlists generated this way was taken to be an estimate of the proportion of homotopic segments which yield unique outputs in a chain, i.e. *r*.

There are too many unique wordlists to generate them all. Thus, to estimate the total number of possible wordlists, the Lincoln Index was used. The Lincoln Index is a well-known method of estimating the total size of a population by sampling with replacement and tallying the repeat elements in the sample. The method is often used in ecology to estimate the population of natural species (Adams, 1951). In the simulation, the Lincoln Index was calculated by generating two random samples of unique synchronically compatible and diachronically local wordlists and inserting the size of the two samples S_1 and S_2, as well as the size of the overlap $S_1 \cap S_2$ into the formula in (31). The total size of the population is approximately equal to the product of the two sample sizes divided by the size of the overlap.

(31)

$$|T| = \frac{|S_1| \cdot |S_2|}{|S_1 \cap S_2|}$$

Lincoln Index: formula for estimating the size of population *T* based on two random samples S_1 and S_2 and their intersection.

The simulations started with mother wordlists exhibiting mean homotopy values between 2 and 10. All of the real languages in the Chapter 4 case study in Chapter 4 of this dissertation fell into this range (see Table 1 for a sample). For each homotopy value, a single mother wordlist was generated. Each mother wordlist was distorted into daughter wordlists through 10 shifts and a number of links varying between 0 and 10. The shifts were included only because links cannot occur on their own and must be part of a chain headed either by a shift or a merger.

The purpose of the simulation is to find the number of unique wordlists derived by this method for varying values of homotopy and links. There were 99 simulations in total, for 9 mean homotopy values from 2 to 10 (inclusive), as well as 11 link values from 0 to10 (inclusive). For each simulation, two sets of sample wordlists were generated. For each simulated set of wordlists, duplicates were removed. The cardinality of the two sets and their intersection was used to estimate the number of unique wordlists derivable through the reconstruction using the Lincoln Index formula in (31).

Daughter wordlists were derived from the mother using a random distribution of links between shifts and random link outputs from a set of size $\hbar$. Because link outputs were sampled from $\hbar$, the resulting wordlists are guaranteed to be synchronically compatible. For each simulation the number of wordlists sampled was set to $2 \cdot |D \cap S\,|_{\lambda}{}^{1/2}$, where $|D \cap S\,|_{\lambda}$ was calculated using (30) with *r* set to 1.[14] This sample size maintained a relatively uniform distribution of standard error for different values of λ and $\hbar$, varying between 1.4% and 4.4% of the estimate.

The number of unique wordlists generated in the simulation and estimated using the Lincoln Index was compared to the output of the Link Formula in (30) where *r* = 1, i.e. where no scaling has been applied. The scaling factor *r* is simply the proportion of unique wordlists in the Link Formula prediction.

The scaling factor *r* was allowed to vary by the mean number of homotopes $\hbar$ and the number of links λ. Because there is some error in the Lincoln Index approximation, no simple function predicted the results perfectly. The closest approximation that I could find using trial and error was one for $\mathrm{r} = \frac{\hbar+3}{9\hbar^2} + \frac{1}{\hbar(\lambda+1)}$. Figure 11 compares the amount of unique diachronically local and

[14] There is no theoretical reason for choosing samples of this size. In general, the larger the samples the more accurate the Lincoln Index results. This particular sample size was the largest of those tested that did not incur a memory error on my 16GB RAM laptop while maintaining a relatively stable error for different values of λ and $\hbar$.

synchronically compatible wordlists found in the simulation (dots) with the prediction made by Formula (30) with this estimate for r (lines).

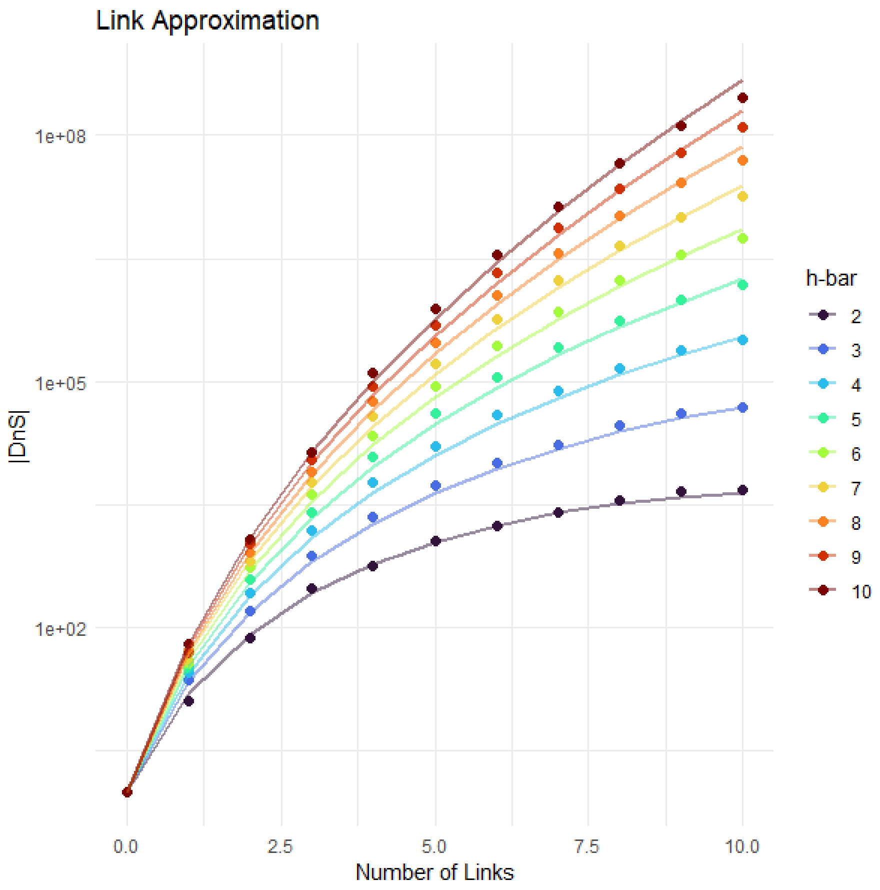


Figure 11: *Link approximation*. $|D \cap S|$ simulated for different values of $\hbar$ homotopes and λ links (points) compared to $|D \cap S|$ approximated using the Link Formula in (30) with $r = \frac{\hbar+3}{9\hbar^2} + \frac{1}{\hbar(\lambda+1)}$ (lines). The y-axis is logarithmic.

The prediction fit the data relatively well. However, future research may find better approximations or solve the problem of redundant wordlists analytically. For the combinations of homotopic segments $\hbar$ and number of links λ tested in the simulation, r was 0.44 on average. In other words, the number of link outputs which yield unique wordlists are slightly less than half of those which yield synchronically compatible wordlists. Because the effect of a change in $\hbar$ on $|D \cap S\,|_\lambda$ is polynomial, failing to account for link output redundancy, e.g. by using the scaling factor r, can lead to catastrophic overestimation of $|D \cap S\,|_\lambda$.

Returning to (28), the Link Formula in (30) can be used to estimate the number of diachronically local and synchronically compatible wordlists introduced by the 4 links and 2 shifts posited in the Great English Vowel Shift. There are $\binom{5}{4} = 5$ ways to distribute the links between the back chain and the front chain. The mean number of homotopes in English is approximately 10.4 (see Section 3.1). For each link, there are approximately $\hbar\mathrm{r} = \frac{\hbar+3}{9} + \frac{\hbar}{\lambda+1} = \frac{10.4+3}{9} + \frac{10.4}{4+1} = 3.6$ link outputs that yield unique synchronically compatible wordlists. In total, there are approximately $(\hbar\mathrm{r})^{\lambda} = 3.6^4 = 162.2$ link outputs for the four links in total. Finally, there are two ways to map the mother-exclusive [ɔː] and [aː] onto the daughter-exclusive [a͜ʊ] and [a͜ɪ] using shifts. Thus, there are approximately $162.2 \cdot 2 \cdot 5 = 1622$ wordlists synchronically compatible with Modern English and diachronically local to Modern English given the Great Vowel Shift in (28). Recall that, to estimate *P*(*D*|*S*) of a reconstruction from Middle English to Modern English, the number of synchronically compatible and diachronically local wordlists needs to be divided by the total number of wordlists synchronically compatible with Modern English.

For practical applications, perhaps the most important property of links is that they generally incur a smaller increase to *P*(*D*|*S*) than mergers. This means that the Merger Formula in (9) provides a conservative estimate for the effect of all sound change on *P*(*D*|*S*). The Merger Formula has the additional benefit of being the simplest of those introduced in this chapter, while also allowing for simplification with the Synchronic Formula and Lexical Replacement Approximation to create rule-of-thumb inequalities, such as the one in (14). Finally, if all sound change is approximated using the Merger Formula, then there is no need to keep track of the shift/merger/link distinction. Nevertheless, in cases where a high level of precision is necessary, or in cases where type II error is a concern, such shortcuts can be detrimental.

One can confirm that mergers are generally more costly, i.e. incur a greater increase on *P*(*D*|*S*), than links by comparing *P*(*D*|*S*) values for reconstructions with the same total number of changes but different combinations of mergers and links. There is no need to consider shifts, since their effect on *P*(*D*|*S*) is extremely small, as demonstrated in Section 3.3. In Figure 12, one can observe the *P*(*D*|*S*) of a reconstruction with mean number of homotopes $\hbar$ at 10, wordlist length *t* at 100, and word complexity *c* at 10,000. The number of mergers and links in the reconstruction varies such that $\varphi + \lambda = 24$, for φ mergers λ links. Thus, movement along the x-axis corresponds to an increase in the proportion of mergers in the sound changes.

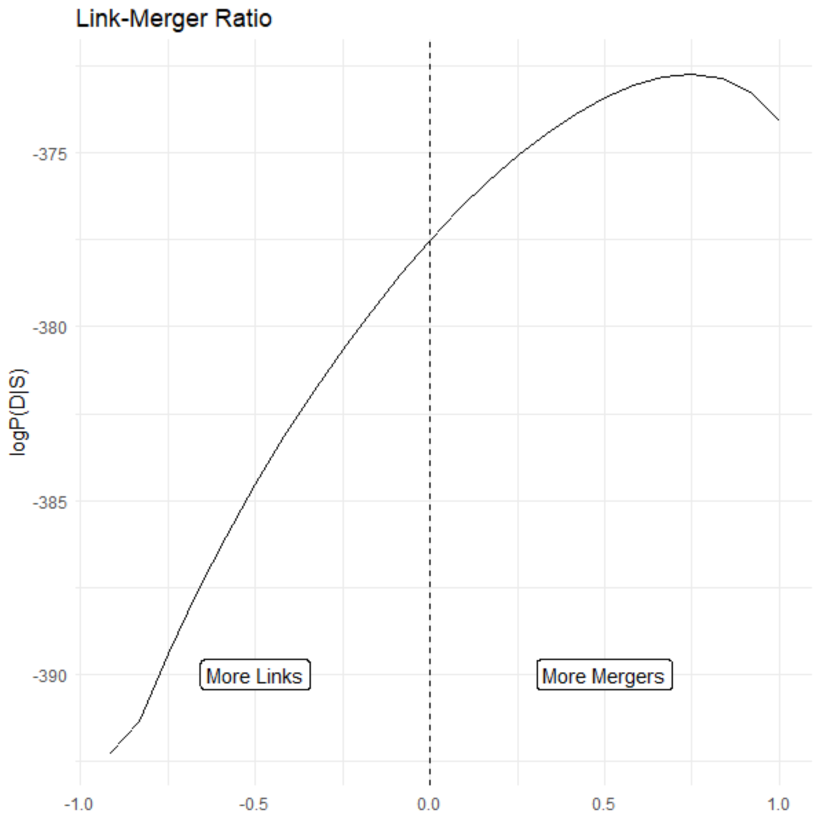


Figure 12: *Link-merger ratio*. $P(D|S)$ is on the y-axis (logarithmic). The axis is a continuum of different combinations of mergers and links which add up to 24, from 24 links (-1) on the left of the graph to 24 mergers (1) on the right. For all values, $\hbar$ at 10, wordlist length *t* at 100, and word complexity *c* at 10,000.

As can be seen in Figure 12, a replacement of a link with a merger usually corresponds to a slight exponential increase to *P*(*D*|*S*), though this becomes less pronounced as the proportion of mergers increases. At some point, positing additional mergers at the cost of links actually slightly decreases *P*(*D*|*S*). As a result, the highest *P*(*D*|*S*) occurs where almost all sound changes are mergers but a few links remain. In Figure 12, this happens at 21 mergers and 3 links, or when 87.5% of changes are mergers. However, when 100% of changes are mergers, *P*(*D*|*S*) is only 1.5 log units lower than the maximum. Treating all sound changes as mergers results in a *P*(*D*|*S*) higher than 17 of the 24 possible combinations or mergers and links. As such, in most situations, using the Merger Formula in (9) for all sound changes leads to a conservative estimate of *P*(*D*|*S*).

3.4.2 Sound Change Interactions

As hinted at the beginning of this section, the architecture of links accounts not only for chain shifts but for sound change interaction in general. Recall that, when estimating the number of synchronically compatible and diachronically local wordlists, the outputs and environments of all sound changes are assumed to be any homotopic segment. The same is true for the inputs of links, which are themselves the outputs of other changes.

As a result, to derive wordlists that display sound change interactions, there is no need to account for sound change ordering separately. In WDT sound change orderings are emergent from the substitution of the individual sound change elements with homotopic segments. For example, observe two vowel raising changes in (32). Assume that these are two links and, therefore, that the outputs and inputs are free to vary in a derivation to a synchronically compatible and diachronically local wordlist.

(32) *Two Links*
(a) a > e / _n
(b) e > i / _n

As it stands, (a) feeds (b), giving (b) additional targets for application, e.g. *an > *en > [in]. In a reverse ordering, (b) would counterfeed (a), resulting in surface opacity, e.g. *an > [en]. The difference in order between the two changes can be achieved either by switching the order of the original sound changes or simply by altering each of the elements in (a) and (b) until the result is the reverse of the original. In one of the derivations to a synchronically compatible and diachronically local wordlist, the input to (a) will be [e] and the output [i], while the input to (b) will be [a] and the output [e], i.e. the reverse of the original. Because any homotopic segment can occupy any position in (32), the reverse ordering of a rule can always be reached simply by altering every element in the two rules. No separate mechanism for sound change interaction is required.

In WDT, the set of synchronically compatible and diachronically local wordlists is modelled to include all unique wordlists that are in line with the phonological properties of the daughter language. The discussion of rule ordering is somewhat tangential to this purpose. All that matters is that the formulas used to estimate the number synchronically compatible and diachronically local wordlists include wordlists that can be derived by rule ordering. Since rule ordering can be modelled by altering the individual elements in the sound changes, no separate mechanism is necessary.

3.5 Phonological Features

This section is devoted to demonstrating that models of sound change employing phonological features are compatible with Wordlist Distortion Theory. For several reasons, the discussion of features in this section is incomplete and cursory. Firstly, formulae operating on phonological features are more mathematically involved than those operating on single segments, making them a poor choice for the introduction of the framework. Secondly, the implementation of phonological features into WDT differs greatly depending on the feature theory assumed. So as to keep the analysis simple and general, phonological features are omitted in the main portion of this chapter, as well as the case study in Chapter 4. Likewise, in order to keep the analysis applicable to various phonological models, only a sketch of phonological feature implementation is presented. The intention is for the outline to guide future evaluations of comparative reconstruction employ feature frameworks in the spirit of Wordlist Distortion Theory.

WDT is compatible with different models of phonological features. It is possible to use features that refer to the presence or absence of a property (binary), or the presence only (privative). Features could exist on a single tier or on multiple tiers (autosegmental). In any case, WDT must come equipped with a definition of a single operation. For example, a single change could be counted as the addition of a feature, the removal of a feature, or, in binary feature frameworks, the change in feature valuation.

However, it should be stressed that both the choice of feature framework and, more importantly, the assignment of features within the framework must be independent of the data. Assigning features based on the assumed history of the language in question is a form of confirmation bias, in the same way that assembling wordlists based on assumed cognacy is cherry-picking. In other words, it is more likely that a pair of wordlists evidences a phonologically naturally change when the notion of phonological naturalness is defined to reflect the data in the wordlists.

The requirement of theory-data independence may appear contrary to many theories of synchronic phonology which assume that some or most aspects of feature systems are language-specific (Blaho, 2008; Mielke, 2008; Clements & Ridouane, 2011; Cowper & Hall, 2015). However, the goals of synchronic phonology differ greatly from those of the current project. Synchronic phonology is concerned with modelling the phonological component of the mind. Features, for such a system, are a way to define psychologically salient groups of segments, evidenced by distribution (Gallagher, 2019), alteration, or contrast (Dresher, 2009). However, there is no reason that the entire history of a language should be reflected in the synchronic grammar of each speaker, nor should large-scale language change be limited by cognitive constraints in the mind. The history of a language spans many generations of speakers and reflects multiple grammars. Each of these grammars can be modelled in a language-specific framework and is expected to behave in accordance with synchronic principles. However, the transition between these is better modelled using a language-independent theory and may result in synchronically unnatural patterns (Mielke, 2008; Beguš, 2018).

Conducting an evaluation of comparative reconstruction within a language-specific feature framework is not only inappropriate but also potentially detrimental. On the one hand, such a feature analysis may introduce sound changes into the analysis that are not reflected in mismatches between the mother and the daughter wordlist. For example, a comparison between Cheyenne and Proto-Algonquin, its direct ancestor, reveals the vowel mergers *o > i and *oː > iː (Goddard, 1986). A language-specific analysis may posit that these mergers are indicative of a total restructuring in the vowel inventory, where most vowels acquire a new feature specification regardless of the phonetics (Oxford, 2015).[15] For example, the feature [low] is hypothesized to have been introduced into the inventory at this time, applying to the vowels [a] and [aː]. Thus, although the merger itself can be conceptualized as the feature operation [labial] > ∅, the overall restructuring also implies featural changes to other vowels, even if these remain phonetically unchanged.

[15] The analysis in question was conducted using the Successive Division Algorithm within the Contrastivist Hypothesis (Dresher, 2009). In this framework, the inventory is successively split into binary groups along progressively less prominent features. The mergers of o > i and oː > iː in Cheyenne imply the demotion of [labial] in the feature hierarchy and the promotion of [low], resulting in an increase to the scope of the latter in the inventory.

Accordingly, a language-specific analysis of this kind may choose to posit additional featural changes to reflect the restructuring in the vowel system. In my opinion, these additional feature changes do not reflect the arbitrariness of the reconstruction. For example, *a in Proto-Algonquin still corresponds to [a] Cheyenne, even if Proto-Algonquin [a] is analyzed as [syllabic] while Cheyenne [a] is analyzed as [syllabic, low]. Therefore, even though a language-specific analysis may require an additional sound change to explain the difference in feature specification between Proto-Algonquin *a and Cheyenne [a], implying additional synchronically compatible and diachronically local wordlists, the additional increase to $P(D|S)$ is not reflected in any observable mismatch between the mother and daughter wordlists. As such, attempting to evaluate the reconstruction by modelling the synchronic phonology in this way may result in an overestimation of the number of synchronically compatible and diachronically local wordlists and an increase in type II error, i.e. a false negative.

On the other hand, a language-specific feature analysis of sound change may ignore attested segmental mismatches between the mother and the daughter wordlist. For example, compare Proto-Malayo-Polynesian *taŋis and Hawaiian [kani] ‘sound’, requiring the change *t > [k] (Trussel & Blust, 2010). By the time the sound change took place, the (pre-)Hawaiian [t] was the only lingual obstruent in the language. The transition from [t] to [k] does not alter the status of the phoneme with respect to contrasts in the inventory, as the resulting segment is still the only lingual obstruent in the language. Because the feature specification of [t] before the change can be argued to be the same as [k] after it (Herd, 2005), an evaluation of the reconstruction from Proto-Malayo-Polynesian to Hawaiian may choose to posit no changes from *t to [k].[16] However, it can be argued that such an analysis ignores the mismatch between the two wordlists. As such, attempting to evaluate the reconstruction by modelling the synchronic phonology in this way may result in an underestimate of the number of diachronically local and synchronically compatible wordlists and an increase in type I error, i.e. a false positive.

Since a language-specific feature analysis is not compatible with WDT, the implementation of features into the framework requires both a universal set of features and a universal set of

[16] Recall that single shifts in the reconstruction do not increase $P(D|S)$ even in a purely segmental analysis of sound change (see Section 3.3). However, this property of shifts is independent of what is being discussed here. In many situations where a segmental analysis must posit multiple shifts and increase $P(D|S)$, a language-specific featural analysis may chose not to posit any, so long the feature specifications are hypothesized not to have changed.

segments defined by the universal features. The analysis of a particular languages may use only a subset of the universal features and segments. However, the segments used must be specified for the same features in all languages. For instance, a segment, say [s], should bear the same feature description regardless of language reconstruction, say [coronal, continuant]. A change between the segment [s] and the segment [z], which itself may be described as [coronal, continuant, voice], can be defined as the insertion of the feature [voice], regardless of the phonological status of [s] in the mother or [z] in the daughter, or the participation of the two segments in phonological processes elsewhere in the language.

Technically, there is no requirement in Wordlist Distortion Theory for the feature set used in the evaluation of comparative reconstruction to be phonetically substantive. So long as the features are chosen independently of the data, any universal feature set is permissable. However, in practice, it is difficult to come up with a language-independent way of assigning features that does not directly follow from articulatory or acoustic descriptions.

The use of features in WDT can be motivated on purely utilitarian grounds, since an analysis using features is expected to yield lower $P(D|S)$ than a corresponding segmental analysis in cases of actual descent. In fact, there are two ways in which a well-implemented feature framework decreases $P(D|S)$. Firstly, features implicitly define a phonological distance metric between segments, giving preference to sound changes between similar segments over sound changes between dissimilar ones. Secondly, when defined as an operation on features, sound change can aggregate multiple segment-to-segment mappings. These two properties of features are discussed in more detail in the following sections. Both properties decrease the number of diachronically local and synchronically compatible wordlists introduced by sound change, at least in situations where the sound changes in question are in line with the feature framework.

3.5.1 Defining Phonological Distance

Observe the simple vowel system in (33). Similar 4-vowel systems have been reported in Yupik (Krauss, 2009), Paiwan (Chang, 2006), and Ivatan (Heye & Hidalgo, 1967). The inventory is conveniently symmetric and can be described with just two features, [back] and [high].

Frameworks utilizing features in similar ways can be found in the literature (Morén, 2003; Iosad, 2014), though other analyses are also compatible with WDT.

(33) *Sample 4-vowel system*

	[high]	[back]
i	✓	
u	✓	✓
a		
ə		✓

In (33), the vowels [i] and [ə] are specified for just one feature, [u] is specified for both, while [a] is specified for neither. Note that only features relevant to the discussion are given in the (33). As such, [a], specified for neither [high] or [back], may also bear the feature [syllabic], for example.

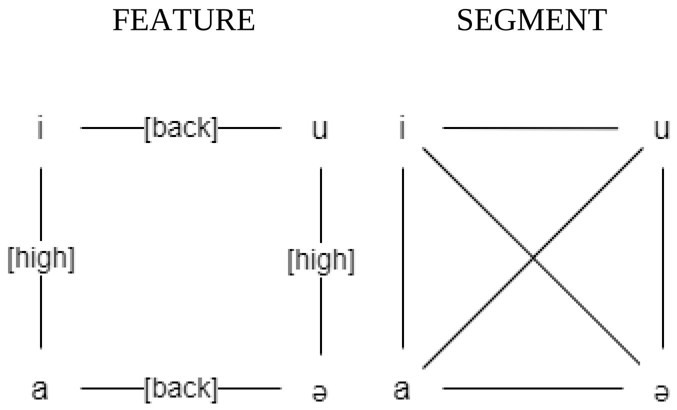
Suppose that the language exhibits a historical merger of the high vowels to [u] before the bilabial nasal, segmentally i > u /_m. In a framework devoid of features, such as the employed in this dissertation, the number of sound change outputs that yield synchronically compatible and diachronically local wordlists to the daughter is simply the number of homotopic segments (see Section 3.2.1). Assuming a basic CV syllable structure, the number of homotopic segments in this case is approximately one less than the number of total vowels, since [i] is ungrammatical in this environment (due to the sound change). In other words, there are three outputs of a merger with the input [i] which are expected to yield wordlists in line with the phonotactics of the daughter language: [u, a, ə].[17]

In a privative feature framework, the merger in question can be thought of as a single transformation operating on the features themselves, [high] > [high, back]/ _[labial, nasal]. Stated informally, the feature [back] is inserted if the feature [high] is in the target immediately followed by both [labial] and [nasal]. Let us suppose that the insertion or removal of a feature counts as a single operation in a featural analysis using WDT. As such, the number of synchronically compatible and diachronically local wordlists introduced by the sound change is the number of single feature changes in the privative framework resulting in grammatical output in the daughter.

[17] In the case of conditional changes, it is also possible to derive synchronically compatible and diachronically local wordlists that operate on the original context (see Section 3.3). As such, the number of synchronically compatible and diachronically local wordlists introduced by the merger is roughly $2\hbar$. These additional synchronically compatible and diachronically local wordlists are not relevant to the discussion and are omitted in this section.

For the example in (32), the number of wordlists synchronically compatible with and diachronically local to the daughter given the reconstruction appears to be two. One could insert the feature [back], resulting in i > u/ _m, as took place in the daughter. Alternatively, one could remove the feature [high] in the same environment, resulting in i > a/_m. Unlike in the segmental description of the same phenomenon, there is no way to map [i] onto [ə] with a single feature operation, since [i] and [ə] differ in both presence of [high] and the absence of [back]. Therefore, a featural analysis of the same sound change implies one fewer diachronically local and synchronically compatible wordlist.

To visualize the difference between segmental and featural approaches with respect to the inventory in (33), turn to Figure 13, which presents a graph of the possible transitions between segments utilizing operations of minimal complexity. Edges in the graph are drawn between segments that could be mapped onto each other using a single segment-to-segment mapping (on the right) and a single featural change (on the left). As discussed in this chapter already, in segmental frameworks, any segment can be converted into any other homotopic segment using only a single sound change. In featural frameworks, a segment can be turned only into segments differing in the presence or absence of a single feature, a subset of the homotopic segments.



FEATURE SEGMENT

Figure 13: *Transitions in a 4-vowel system*. Graphs illustrating the possible mappings through a single sound change in a segmental framework (right) and a single feature insertion/deletion change in a featural framework (left) for the vowel inventory in (33).

In effect, features provide a way of quantifying phonological distance. For example, it could be said that, in Figure 13, [i] and [a] are a distance of 1 featural change apart, while [i] and [ə] are a distance 2 of two featural changes apart. Because the original change took place between segments of a distance of 1 apart, only changes of a distance of 1 (or less) apart produce wordlists diachronically local to the daughter.

In this way. the phonological distance metric introduced by a feature framework allows for the discrimination between natural and unnatural sound changes. Because the prediction about sound change naturalness is not a part of a system with simple segment-to-segment mappings, systems with features are more restrictive. More restrictive frameworks allow for fewer possible transformations. Fewer possible transformations, in turn, result in fewer diachronically local wordlists derivable through these transformations. As such, restrictive frameworks in general, and phonological frameworks employing features in particular, are expected to yield lower $P(D|S)$ in cases of actual descent.

The decrease in $P(D|S)$ for featural frameworks applies only if the data requires a mapping between two phonologically similar segments. Imagine instead that a language with the vowel set in (33) evidenced the change i > ə /_#. In a segmental framework, the number of derivations yielding a diachronically local wordlist is 4, one for every (homotopic) segment. In a featural framework, the same sound change can be conceptualized as two operations on features, [high] > ∅ /_# and ∅ > [back] / _#, the insertion of [back] and the removal of [high]. The number of diachronically local wordlists introduced by this new change is the same as the number of segments that could be reached from [i] through at most two feature substitutions of this kind. As it turns out, this condition applies to all 4 vowels. Therefore, for changes where the input and that output are sufficiently dissimilar, such as in the case of distant descent or no descent, a featural framework provides little to no decrease to $P(D|S)$ when compared to a segment-to-segment mapping approach.

As the number of segments and features increases, so does the difficulty in estimating the number of synchronically compatible and diachronically local wordlists introduced by the change. For an illustration, turn to a different vowel inventory in (34). 5-vowel inventories as in (34) are extremely common and occur in approximately a third of the world’s languages (Schwartz et al, 1997).

(34) *Sample 5-vowel system*

	[high]	[back]	[low]
i	✓		
u	✓	✓	
e			
o		✓	
a		✓	✓

A five-vowel inventory requires at least 3 features to uniquely specify every segment. The example in (34) uses the features [high], [back], and [low]. The vowels [i] and [u] share the feature [high]. The vowels [u], [o], and [a] share the feature [back]. To distinguish [o] and [a], an additional feature [low] is needed. Other ways of assigning features are possible, including those that use more than three features.

Figure 14 presents a graph of the possible transitions between segments in (34) utilizing operations of minimal complexity. Edges in the graph are drawn between segments that could be mapped onto each other using a single segment-to-segment mapping (on the right) and a single featural change, i.e. deletion or insertion of a feature, (on the left).

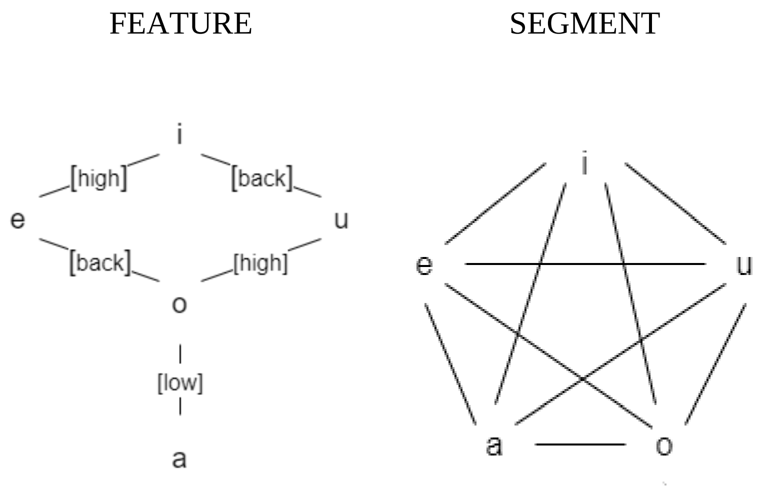


Figure 14: *Transitions in a 5-vowel system*. Graphs illustrating the possible mappings through a single sound change in a segmental framework (right) and a single feature insertion/deletion change in a featural framework (left) for the vowel inventory in (34).

Once again, a framework employing segment-to-segment mappings effectively treats transitions between all pairs of segments as equally arbitrary. In contrast, a framework employing features makes different predictions for changes with different input vowels. Observe that the vowels [i], [e], and [u] can each be converted into two vowels through a single feature change; the vowel [o] can be converted into three in [e], [u], and [a]; and the vowel [a] can be converted into only one, [o]. Because [o] resides in a denser neighborhood of the vowel inventory, reconstructions that posit a merger involving this vowel, e.g. o > u /_m, are particularly probabilistically arbitrary. In other words, a randomly generated wordlist, if analyzed in accordance with (34), is more likely to evidence a minimal feature change from [o], as [o] is adjacent to all three of [e], [u] and [a]. In contrast, a randomly generated wordlist is less likely to evidence a minimal feature change from [a], as [a] is adjacent to [o] only.

Recall that segments exhibit other asymmetries as well. Most notably, the asymmetry in phonotactic segment distribution is relevant for an evaluation in WDT, whether feature-based or segment-based. Similar to the asymmetry in segment neighborhoods, the asymmetry in segment distribution renders some segments to be more likely to exhibit evidence of a merger than others. A segment that is restricted to a narrow environment has fewer opportunities to evidence changes in a randomly generated wordlist than a segment occurring more freely. For the asymmetry in segment distribution, rather than treating all segments individually, this dissertation introduced the notion of homotopes (see Section 3.1.1). The mean number of homotopes measure $\hbar$ acts as a stand-in for the average segment in the language, regardless of phonological position. This measure allows the researcher to treat all sound changes equally, with the assumption that the effect of a random change on $P(D|S)$ is approximately average.

Something similar can be done for the asymmetry in segment neighborhoods in frameworks that employ features. In addition to the mean number of homotopes, one could calculate the mean number of segments reachable through a single feature operation (for a privative system, feature deletion or feature insertion). Similar to the conditional entropy calculation in 3.1.1, this should be a geometric mean weighted by segment frequency. The measure should answer the question: given a randomly selected segment from the wordlist, how many segments exist differing from it in one feature or less? As an example, the vowels in the 4-vowel system in (33) are on average separated from 3 vowels by a distance of one feature change or less. By comparison, the vowels in the 5-vowel system in (34) are on average separated from approximately 2.93 vowels by a distance of one feature or less. Therefore, *ceteris paribus*, a single feature operation for the inventory in (33) is slightly more powerful than an operation for the inventory in (34), as it is able to derive more segments. The greater the number of featural connections between segments, the less restrictive the predictions of the framework and the higher the resulting $P(D|S)$ value in a reconstruction.

Because operations on features are not independent, a feature implementation of the framework must calculate multiple measures of segment connectedness. More concretely, if one feature change can on average derive *m* segments, it is not necessarily the case that two feature changes can derive m^2 segments. For the 4-vowel inventory in (33), going from 0 feature changes to 1 increases the number of reachable segments by a factor of 3 (from 1 to 3), while going from 1 feature change to 2 increases the number of reachable segments only by a factor of 1.33 (from 3 to 4). In effect, because the entire vowel inventory contains only 4 vowels, the second feature change introduces less arbitrariness than the first. In the same way, for the 5-vowel inventory in (34), on average one feature change or less can derive 2.93 vowels, while two feature changes or fewer can derive 4.32 vowels.

It must be the case that, if *n* feature changes can derive *m* segments, *n* + 1 feature changes can derive $\geq m$ segments. Beyond this, nothing is certain. The number of segments that can on average be reached through some number of feature changes depends on the distribution of features in the inventory and must be calculated on a case-by-case basis. Thus, an evaluation of comparative reconstruction employing features must posit multiple approximations for the number of synchronically compatible and diachronically local wordlists introduced by feature changes of varying complexity, e.g. f_1 for sound changes described as a single feature transformation, f_2 for sound changes described as two feature transformations, etc.

To summarize, the number of synchronically compatible and diachronically local wordlists introduced by sound changes in a featural system differs not only by the number of changes posited and the number of features affected in the changes but also by the distribution of features in the inventory. It is intuitive that sound changes altering more features are more costly, i.e. they increase $P(D|S)$ by a greater factor. However, it is perhaps less intuitive, that inventories with more feature minimal pairs, i.e. segments separated by exactly one feature change, increase $P(D|S)$ by a greater factor. For strongly connected inventories, those with more feature minimal pairs, phonological proximity is more likely to occur by accident, resulting in a higher *P*(*D*|*S*) per sound change on average. For weakly connected inventories, those with fewer feature minimal pairs, accidental correspondences are less likely in a feature-based analysis. While the analysis in this section employed a privative feature system, the arguments hold true for any featural analysis.

3.5.2 Aggregating Sound Changes

Another advantage of featural implementations of sound change in terms of *P*(*D*|*S*) is the ability to aggregate segment-to-segment mappings under a single sound change. What segmental frameworks must analyze as multiple changes, either due to multiple segmental outputs or multiple segmental environments, can often be conceptualized as a single change in featural frameworks. The aggregating of sound changes in this way greatly reduces the effect of sound change on *P*(*D*|*S*), on average yielding less arbitrary reconstructions for genetically related languages and increasing the statistical power of the methodology.

For example, imagine a reconstruction positing the changes i > u /_m and a > ə /_m for the 4-vowel inventory in (33). A framework using simple segment-to-segment mappings has no choice but to analyze these changes separately. In such a framework, the effect of the first change on *P*(*D*|*S*) is separate from the effect of the second change. However, notice that, as defined in (33), [i] and [u] differ in the same feature as do [a] and [ə]. Adding the feature [back] to [i] yields [u]; adding the feature [back] to [a] yields [ə]. Therefore, in a featural system, the two mappings can be conceptualized as a single operation on features: ∅ > [back] /_[labial, nasal].

As well as grouping by input, sound changes can be grouped by environment. For example, imagine two additional changes for the 4-vowel inventory in (33), i > u /_n and a > ə /_n. A framework using simple segment-to-segment mappings must analyze these changes independently of one another and independently of the first two changes (vowel backness merger before [m]). A framework using features can conceptualize all four mappings as a single operation on features, ∅ > [back] /_[nasal], provided there are no other nasals in the language.

In a sense, this grouping of sound changes comes at no cost. In other words, there is no additional increase to *P*(*D*|*S*) resulting from ‘aggregated’ sound changes as opposed to sound changes that describe single segment-to-segment mappings. In fact, when analyzed within a feature framework, there is no difference between aggregated sound changes and non-aggregated sound changes. The transformations are conceptualized as operating on properties of segments rather than the segments themselves. Whether changes to the aspects of segments encode a single segment-to-segment mapping or multiple mappings is not expressed directly in the formalism. So long as the choice of feature theory is not *ad hoc*, a single change as conceptualized in the framework should count as a single change in WDT, regardless of its overt effect on the segment inventory.

3.5.3 Section Summary

To summarize, this section argues that an evaluation of comparative reconstruction is compatible with phonological features. In fact, analyzing sound change as operating on features rather than segments is expected to reduce *P*(*D*|*S*) in two ways. Firstly, features provide a natural method of quantifying phonological distance, giving preference to phonologically natural sound changes over phonologically unnatural ones. Secondly, the use of features allows the analysis to aggregate multiple segment-to-segment mappings, decreasing the number of sound changes required by the reconstruction. As a result, in cases of actual descent, the use of features is expected to reduce the likelihood of type II error and increase the statistical power of the analysis.

The drawback of using features in the evaluation of comparative is the added complexity to the model. The increase to *P*(*D*|*S*) resulting from feature transformations is dependent on several factors which are not relevant to simple segment-to-segment mappings, such as the number of features involved in the change, the number of segments in the language, and the distribution of features in the inventory. In addition, the evaluation differs greatly depending on the set of features as well as the type of features assumed. As such, while phonological features and Wordlist Distortion Theory are compatible, a complete description of features in the framework is beyond the scope of this dissertation.

3.6 Diachronic Formula

As a reminder, Wordlist Distortion Theory is largely concerned with presenting formulae that approximate the result of sampling wordlists with particular phonological properties and checking their diachronic distance from a purported ancestor. The synchronic formula, presented in Section 2.1 deals with the synchronically compatible set, the set of wordlists sampled from. The various formulas for calculating the number of diachronically local and synchronically compatible wordlists introduced by lexical replacement, semantic changes, shifts, mergers, and links, are meant to estimate the number of wordlists that are as close or closer to the mother than the daughter, those that pass the check after sampling. This section discusses how these various diachronic formulas ought to be combined in practice.

This dissertation assumes that the arbitrariness introduced by one type of transformation is independent from the arbitrariness introduced by another type. Although this assumption makes calculation simpler, it may not be evidenced by the wordlist in question. For example, given the limited amount of data in (35), at least two different reconstructions are possible. One reconstruction might posit a shift [eː] > [iː] followed by a merger [ɛː] > [iː]; the other might posit a shift [eː] > [iː] followed by the (admittedly unlikely) semantic change ‘meet’ > ‘meat’. Both reconstructions yield the Modern English wordlist as output.

(35) *Ambiguous reconstruction*

gloss	*Middle English*	*Modern English*
meet	meːt	miːt
meat	mɛːt	miːt

If the Sound Change Formula and the Semantic Change Formula are applied separately, the Modern English wordlist in (35) would be counted in the set of diachronically local and synchronically compatible wordlists twice. It is possible to calculate the number of wordlists derivable through different combinations of different type (semantic change, sound change, lexical replacement) and subtract the result from the set of diachronically local and synchronically compatible wordlists. However, for simplicity, this dissertation assumes that wordlists are distinct if they are derived from reconstructions that yield distinct wordlists for some input. For the example in (35), the sound changes [eː] > [iː] and [ɛː] > [iː] are normally expected to yield distinct wordlist from the sound change [eː] > [iː] followed by the semantic change ‘meet’ > ‘meat’. It just happens to be the case that the daughter wordlist is too short to evidence the difference between the two reconstructions. For instance, the additional word-pair [hɛːθ]~[hiːθ] 'heath' would serve to disambiguate the two reconstructions, as it can be derived through the same sound change as [mɛːt]~[miːt] 'meat' in [ɛː] > [iː] but not the semantic change in 'meet' > 'meat'.

As such, transformations of different types (semantic change, sound change, lexical replacement) are treated as independent in the calculations, with the exception of shifts and mergers (see Section 3.3.1). Future implementations of WDT may want to address this issue in a different way, such as combining all calculations into a single formula or using additional formulas to calculate the degree of overlap.

Because the outputs of the individual diachronic formulas are treated as independent, to estimate the total number of diachronically local and synchronically compatible wordlists, one need only to multiply the number of diachronically local and synchronically compatible wordlists introduced by individual operations: lexical replacement, semantic change, mergers and shifts (one formula), sound change environments, and links. The product of all domain-specific formulas for calculating $|D \cap S|$ is formalized in (36).

(36)

$$|D \cap S| = \prod_{x \in X} |D \cap S|_x$$

Diachronic Formula: Formula for estimating the total number of diachronically local and synchronically compatible wordlists given the number of diachronically local wordlist introduced by *x* in *X*, where *X* is the set containing the number of diachronically local and synchronically compatible wordlists calculated by domain-specific formulas.

Note that the relationship between the elements in (36) is multiplicative. As such, it can be expressed in log-space. The log of $|D \cap S|$ is simply the sum of the log of individual contributions to $|D \cap S|$ from different kinds of transformations, as shown in (37).

(37)

$$\log|D \cap S| = \sum_{x} \log|D \cap S|_x$$

Although multiplication in linear space and addition in log space are identical in theory, calculations in log-space are preferable in practice. For WDT, where the raw values are extremely large, working with the formula in (37) is particularly useful. Furthermore, some of the individual formulae introduced thus far, those that capture multiplicative relationships only, allow for operations in log space themselves, namely the Synchronic Formula (Section 2.1), the Word Complexity Formula (Section 3.1.2), the Merger Formula (Section 3.2.1) and the Environment Formula (Section 3.2.1). Just as with (37), the formulae can be converted into log space for convenience.

3.7 Chapter Summary

This chapter concludes the theoretical underpinnings of Wordlist Distortion Theory by showing how to calculate the number of diachronically local and synchronically compatible wordlists introduced by sound changes. It is argued that, just as lexical replacement and semantic change, sound change affects $P(D|S)$ and alters the reliability of the reconstruction. Mergers, contrast eroding sound changes, increase $P(D|S)$ by a greater amount than do contrast preserving changes (shifts) and contrast transforming changes (links).

This chapter also discusses several corollaries of Wordlist Distortion Theory with respect to sound change. For example, it is shown that the number of sound changes that may be posited in a reconstruction without increasing $P(D|S)$ is linearly correlated with average word length in the language. Additionally, it is shown that the effect of shifts on reconstruction reliability is incredibly minor, to the point that, even in reconstruction for very small datasets, across-the-board shift incurs only a small increase to $P(D|S)$.

Finally, this chapter demonstrates that Wordlist Distortion Theory is compatible with phonological feature frameworks, though no such framework is employed in this dissertation. It is shown that the effect of individual sound changes on $P(D|S)$, when analyzed as a change operating on features, depends on the segmental inventory of the language, the distribution of features in the inventory, and the type of features used. The theoretical backbone presented in Chapters 2 and 3 is sufficient for a working analysis of comparative reconstruction, as is demonstrated in the following chapter.

Chapter 4
Case Study: Austronesian

This chapter presents a simulated annealing machine learning algorithm which generates reconstructions automatically and evaluates them using Wordlist Distortion Theory. The algorithm operates by stochastically suggesting sound changes from a mother wordlist to a daughter wordlist. Suggested sound changes are rejected or accepted based on the change that they incur to the $P(D|S)$ in the reconstruction, with preference given to changes that reduce $P(D|S)$. The algorithm was tested in a case study of 74 attested Austronesian languages and 5 Austronesian proto-languages, and further extended to 2 attested Ongan languages and 2 Ongan proto-languages. The Ongan languages were included to test the putative genetic connection between Austronesian and Ongan language families (Blevins, 2007), a hypothesis that is not widely accepted in the field (Blust, 2014). A *Python* script implementation of the simulated algorithm is made publicly available online.

With respect to the Austronesian family, the results of the case study are in line with the general understanding in the field. The reconstructions between an Austronesian proto-language and a direct descendant produced $P(D|S)$ values well below the threshold of reliability chosen. Additionally, $P(D|S)$ appears to be correlated with proto-language recency, with reconstructions from recent ancestors yielding lower $P(D|S)$ than reconstructions from distant ancestors. By comparison, an English control dataset did not elicit reliable reconstructions from any of the Austronesian proto-languages tested.

Looking at the results more closely reveals several insights about the Austronesian language family. For example, contact of the Philippine languages of Sulawesi with local non-Philippine populations appears to have noticeably eroded evidence of Philippine descent. Similarly, the sporadic appearance of a nasal/oral grade distinction in Oceanic appreciably eroded evidence of Malayo-Polynesian and Austronesian descent. The results also shed light on some of the classification debates in Austronesian: Batanic languages are argued to be Philippine, while Chamorro is argued to be Malayo-Polynesian. Overall, the evidence in support of the Austronesian proto-languages and their connection to respective descendants is overwhelming. This case study is the first to confirm the genetic relatedness between languages through a probabilistic evaluation of comparative reconstructions.

With respect to the putative Ongan-Austronesian connection, the results are not as clear. Reconstructions from Proto-Ongan-Austronesian to the Austronesian languages appear promising and yield $P(D|S)$ values below the threshold of reliability for approximately one third of the Austronesian languages tested, but not the remaining two thirds. Because the Ongan wordlists used for the proposal are not available, a precise evaluation of the reconstructions to the Ongan languages are not possible. Estimating a reasonable range of $P(D|S)$ values for these two datasets suggests that reconstructions from Proto-Ongan-Austronesian to Jarawa could be reliable, whereas reconstructions from Proto-Ongan-Austronesian to Onge are almost certainly outside the range of statistical reliability.

This chapter also presents results from an additional case study performed on the same data using a multilateral comparison algorithm (Kessler & Lehtonen, 2006; Ceolin, 2019). The results of the multilateral comparison are mostly in line with the WDT evaluations of simulated reconstructions. The *p* value as estimated using multilateral comparison is strongly correlated with $P(D|S)$ for the simulated reconstructions. However, simulated annealing coupled with WDT evaluation appears to give more conservative results in general. No correlation was found between phonetic distance and the number of sound changes required in the reconstruction, confirming that multilateral comparison and the comparative method are largely independent in focus.

This chapter differs from Chapters 2 and 3 of this dissertation, in that it does not present any new aspects of Wordlist Distortion Theory. Rather, the purpose of Chapter 4 is to showcase the utility of WDT in a computational setting. The computational implementation of WDT is a simplified version of the framework discussed so far. As in Chapter 3, sound changes are analyzed as segment-to-segment mappings. However, the distinction between mergers, shifts, and links is not maintained. Instead, the Merger Formula is used as an approximation for the effect of all sound changes, generally expected to result in a conservative estimate of $P(D|S)$ (see Section 3.4). The implementation considers lexical replacement but not semantic change. These simplifications were instituted primarily for computational convenience. Nevertheless, the results of even a relatively simple implementation appear promising.

The structure of the remainder of the chapter is as follows. Section 4.1 introduces the datasets employed in the case study, comprising data from Austronesian languages and proto-languages

(Trussel & Blust, 2010), data from Ongan languages and proto-languages (Blevins, 2007), and an English control dataset. Section 4.2 presents the case study methodology, primarily focusing on a description of the simulated annealing algorithm and its implementation. Section 4.3 discusses case study predictions. Section 4.4 presents results of the case study separated by Austronesian language branch, as well as results with respect to the putative Austronesian-Ongan. Section 4.5 presents the results of a multilateral comparison conducted on the same dataset and contrasts these results with those derived through simulated annealing. Finally, section 4.6 concludes by outlining the performance of the simulated annealing algorithm more broadly.

4.1 Language Background

The following section introduces the languages and language data analyzed in the case study. Discussion begins with the Austronesian languages. To mirror the results section, each Austronesian language group, as defined in the Austronesian Comparative Dictionary (Trussel & Blust, 2010) is introduced separately. A discussion of the Ongan language family and the dataset used to evaluate the Ongan-Austronesian hypothesis follows. Finally, an English control dataset is also introduced here.

4.1.1 Austronesian Background

The Austronesian language family is spoken by over 380 million speakers, or approximately 5% of the world's population, mostly in insular South-East Asia and Oceania, but also in the Asian mainland and as far west as Madagascar (Blust, 2013). With approximately 1200 languages (Eberhard, et al., 2023), the family is one of the largest in the world. The Austronesian languages are generally known for their relatively small phonological inventories, a property most prominent in Polynesia, where Hawaiian, for example, has been argued to exhibit just 13 phonemes in total. Consonant clusters in Austronesian languages are generally reserved to the word-medial position. The syllable structure CV(C), still found in many languages of the family, appears to be inherited from Proto-Austronesian. Roots tend to be disyllabic, and morphology agglutinating-synthetic. Reduplication and infixation are extremely common morphological processes in all of the Austronesian branches. Finally, the Austronesian languages are also known for their complex interaction with social structure, including genderlects, secret languages, ritual languages, and taboo (Blust, 2013).

The wordlists for the case study were extracted from the online version of the Austronesian Comparative Dictionary (Trussel & Blust, 2010). The data in the Austronesian Comparative Dictionary is listed in several different formats. For the current project, the *Cognate Sets* format was most useful, where each dictionary entry consists of a proto-form, in one of the Austronesian proto-languages, and all of its attested cognates. Cognacy between mother and daughter forms was, therefore, given in the dataset itself and did not need to be established independently.

In total, words and glosses were extracted from 1035 attested Austronesian languages and varieties, as well as 37 reconstructed Austronesian proto-languages. The reconstructed languages in the Austronesian Comparative Dictionary include Proto-Austronesian itself but also many lower level proto-languages of varying size, e.g. Proto-Micronesian, proto-Chamic, and proto-Lampungic. The number of entries differed by language for both attested and proto-languages. The best represented attested language was Malay, with 2332 entries, while the best represented proto-language was Proto-West-Malay-Polynesian, with 3681 entries. A total of 167 languages and 7 proto-languages were represented by only a single entry.

To ensure that wordlists contained enough data for accurate reconstruction as well as word complexity estimation, Austronesian languages and proto-languages were included in the case study only if their wordlists contained 300 or more entries. In total, 74 attested Austronesian languages met this criterion and were included in the case study. The Austronesian Comparative Dictionary sorts Austronesian languages into six groups, roughly according to their place in the Austronesian family tree. Table 2 presents the Austronesian groups along with the language count for the current case study. The six groups are a convenient way to delineate the Austronesian family and comprise a mix of ‘true’ language families (Philippine, Oceanic, South Halmahera-West New Guinea) and groupings which are known, or have been argued to be, paraphyletic (Formosan, Western, Central). No South Halmahera-West New Guinea language was included in the case study, since no wordlist in the group reached 300 entries.

Table 2: Austronesian groups in the case study

group	*number*	*example*
Philippine	29	Tagalog
Western (not Philippine)	20	Malay
Oceanic	15	Fijian
Central	6	Manggarai
Formosan	4	Paiwan
South Halmahera-West New Guinea	0	Buli

As in the case of attested languages, the proto-languages varied in number of entries. A total of 5 proto-languages were represented by at least 300 entries and were included in the case study. The proto-languages are given in Table 3 along with the number of reconstructed forms per proto-language.

Table 3: Proto-language attestation

proto-language	*entries*
Proto-Austronesian	1769
Proto-Malayo-Polynesian	2790
Proto-West-Malayo-Polynesian	3681
Proto-Philippine	1971
Proto-Oceanic	1746

Figure 15 displays the internal structure of the Austronesian language family (Trussel & Blust, 2010). Only divisions that are relevant to either the proto-languages or the 5 Austronesian groups included in this case study are shown, and much of the lower-level structure of the family is omitted. Proto-Austronesian is the ancestor of all of the languages in the Austronesian database. The Formosan languages do not constitute a valid genetic subgrouping and are directly descended from Proto-Austronesian. Proto-Malayo-Polynesian is the ancestor of all Austronesian languages outside of Taiwan. Within the Malayo-Polynesian languages, most of the Austronesian languages west of Melanesia are descended from Proto-West-Malayo-Polynesian. The West Malayo-Polynesian languages also include the Philippine languages, which can be traced to a more recent ancestor in Proto-Philippine. Some languages of the Lesser Sunda Islands are classified as Central Malayo-Polynesian and not descendants for Proto-West-Malayo-Polynesian. Most of the Austronesian languages of Oceania are also not descendants of Proto-West-Malayo-Polynesian but do share a common ancestor in Proto-Oceanic.

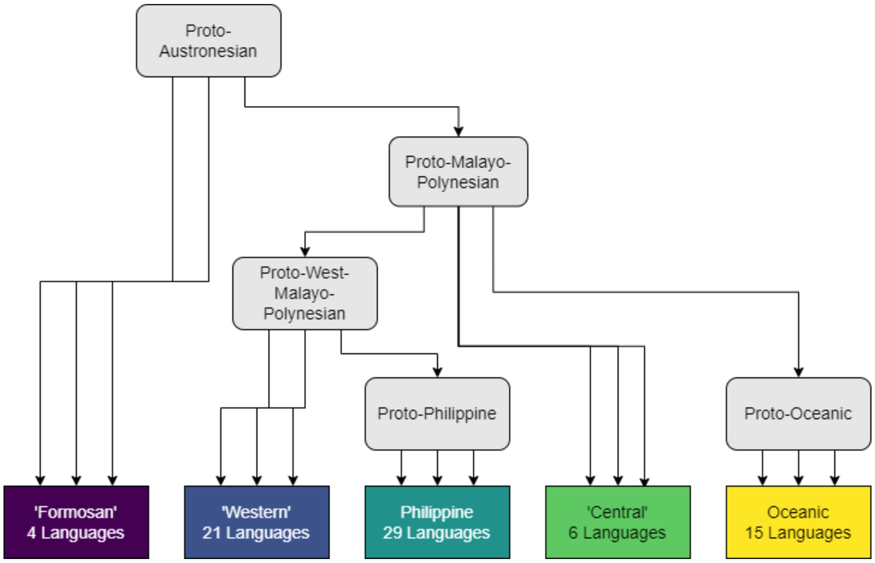


Figure 15: *Austronesian tree*. The genetic relationship between the five Austronesian proto-languages and five subgroups of Austronesian in the current case study.

Recall that the likelihood that a reconstruction can be substantiated by a random wordlist, $P(D|S)$, is affected by both diachronic and synchronic factors. Diachronic transformations – sound changes, replacements, semantic changes, etc. – determine the number of diachronically local and synchronically compatible wordlists, the numerator of the General Formula (see Section 1.2). Identifying these changes is the function of the simulated annealing algorithm presented later in the chapter. The denominator of the General Formula, the number of synchronically local wordlists in the language, is determined by synchronic and practical factors, namely mean word length w, mean number of homotopes $\hbar$, and wordlist size t.

In the languages and language groups in the case study, synchronic factors varied greatly. To illustrate, Figure 16 presents mean wordlist length by Austronesian group. As can be seen, languages in the Philippine and Western groups are well attested in the dataset, with wordlists of 894 and 804 entries on average respectively, whereas languages in the Formosan, Oceanic, and Central groups are not as well attested, with wordlists of 447, 444, and 391 entries on average respectively. Even under the assumption that the proto-languages and attested languages in the dataset are equally related, it is still the case that reconstructions are easier to justify with more word-pairs available for comparison. Therefore, $P(D|S)$ is expected to be lowest for the Philippine languages and highest for Central languages overall.

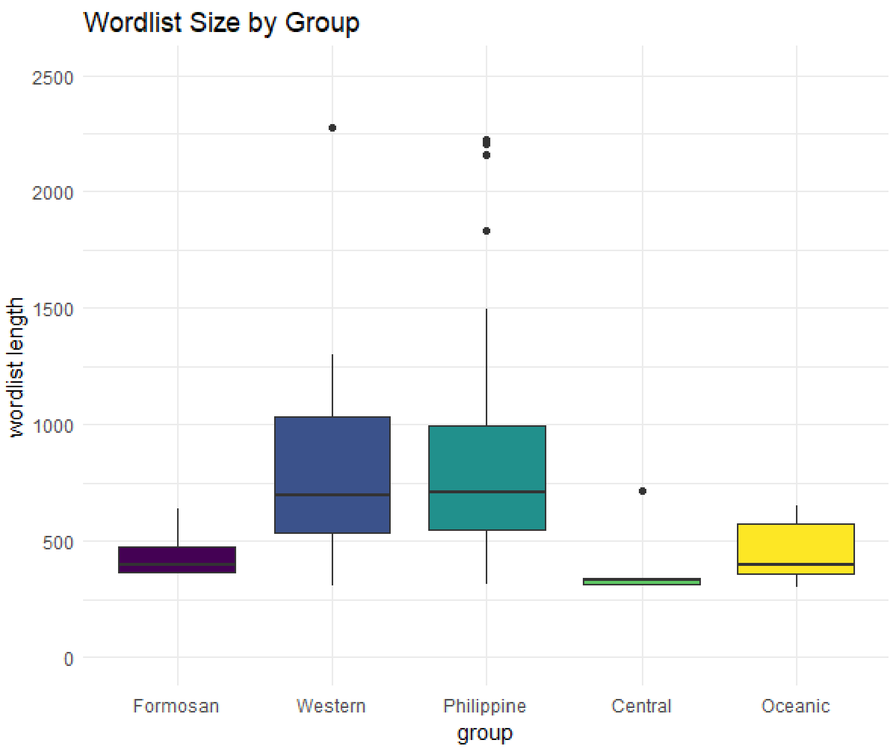


Figure 16: *Wordlist size*. Boxplots of wordlist length *t* (y-axis) for the 74 languages in the case study by language group (x-axis). The boxes cover the interquartile range, with the whiskers covering the entire range excluding outliers; outliers are represented by dots. The horizontal line within the boxes corresponds to the median.

The mean number of homotopes per wordlist and mean word-length can be combined into a single word complexity measure *c* (see Section 3.1.2), which is an estimate of how many phonological forms of mean word length are possible in the language. Figure 17 presents the average word complexity for the Austronesian groups.

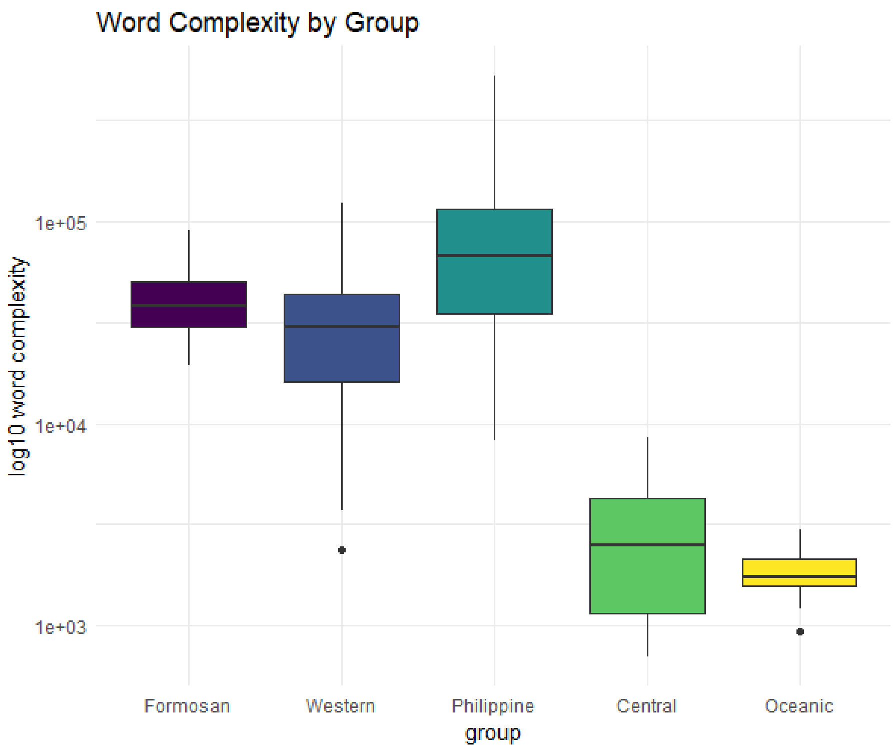


Figure 17: *Word complexity*. Boxplots of word complexity c (y-axis, logarithmic) for the 74 languages in the case study by language group (x-axis).

As can be seen in Figure 17, languages in the Philippine, Western, and Formosan groups exhibit a high word complexity, whereas languages in the Oceanic and Central groups exhibit a low word complexity. Note that the figure is logarithmic, and the differences in the plot appear deceptively small. To illustrate, the Central language Ngadha exhibits an estimated word complexity of 699, the lowest in the dataset, approximately 750 times smaller than for the Philippine language Bikol, which exhibits a word complexity of 523691, the highest in the dataset. Effectively, each word-pair reconstructed in Bikol decreases $P(D|S)$ by roughly the same factor as 2 word-pairs reconstructed in Ngadha, since $\log_{699} 523691 = 2.011$.

Due to confounding variables such as wordlist size and word complexity, it makes little sense to compare $P(D|S)$ by proto-language across language groups, e.g. Oceanic vs Philippine with respect to Proto-Malayo-Polynesian. It is difficult to determine if a low $P(D|S)$ for one group as opposed to another should be attributed to a closer genetic relationship with the mother, higher word complexity, or greater wordlist size. Instead, this chapter splits hypotheses and results by group as defined in the Austronesian Comparative Dictionary.

4.1.1.1 Philippine Group Background

The case study includes data from 29 Philippine languages. The full list of these is given in Table 4, along with the major Philippine sub-branch (Blust, 2013), the wordlist length *t*, mean number of homotopes *ℏ*, and mean word length *w*. Mean wordlist length for the Philippine languages in the case study is 893 (range: 311 – 2226). Mean number of homotopes for the Philippine languages in the case study is 7.56 (range: 5.92 – 8.99). Mean word length for the Philippine languages in the case study is 5.49 (range: 4.71 – 6.31). The Philippine group exhibits the highest average wordlist length, word length, and mean homotopy among the Austronesian branches in the case study.

Table 4: Philippine languages in the case study

language	*branch*	*t*	*ℏ*	*w*
Itbayaten	Batanic	998	7.54	5.81
Yami	Batanic	324	7.00	5.05
Tboli	Bilic	335	6.80	4.71
Tiruray	Bilic	563	6.93	5.35
Ayata Abellan	Central Luzon	311	6.82	4.99
Kapampangan	Central Luzon	588	7.82	5.52
Agutaynen	Central Philippine	535	6.53	6.20
Aklanon	Central Philippine	1135	8.53	5.32
Bikol	Central Philippine	1835	8.05	6.31
Binukid	Central Philippine	549	7.27	5.48
Cebuano	Central Philippine	2205	8.43	5.68
Hanuno	Central Philippine	1007	7.76	5.43
Hiligaynon	Central Philippine	526	7.47	5.72
Tagalog	Central Philippine	2226	8.22	5.73
Tausug	Central Philippine	547	6.85	5.81
Manobo (Western Bukidnon)	Central Philippine	820	7.74	5.25
Mansaka	Central Philippine	665	7.14	5.13
Maranao	Danao	1494	7.13	5.06
Mongondow	Gorontalic	771	7.02	5.61
Tontemboan	Minahasan	384	6.66	5.30
Bontok	Northern Luzon	911	7.54	5.48
Casiguran (Dumagat)	Northern Luzon	923	8.60	5.21
Ibaloy	Northern Luzon	570	7.18	6.06
Ifugaw	Northern Luzon	643	8.54	5.06
Ilokano	Northern Luzon	2159	8.99	5.70
Isneg	Northern Luzon	837	7.99	5.79
Kankanaey	Northern Luzon	709	8.58	5.38
Sangir	Sangiric	630	5.92	5.88
Pangasinan	Southern Cordilleran	715	8.12	5.15

Figure 18 depicts the distribution of the Philippine branches on a map of the Philippines and neighboring islands. The case study includes all of the major branches of the Philippine group. Non-Philippine Western languages spoken on the Indonesian islands of Borneo and Sulawesi to the south are not depicted (See Section 4.1.1.3). Also not depicted are the languages in the Central group, spoken on the Maluku islands to the south-east (see Section 4.1.1.4), and languages in the Formosan group, spoken on Taiwan to the north (see Section 4.1.1.5). The non-Austronesian languages of the Chinese mainland are left blank.

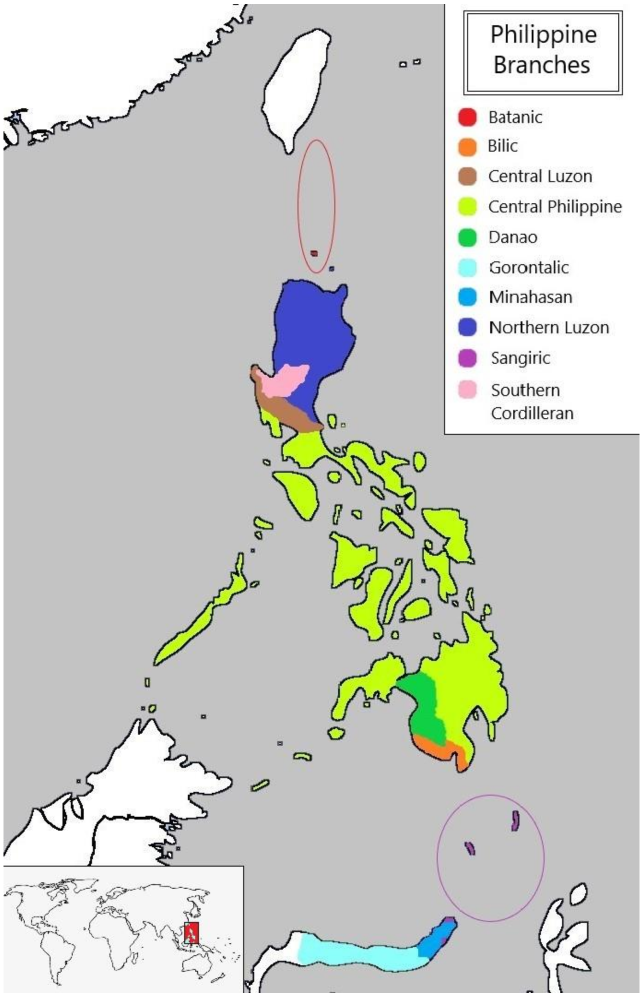


Figure 18: *Philippine branches*. Map of the primary branches of the Philippine language family covered in the case study. Boundaries are approximate but are in line with the available descriptive literature (Sneddon, 1989; Noorduyn, 1991; Blust 2013; Robinson & Lobel, 2013). Areas with non-Philippine languages are left blank.

As can be seen in Figure 18, Central Philippine is the most geographically widespread of the Philippine branches and is spoken on all of the major Philippine islands. The Northern Luzon, Central Luzon, and Southern Cordilleran languages are spoken in the north, on the largest Philippine island of Luzon. The Bilic and Danao branches are spoken in the south, on the second largest island of Mindanao.

The Gorontalic, Sangiric and Minahasan languages are generally not spoken in the Philippines. The speakers of these languages reside either on the Indonesian island of Sulawesi or, in the case of Sangir, on the Sangir islands chain just north of it. All three branches have traditionally been regarded as Philippine in the literature (Reid, 1971; Snodden, 1984; Blust, 2013). It has been argued that the Minahasan and Sangiric populations migrated to Sulawesi from Mindanao, while the Gorontolic speakers came from elsewhere in the Philippines at another date (Blust, 2019). While the northern peninsula of Sulawesi is inhabited predominately by speakers of Philippine languages, the remainder of the island is home to non-Philippine languages of the Western Malayo-Polynesian group, such as Wolio, Bare’e, Ta’e, Makassarese, and others (see Section 4.1.1.3).

The Batanic branch is spoken on the Batanes islands north of Luzon and, in the case of Yami, on the Tau Island in Taiwan. Because the Batanes islands are likely to be the first step in the migration of the Malayo-Polynesian speakers out of Taiwan, it has long been assumed that the Batanic languages split relatively early in the Austronesian family tree. Thus, proto-Batanic has been argued to be sister to Proto-Malayo-Polynesian, or a primary branch of Malayo-Polynesian (Ross, 2005). However, more recent studies re-evaluate Batanic as Philippine (Blust, 2013:747; Gallego, 2014; Ross, 2020). The current understanding is that the languages spoken on the islands are not descendent from the original Austronesian population of the Batanes but rather the result of a subsequent spread out of Luzon northward.

4.1.1.2 Oceanic Group Background

The case study includes data from 15 Oceanic languages. The full list of these is given in Table 5, along with the major Oceanic sub-branch (Blust, 2013), the wordlist length *t*, mean number of homotopes $\hbar$, and mean word length *w*. Mean wordlist length for the Oceanic languages in the case study is 444 (range: 301 – 652). Mean number of homotopes for the Oceanic languages in the case study is 5.40 (range: 4.87 – 6.51), the lowest among Austronesian groups in the case study. Mean word length for the Oceanic languages in the case study is 4.45 (range: 4.09 – 4.70).

Table 5: Oceanic languages in the case study

language	*branch*	*t*	*ħ*	*w*
Fijian	Central Pacific	617	5.39	4.70
Tolai	Meso-Melanesian	301	6.46	4.12
Motu	Papuan Tip	382	5.28	4.46
Hawaiian	Polynesian	375	4.87	4.32
Maori	Polynesian	399	5.02	4.41
Niue	Polynesian	358	5.07	4.59
Rennellese	Polynesian	423	5.10	4.59
Samoan	Polynesian	572	5.11	4.37
Tongan	Polynesian	591	5.44	4.52
Are’are	Southeast Solomonic	301	4.92	4.61
Arosi	Southeast Solomonic	652	5.39	4.43
Lau	Southeast Solomonic	368	5.39	4.39
Nggela	Southeast Solomonic	560	5.26	4.67
Sa’a	Southeast Solomonic	418	5.77	4.56
Mota	Torres-Banks	336	6.51	4.09

Figure 19 depicts the distribution of the Oceanic branches on a map of Melanesia and Oceania. Because the Polynesian branch is stretched out over a large part of the Pacific, the easternmost extent of this branch is omitted for clarity. Oceanic branches not represented in this case study, such as the East Vanuatu branch in Vanuatu and the South New Caledonian branch in New Caledonia, are not depicted. Also not depicted are the non-Austronesian languages of mainland Papua New Guinea and Australia.

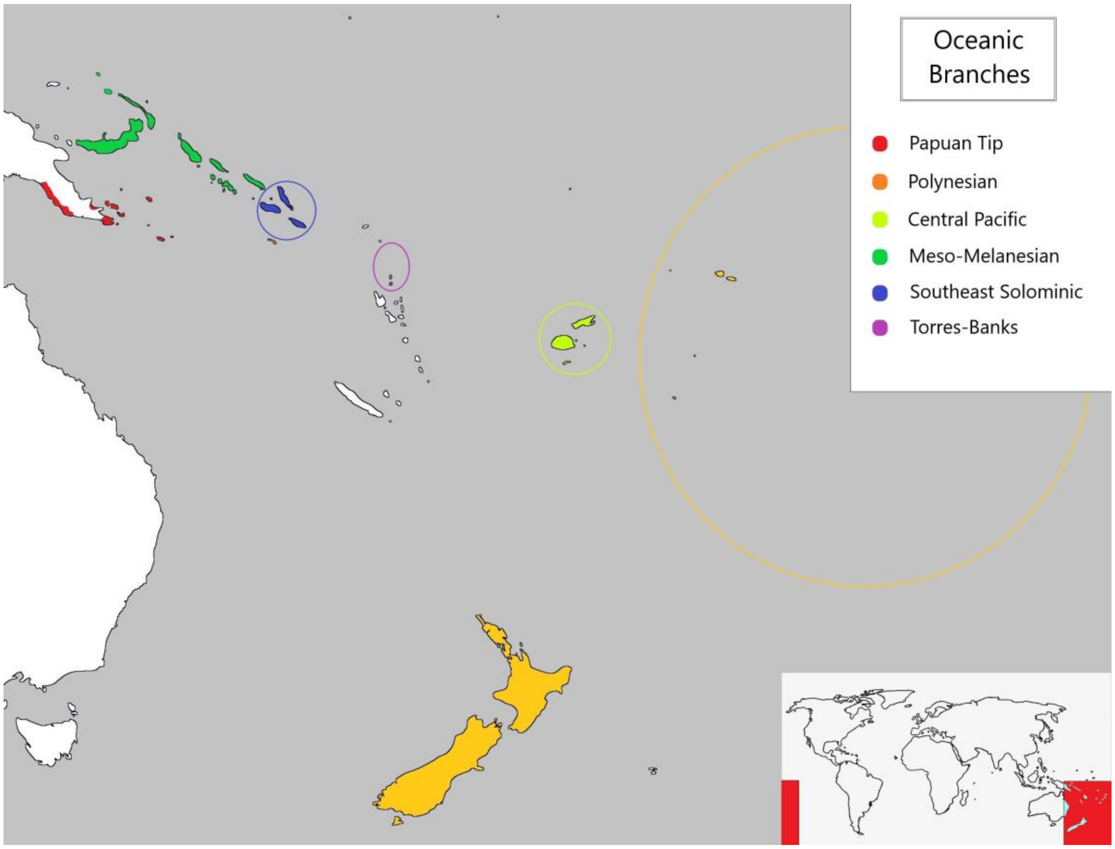


Figure 19: *Oceanic branches*. Map of the primary branches of the Oceanic language family covered in the case study. Boundaries are approximate but are in line with the available descriptive literature (Lynch et al., 2001; Terrill, 2011; Blust 2013; Guerin, 2017). Areas with non-Oceanic languages are left blank. Oceanic branches not covered in the case study are not depicted. The easternmost regions of the Polynesian branch are omitted for clarity.

As can be seen in Figure 19, the Papuan Tip branch of Oceanic is spoken on the eastern tip of the Island of New Guinea. The remainder of the island is mostly inhabited by speakers of non-Austronesian Papuan languages; contact between Papuan and Oceanic languages in the area is well-established (Terrill, 2011; Barlow, 2023). The Meso-Melanesian branch is spoken on the New Ireland, New Britain, and Bougainville islands of Papua New Guinea, as well as the northern half of the Solomon Islands, immediately east of New Guinea. East of the Solomon islands, on the islands of Torres and Banks in northern Vanuatu, one can find the Torres-Banks branch of Oceanic. The Central Pacific branch is spoken in Fiji and the nearby Rotuman, further east into Oceania. The Polynesian branch is spread out over the numerous equatorial islands of the Pacific, as well as New Zealand.

4.1.1.3 Western Group Background

Technically, the Western Malayo-Polynesian language group includes the Philippine languages as well. However, as done in the Austronesian Comparative Dictionary (Trussel & Blust, 2010), this section recounts only the Western Malayo-Polynesian languages that are not Philippine, i.e. descendants of Proto-Western-Malayo-Polynesian but not Proto-Philippine. The languages do not share an ancestor to the exclusion of Philippine.

Moreover, it must be noted that Proto-Western-Malayo-Polynesian itself is not without controversy. It is often assumed in the field that the Austronesian language family is 'right-branching' and that the eastward migration of Austronesian peoples resulted in a situation where the easternmost languages can be grouped into progressively larger families within Austronesian, while the western languages generally do not constitute a valid grouping. Therefore, the idea of a single Western Malayo-Polynesian group has attracted skepticism in the literature (Ross, 2005; Klamer, 2019). Moreover, the few innovations that Western-Malayo-Polynesian languages do share are chiefly morphological, rather than phonological, and have also been attributed to language contact rather than descent (Blust, 2013:31). Under the hypothesis that the Western grouping is areal rather than genetic, the primary branches of Western-Malayo-Polynesian are actually primary branches of Malayo-Polynesian itself. Nevertheless, because the Austronesian Comparative Dictionary (Trussel & Blust, 2010) lists 'Western' as a separate group within Austronesian while also including a Proto-West-Malayo-Polynesian wordlist, Western Malayo-Polynesian is assumed to be a genetic grouping here as well.

The case study included data from 20 Western non-Philippine languages. The full list of the these is given in Table 6, along with the major sub-branches (Blust, 2013), the wordlist length t, mean number of homotopes $\hbar$, and mean word length w. Mean wordlist length for the Western languages in the case study is 804 (range: 307 – 2276). Mean number of homotopes for the Western languages in the case study is 6.70 (range: 5.15 – 7.69). Mean word length for the Western languages in the case study is 5.4 (range: 4.7 – 6.0).

Table 6: Western Malayo-Polynesian languages in the case study

language	*branch*	*t*	*ħ*	*w*
Balinese	Bali-Sasak	1027	7.51	5.17
Sasak	Bali-Sasak	803	7.29	5.19
Karo Batak	Barrier Island-Batak	801	7.44	5.55
Toba Batak	Barrier Island-Batak	1087	7.06	5.98
Kadazan Dusun	Dusunic	524	6.90	5.44
Malagasy	Greater Barito	594	5.31	5.98
Ngaju Dayak	Greater Barito	680	7.01	5.39
Bare'e	Kaili-Pamona	644	5.40	4.88
Tae'	Kaili-Pamona	708	6.10	5.55
Kayan	Kayan-Murik-Mudang	463	6.62	4.73
Iban	Malayo-Chamic	1162	7.13	5.24
Malay	Malayo-Chamic	2276	7.49	5.82
Kelabit	North Sarawak	541	7.05	4.92
Buginese	South Sulawesi	307	5.71	5.17
Makassarese	South Sulawesi	661	5.51	5.94
Wolio	Wotu-Wolio	380	5.15	4.75
Chamorro	unclear	316	6.59	5.15
Javanese	unclear	1069	7.69	5.17
Old Javanese	unclear	1299	7.47	5.73
Sundanese	unclear	742	7.58	5.55

Figure 20 depicts the distribution of the Western (non-Philippine) branches on a map of Indonesia and Malaysia. The Malagasy language, a Greater Barito language spoken on Madagascar, and Chamorro, an unclassified language spoken on the Mariana islands in Oceania, are too far away and are omitted for clarity. Western branches not represented in this case study, such as the Ida'an branch in Malaysian Borneo and the Lampungic branch in Sumatra, are not depicted. Also not depicted are the Philippine languages in the Philippines to the north (see Section 4.1.1.1) and the Central languages in eastern Indonesia (see Section 4.1.1.4). Non-Austronesian languages of mainland South-East Asia are left blank.

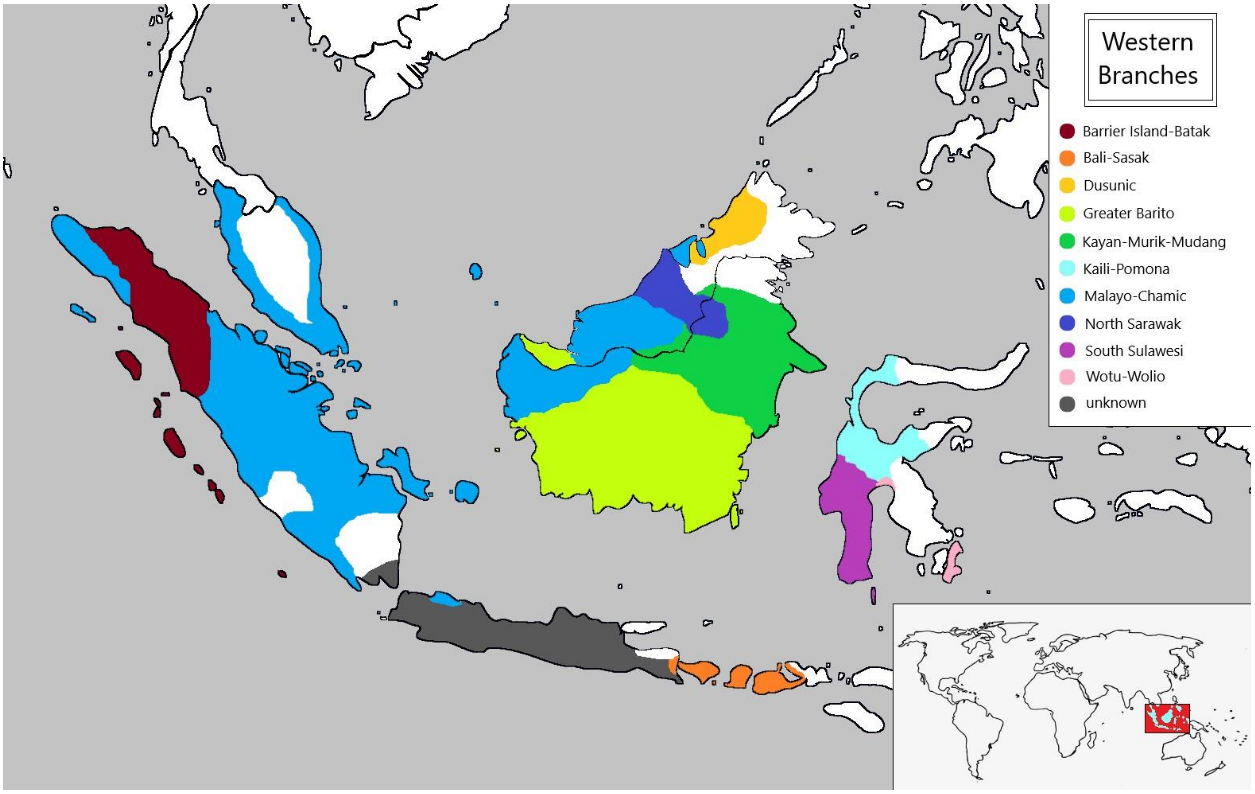


Figure 20: *Western branches*. Map of the primary branches of the Western language family covered in the case study (Philippine languages excluded). Boundaries are approximate but are in line with the available descriptive literature (Benjamin, 2012; Blust 2013; Smith, 2017). Areas with non-Western languages are left blank. Madagascar, inhabited by speakers of Malagasy (Greater Barito), and the Mariana islands, inhabited by speakers of Chamorro (unclassified), are omitted for clarity.

As can be seen in Figure 20, the Malayo-Chamic branch is the most widespread branch of Western languages, spoken on the major Indonesian islands of Borneo, Sumatra, and Java, as well as the Malaysian peninsula. The remainder of the Malaysian peninsula is inhabited by non-Austronesian speakers of the Aslian branch of Austroasiatic. The Greater Barito Group is also widespread, with Ngaju Dayak spoken on Borneo and Malagasy on Madagascar (not shown in Figure 20). The island of Borneo is also inhabited by speakers of the Kayan-Murik-Mudang branch, the North Sarawak branch, and the Dusunic Branch. The Kaili-Pamona, South Sulawesi and Wotu-Wolio branches are spoken on the island of Sulawesi, east of Borneo. The Barrier Island-Batak branch is spoken on the island of Sumatra, west of Borneo. The Bali-Sasak languages are spoken on the Indonesian Lower Sundanese islands, immediately north-west of Australia.

The classification of Sundanese among the Western Malayo-Polynesian languages is unclear. The language is spoken by over 40 million people all over Indonesia, but mostly on the Islands of Java and Sumatra in the south-west. Sundanese is uncontroversially a Western Malayo-Polynesian language (to the extent that the grouping itself is uncontroversial) and appears to be closely related to the Greater Barito and Malayo-Chamic languages of Borneo, though the exact relationship between these is unknown (Blust, 2013).

The issues surrounding the classification of Javanese are similar. The language, spoken by over 80 million primarily on the island of Java, is commonly assumed to be a close relative of the Malayo-Chamic languages, though the precise relationship between the two is unclear (Blust, 2013). Old Javanese is a well-attested literary language of the Javan Middle Ages, heavily influenced by Sanskrit. Javanese and Old Javanese are listed separately in the Austronesian Comparative Dictionary and are, therefore, kept separate here. Old Javanese is the only non-living variety included in the case study.

The classification of Chamorro is contentious. The language is spoken on the Mariana Islands (not shown in Figure 20) and is one of the two languages of Oceania commonly considered non-Oceanic. It is now agreed upon that the Chamorro is neither Formosan nor Philippine. However, it is debated whether Chamorro is a sister to Malayo-Polynesian (Starosta, 1995) or a primary branch of Malayo-Polynesian (Blust, 2000; Reid, 2002). The Austronesian Comparative Dictionary lists the language as 'Western'. However, this appears to be largely because the more widely accepted positions of Chamorro within the Austronesian tree do not fit neatly into any of the major groups.

4.1.1.4 Central Group Background

Like Proto-Western-Malayo-Polynesian, doubt has been cast on the proto-Central-Malayo-Polynesian hypothesis (Klamer, 2019). However, because proto-Central-Malayo-Polynesian is not included in the current dataset, this debate cannot be addressed here. For our purposes, it is enough to say that Central languages are descended from Proto-Malayo-Polynesian but not Proto-West-Malayo-Polynesian, a fact which is by itself uncontroversial. Whether the Central languages also share an ancestor more recent than Proto-Malayo-Polynesian cannot be ascertained without a sufficiently large proto-Central-Malayo-Polynesian wordlist.

The case study includes data from 6 Central languages. The full list of these is given in Table 7, along with the major sub-branches (Blust, 2013), the wordlist length *t*, mean number of homotopes *ħ*, and mean word length *w*. Mean wordlist length for the Central languages in the case study is 391 (range: 311 – 714) the lowest among the Austronesian groups in the case study. Mean number of homotopes for the Central languages in the case study is 5.83 (range: 4.67 – 7.00). Mean word length for the Central languages in the case study is 4.43 (range: 3.92 – 4.88), the lowest among Austronesian groups in the case study.

Table 7: Central Malayo-Polynesian languages in the case study

language	*branch*	*t*	*ħ*	*w*
Tetun	Central Timor	314	6.05	4.65
Kambera	Sumba-Hawu	340	4.68	4.88
Buruese	West Central Maluku	311	6.38	4.38
Manggarai	West Flores	714	7.00	4.65
Ngadha	West Flores	341	5.32	3.92
Rotinese	West Timor	327	5.56	4.11

Figure 21 depicts the distribution of the Central branches on a map of the Lesser Sunda Islands and Moluccas in Indonesia. Central branches not represented in this case study, such as the Piru Bay branch on Ambon Island and the Three Rivers branch on Seram island, are not depicted. Also not depicted are the Western language of Bali, Sulawesi and Borneo to the west (see Section 4.1.1.3). Non-Austronesian Papuan languages spoken on some of the islands are left blank.

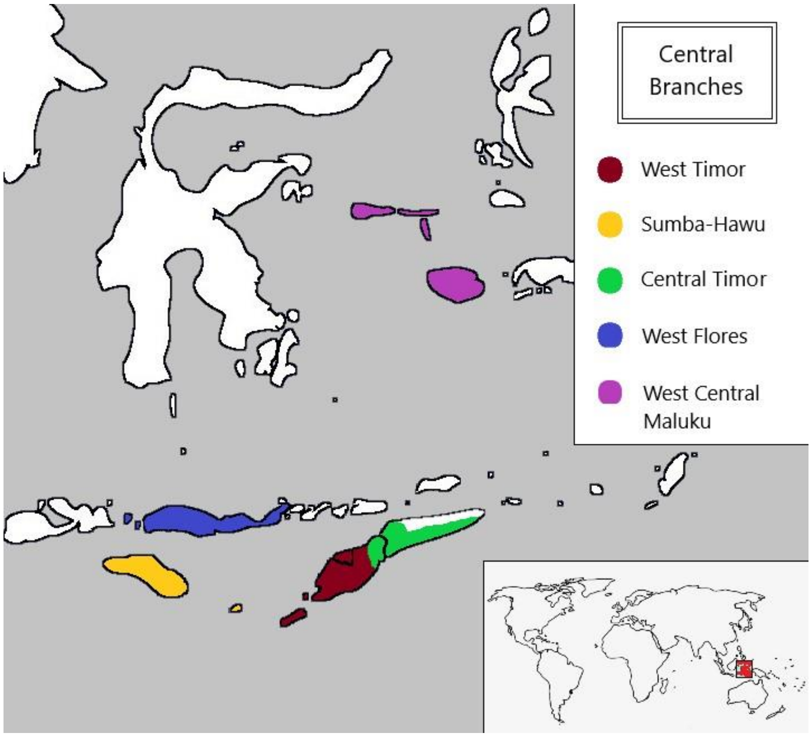


Figure 21: *Central branches*. Map of the primary branches of the Central language family covered in the case study. Boundaries are approximate but are in line with the available descriptive literature (Blust 2013; Kalping & Klamer, 2018). Areas with non-Central languages are left blank.

As can be seen on the map in Figure 21, the West Flores and Sumba-Hawu branches are spoken on the islands of the same name in the Lesser Sunda Islands of southern Indonesia. The West Timor and Central Timor languages are spoken on the island of Timor, just east of The Lesser Sunda Islands. The West Central Maluku languages are spoken on Buru and Maluku islands between Sulawesi and New Guinea, north of Timor.

4.1.1.5 Formosan Group Background

The group known as ‘Formosan’ comprises the Austronesian languages spoken on the island of Taiwan (formerly ‘Formosa’). Yami, spoken in Tau Island just off the coast of Taiwan, is an exception, as it is a Philippine language of the Batanic branch. The Formosan languages are not closely related and are grouped together for convenience. These languages are simply the descendants of the Austronesian varieties which were spoken in Taiwan before the Malayo-Polynesian expansion approximately 4,000 years ago (Blust, 2019). Therefore, the Formosan branches comprise, along with Malayo-Polynesian, the primary branches of the Austronesian language family.

The case study includes data from 4 Formosan languages. The full list of these is given in Table 8, along with the major sub-branches (Blust, 2013), the wordlist length t, mean number of homotopes $\hbar$, and mean word length w. Mean wordlist length for the Formosan languages in the case study is 447 (range: 360 – 635). Mean number of homotopes for the Formosan languages in the case study is 7.10 (range: 6.24 – 7.56). Mean word length for the Formosan languages in the case study is 5.42 (range: 4.98 – 5.73).

Table 8: Formosan languages in the case study

language	*branch*	t	$\hbar$	w
Amis	East Formosan	425	7.58	5.25
Kavalan	East Formosan	368	7.27	4.98
Paiwan	Paiwan	635	7.34	5.73
Thao	Western Plains	360	6.24	5.71

Figure 22 depicts the distribution of the Formosan branches on a map of Taiwan. Formosan branches not represented in this case study, such as the Tsouic and Bunun branches spoken in central Taiwan, are not depicted. Non-Austronesian languages spoken of the Chinese mainland are left blank.

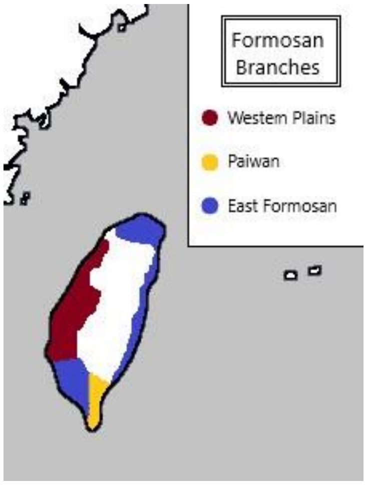


Figure 22: *Formosan branches*. Map of the primary branches of the Formosan languages in Taiwan covered in the case study. Boundaries are approximate but are in line with the available descriptive literature (Blust 2013). Areas with non-Austronesian languages are left blank.

4.1.2 Ongan Background

The Ongan language consists of two sister languages, Jarawa and Onge. As can be seen on the map in Figure 23, the language family is spoken on the Indian Andaman Islands in the Indian ocean north of the Indonesian island of Sumatra. Jarawa is spoken on the Middle and South Andaman islands, while Onge is spoken on the Little Andaman Island. Sentinelese, spoken on the North Sentinel Island nearby is often presumed to be Ongan, though this is unconfirmed due to a lack of outside communication with the native population of the island (Blevins, 2007). The Ongan language family is not to be confused with the Great Andamanese Language family, spoken on the northern half of the Andaman Islands. The two families appear to have little in common and are considered to be unrelated (Abbi, 2009; Blevins, 2020).

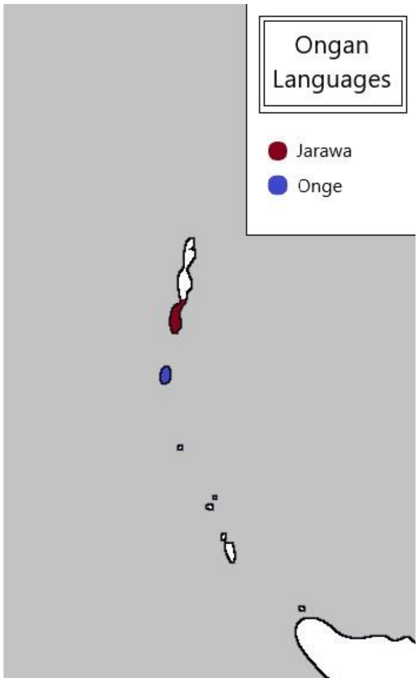


Figure 23: *Ongan languages*. Map of the two Ongan languages in the Andaman islands. Boundaries are approximate but are in line with the available descriptive literature (Blevins, 2007). Areas with non-Ongan languages are left blank.

The Ongan languages have relatively small consonant inventories with no oral fricatives (Blevins, 2007). The two languages exhibit a CV(C) syllable template, with minimally bimoraic roots (Blevins, 2007). Ongan morphology can be described as agglutinating (Abbi, 2009).

The current case study explores the putative Austronesian-Ongan connection. The Austronesian-Ongan hypothesis is argued primarily on the basis of lexicon and morphology. Additionally, it has been noted the Ongan people and the pre-Austronesian inhabitants of the Philippines share some anthropological features (Blevins, 2007). The connection Austronesian-Ongan connection is not the only macro-family proposal to include Austronesian. Other hypotheses include Austronesian-Austroasiatic (Reid, 1994; Reid, 1999), Austronesian-Japanese (Kawamoto, 1984), and Austronesian-Kra-Dai (Benedict, 1976). None of these macro-family proposals, including Ongan-Austronesian, are currently accepted by specialists (Blust, 2013; Blust, 2014).

Crucial to the current project are the Proto-Ongan-Austronesian and Proto-Ongan forms included in the proposal, along with cognates in the two attested Ongan languages where available. A total of 109 cognate sets are given in the original paper (Blevins, 2007). Not all of these include proto-forms and, after removing duplicates and Austronesian forms not found in the Austronesian Comparative Dictionary (Trussel & Blust, 2010), the remaining 83 Proto-Ongan-Austronesian and Proto-Ongan forms were added to the dataset.

According to the hypothesis, the Ongan language family and Austronesian languages family are sisters, as is illustrated in Figure 24. Thus, all descendants of Proto-Austronesian are also presumed to be descendants of Proto-Ongan-Austronesian.

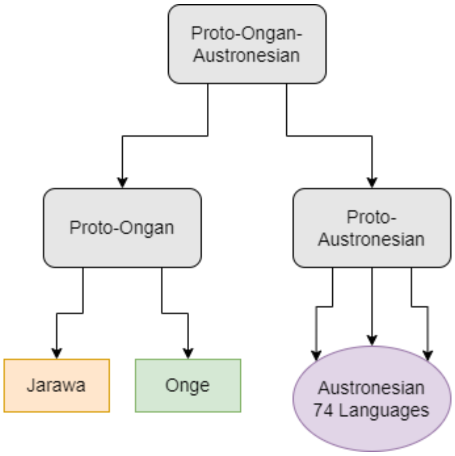


Figure 24: *Austronesian-Ongan tree*. The hypothesized genetic relationship between the Austronesian and Ongan languages used in the case study.

The two Ongan languages used in the case study are listed in Table 9 the wordlist length *t*, mean number of homotopes *ħ*, and mean word length *w*.

Table 9: Ongan languages in the case study

language	*t*	*ħ*	*w*
Jarawa	>61	5.13	4.28
Onge	>85	4.43	4.71

The Ongan languages in the case study have particularly low values for mean number of homotopes. Only four Austronesian languages in the case study exhibit a smaller mean number of homotopes than Jarawa, and no Austronesian language in the dataset exhibits a mean number of homotopes smaller than Onge at 4.43. Combined with the relatively low word-length in the two languages, the low homotopy yields a low word complexity. As a result, even more so than for the Oceanic and Central languages in the Austronesian datasets, these languages require more data to evidence reliable reconstructions than languages with higher word complexity or higher number of attested forms.

Notice that wordlist length *t* is listed as a lower bound in Table 9. The original proposal only lists forms in Jarawa and Onge if they are presumed to be cognate with one of the forms in Proto-Ongan-Austronesian (Blevins, 2007). Thus, forms without a discernable cognate in the Proto-Ongan-Austronesian wordlist were not included. While this is common practice in the field, the lack of complete daughter wordlists is particularly detrimental to a probabilistic analysis such as in Wordlist Distortion Theory. Recall that, as discussed in Section 2.2, evaluating reconstructions to daughter wordlists containing only forms derivable through sound change is a form of cherry-picking, as it assumes that no lexical replacement took place.

To properly evaluate reconstructions to Jarawa and Onge using Wordlist Distortion Theory, in addition to the forms that can be derived from Proto-Ongan-Austronesian through regular sound change it is crucial to know the forms that cannot be derived through regular sound change i.e. those that require lexical replacement as an explanation. The numbers presented in Table 9 correspond to the number of Jarawa and Onge forms given in the paper, whereas the number of Jarawa and Onge forms consulted for the reconstruction remains unknown. Section 4.4.2.2. presents a method of circumventing this issue by providing a range of estimates for $P(D|S)$ in reconstructions involving these two languages.

4.1.3 Control

The case study also included English control wordlists. The purpose of the controls was to establish that the methodology presented in the case study is able to yield a negative result where a negative result is expected. Save for the Proto-World hypothesis (Ruhlen, 1994), not accepted in the field, to my knowledge no proposal links the Indo-European language family (containing English) to the Austronesian language family. Therefore, a reliable reconstruction from any of the proto-languages in the case study to the control wordlists is unlikely.

Because the purpose is to showcase a negative result, there is no requirement for the word-pairs between the control and the proto-languages to match in meaning. As such, word-pairs were constructed by matching the index of the proto-form in the Austronesian Comparative Dictionary (Trussel & Blust, 2010)[1] with the word position in the first two chapters of the fantasy novel *Lord of the Rings* by J.R.R. Tolkien. Punctuation and capitalization in the control 'descendant' were removed but the forms were otherwise kept as in the classic. For the purposes of a random control, English orthography serves the same function as a phonetic transcription.

As an example, Proto-Austronesian *alikas 'quick, fast', entry 190 in the Austronesian Comparative Dictionary, corresponds to [bilbo], the 190th word in *Lord of the Rings*. Likewise, Proto-Austronesian *baNar 'a thorny vine', entry 1115 in the Austronesian Comparative Dictionary, corresponds to [cousin], the 1115th word in the *Lord of the Rings*.

For each of the proto-languages tested, word-pairs were constructed only for indices contained in the proto-language. For example, entry 1115 in the Austronesian Comparative Dictionary 'a thorny vine' contains Proto-Austronesian *baNar and proto-Malayo-Polynesian *banaR but no entries for the other proto-languages in the case study. As such, both PAN *baNar and PMP *banaR were compared to the form [cousin] from *Lord of the Rings*, but the entry was not included in reconstructions from Proto-Oceanic, Proto-Philippine, or Proto-West-Malayo-Polynesian. Technically, this is a form of cherry-picking since the wordlist entries should be independent of the data (see Section 2.2). However, because no effort was made to choose phonologically matching proto-forms and descendants, a false positive result is unlikely to emerge from this manipulation. The control wordlists were constructed in this way to mirror wordlist length found in the case study.

[1] The original dataset does not have indices. Indices were assigned automatically when the data was extracted from the Austronesian Comparative Dictionary website (Trussel & Blust, 2010): https://www.trussel2.com/ACD/acd-s_a1.htm

In total there were 7 control wordlists, one for every proto-language in the case study (including Proto-Ongan-Austronesian and Proto-Ongan). The wordlists differed only in which subsets of *Lord of the Rings* words they included. Mean homotopy in the control wordlists ranged between 6.29 and 7.91, while average word-length ranged between 4.20 and 4.36. Both measures are comparable to the values found for natural languages in the case study.

4.2 Case Study Methods

4.2.1 Data Preparation

The data in both the Austronesian and Ongan datasets is meant to be interpreted by humans and includes information not necessary and, in fact, detrimental to an automated analysis. As such, the raw data was simplified in a several ways. To start, all morphological boundaries were ignored. Symbols indicating morphological boundaries, such as '-' or '/', were simply removed from the transcription. For example, Nggela [vine-vine] 'obstinate' was rendered as [vinevine]. In practice, this is a requirement for proto-forms to be morphologically comparable to their descendants. Thus, although the Nggela form is listed as derived from Proto-Oceanic *pine 'female' (presumably through reduplication), it cannot be reasonably reconstructed from the mother through sound change alone, as such an analysis would require a counter-productive amount of epenthesis. Without an explicit mechanism for describing morphological change, cases where mother-daughter pairs differ morphologically cannot be reconstructed. Contrast the Nggela example with the Maori descendant [hine] 'girl', which was successfully reconstructed from Proto-Oceanic *pine through the simple sound change *p > h.

Segments given as optional in the proto-form were removed entirely. Thus, the Proto-Malayo-Polynesian form *pa(n)tar 'shelf', which exhibits evidence of a nasal in Toba Batak [pantar] but not Fijian [vata], was listed as simply *patar in the dataset. The same is true for two Austronesian infixes: *-in-, which marks the perfective aspect, and *-um-, which marks transitivity (Blust, 2013). Since the proto-forms list the prefixes as optional, only the form without the infix was kept. For example, Proto-Oceanic *h<um>una 'to do or go first' is listed as *huna in the case study dataset. In entries containing more than one proto-form, only the first form was kept. For example, Proto-Oceanic *kali ~ keli ‘dig’ is listed as *kali only, since it appears first.

These changes were made to both the Austronesian and the Ongan datasets. Future implementations of this method deal with competing proto-forms, optional segments, and morphological boundaries in a different way. No matter the strategy employed, care should be taken to avoid *ad hoc* data selection, such as removing optional segments inconsistently, or drawing morphological boundaries that are not independently justified in the daughter language.

It should be noted that the raw datasets do not list an etymological equivalent of every form in every attested language, presumably due to lexical replacement. As such, many of the entries remain unfilled. The unfilled entries should not be thought of as phonologically null. As such, it is not the case that the algorithm should accrue forms from zero through epenthesis. Instead, a word-pair with an empty mother entry but a non-empty daughter entry was listed as permanent lexical replacement.[2] In other words, no amount of sound change could derive such word-pairs. These unfilled pairs still affected the $P(D|S)$ calculation, as they contributed to wordlist length t and the number of lexical replacements l, but they did not interact with the machine learning algorithm in any other way.

No further adjustments were made in the case study to the transcription of either the attested languages or the proto-languages. For attested languages, the Austronesian Comparative Dictionary provides forms in a modified native orthography and not a phonetic transcription (Blust, 2013). Most entries in the database look identical to the standard orthographic notation for the respective language. However, a number of phones in all languages have been replaced with a phonetic symbol; these are the IPA [ŋ], [ʔ], and [ə], as well as *ñ*, corresponding to the palatal nasal [ɲ]. The substituted segments are cross-linguistically common phonemes which vary greatly in their orthographic representation. The substitution facilitates cross-language comparison of morphemes while still allowing for the use of language specific dictionaries.

[2] Word-pairs with a non-empty mother entry but an empty daughter entry were not considered in the analysis. The framework evaluates the likelihood that a randomly generated daughter wordlist can substantiate a reconstruction from the mother. Positing proto-forms for entries not attested in the daughter does not affect this value.

For proto-languages, the forms for proto-languages are given in specialized Austronesian orthography that is well established in the field. The exact pronunciation of the proto-phonemes is unknown. For the symbols not widely employed outside of Austronesian studies, the assumed IPA equivalents based on conjectured phonetic descriptions are given in Table 10 (Blust, 2013).

Table 10: Austronesian orthography

Austronesian	*IPA*
C	t͡s
c	t͡ʃ
z	d͡ʒ
D	ɖ
j	ɟ
S	s
s	ʃ
N	ʎ
r	ɾ

The Proto-Ongan-Austronesian wordlist (Blevins, 2007) has been made to align with the existing Austronesian literature and, therefore, suffers from the same issue. Thus, the orthographies of proto-languages in the dataset are consistent with each other. The orthographies of living languages are guaranteed to be a little consistent with each other, due to the cross-linguistic replacement of some phonemes with phonetic symbols. Most importantly, the orthography of living languages is not necessarily consistent with the orthography of proto-languages.

Ideally, the entire dataset would share a single transcription standard, as differences in orthography can obscure the relationship between the mother and the daughter languages. In fact, for previous quantitative methods in historical linguistics, a mismatch in orthography between wordlists is disastrous. For example, a multilateral comparison algorithm has no way of determining if a mismatch is a result of non-cognacy or disparate orthographies. As a result, it is typical for projects which utilize multilateral comparison to convert the input into a standard orthography, such as ASJPcode (Brown et al., 2009). Nevertheless, as will be shown later in this chapter, for the data at hand, multilateral comparison can yield reasonable results without any additional modification.

For WDT, while differences in orthographic standard are not ideal, they also do not spell catastrophe. If the representation of a segment in the mother does not match the representation in the daughter, a single sound change from the mother representation to the daughter representation can align the two orthographies. For example, observe the Amis (Formosan) and

Proto-Austronesian wordlists in (1) (Trussel & Blust, 2010). There is a mismatch between Proto-Austronesian *S and Amis *s*, though both are phonetically [s]. This mismatch forces the algorithm to posit the unconditioned change *S > s. This additional change does not reflect the history of Amis, as the pronunciation of the segment has remained unchanged. However, it allows the algorithm to successfully ‘reconstruct’ the forms in (1).

(1) *Proto-Austronesian to Amis*

gloss	*PAN*	*Amis*
sweep	*aSik	asik
heaven	*kawaS	kawas
scrape	*kiSkiS	kiskis
mosquito	*likeS	likes

Therefore, so long as the simulated annealing algorithm is able to detect systematic mismatches between orthographies and posit appropriate sound changes between the mother and the daughter, the data can be analyzed without conversion to a single orthographic standard. Nevertheless, where type II error is a concern, using a dataset with standardized orthography is preferable even in WDT.

4.2.2 Simulated Annealing

The following section introduces simulated annealing, the machine learning algorithm used in the case study to generate reconstructions stochastically. The section begins with background on simulated annealing in general. Thereafter, the discussion shifts to the implementation of simulated annealing in the case study.

4.2.2.1 Simulated Annealing Background

The machine learning algorithm employed in the current case study is known as simulated annealing. Simulated annealing is a non-deterministic statistical technique, useful for optimizing outputs in situations where locating global minima is intractable, such as in NP-hard problems (Kirkpatrick et al., 1983). The technique is employed in fields as diverse as image processing (Bertsimas & Tstsiklis, 1993) and nuclear reactor optimization (Triki et al., 2005), but also cognate detection in historical linguistics (Rama & List, 2019). The main advantage of simulated annealing is that is performs well in a variety of tasks with little modification, occasionally on par with problem-specific algorithms (Aarts & Laarhoven, 1989), while also being simple to implement in code. The main drawback of simulated annealing is that it is slow and computationally expensive.

At its core, simulated annealing resembles an iterative improvement algorithm, where the desired function is optimized by stochastically moving through configuration space with a bias for decreasing the cost function. For example, in the case of the travelling salesman problem (Lenstra & Kan, 1975), where the task is to find the shortest path through some number of points (e.g. a salesman's itinerary through some number of cities), the algorithm moves through the space of possible paths with a bias toward those that are shortest. A major issue for iterative improvement algorithms are local minima, i.e. non-optimal configurations where every movement in configuration space is penalized in the cost function. For example, in the travelling salesman problem, there may exist some itinerary through all cities such that no single change to the itinerary (as defined in the algorithm) reduces the distance travelled, even though a shorter itinerary is possible. In cases of local minima, a simple iterative improvement algorithm would be effectively 'stuck', as none of the suggested improvement would prompt a change.

Simulated annealing addresses the issue of local minima by introducing the concepts of *temperature* and *cooling schedule*. The temperature at a given point in time corresponds to the willingness of the algorithm to accept sub-optimal steps. The cooling schedule is the function which determines the change in temperature from one step to the next. The term 'annealing' is taken from annealing in physics, where a substance is heated and gradually cooled. In physical annealing, an initial high temperature allows for sufficient atomic rearrangement, while the gradual cooling encourages formation of energy-efficient configurations, i.e. crystals (Rutenbar, 1989). By comparison, in simulated annealing, initial high temperatures allow for the exploration of the entire configuration space, while the gradual cooling encourages convergence to a (hopefully global) minimum.

Simulated annealing can be used in any optimization problem, so long as six requirements are met (Bertsimas & Tsitsiklis, 1993); these are listed in (2).

(2) *Simulated annealing requirements*

(a) a finite set *G*
(b) a real-valued cost function *J* defined on *G*
(c) for each $i \in G$, a set $G(i) \subset G$ - *i*, called the set of neighbors of *i*
(d) for each *i*, a collection of positive coefficients q_{ij}, where $j \in G(i)$, such that $\sum_{j\in\mathrm{G(i)}} qij = 1$
(e) a non-increasing function *T*, called the cooling schedule, where *T*(*t*) is the temperature at time *t*.
(f) an initial state $x(0) \in G$

The finite set *G* in (2a) is the configuration space that the algorithm searches through. The goal of the algorithm is to find the configuration with the global minimum as defined by cost function *J* in (2b). Each configuration *i* has an adjacent neighborhood of configurations *G*(*i*) which is a proper subset of *G*, as in (2c). If there is more than one neighbor for *i*, transition probabilities q_{ij} onto all neighbors *j* must be defined and sum to 1, as defined in (2d). At each point in time *t*, there must be a temperature *T*(*t*) that is equal to or lower than the temperature at the previous point in time *t*-1, as stated in (2e). Finally, as in (2f), the algorithm must start at some initial configuration.

The algorithm begins at the initial state *x*(0). At each step *x*(*t*), there is some configuration *i*. To determine the configuration at the next step *x*(*t*+1), a neighbor *j* of *i* is selected based on the transition probabilities of *i*. The cost of *i* and *j* are compared. Assuming that the goal is to decrease the cost function *J*, if *J*(*j*) is lower than or equal to *J*(*i*), the configuration of *x*(*t*+1) is set to *j*. If *J*(*j*) is higher than *J*(*i*), the configuration of *x*(*t*+1) is set to *j* with probability *T*(*t*). Otherwise, the configuration of *x*(*t*+1) remains *i*.

Stated informally, the algorithm searches the configuration space by moving from configuration to configuration. Movement is random but limited to configurations that are ‘neighbors’, as must be defined independently. If a move is suggested and the neighboring configuration is better or equal to than the current, the move is accepted. If a move is suggested but the neighboring configuration is worse than the current, the move may still be accepted. The chance of accepting a sub-optimal move depends on the current temperature.

The notions of temperature and cooling schedule is what sets simulated annealing apart from a simple iterative improvement algorithm. It is widely agreed upon that, as in physical annealing, cooling in simulated annealing should be monotonic and gradual (Aarts & Laarhoven, 1989).

However, there is no consensus about what cooling schedule is best (Nourani & Andresen, 1998; Atiqullah, 2004), though it has been shown experimentally that the choice of cooling schedule greatly impacts performance (Bertsimas & Tsitsiklis, 1993). A variety of different cooling schedules have been tried in the field: constant, exponential, logarithmic, and linear (Nourani & Andresen, 1998). One of the most common cooling schedules is geometric, where temperature at point $T(t+1) = T(t) * a$, for some value a, between 0 and 1, or more commonly between 0.5 and 0.99 (Triki et al., 2005).

4.2.2.2 Simulated Annealing Implementation

For the purposes of this dissertation, the configuration space is the set of all reconstructions from a mother wordlist to a daughter wordlist. Each member of the configuration space is an exhaustive list of transformations which successfully derives the latter from the former. The simulated annealing algorithm moves through this set in search of the reconstruction with the lowest *P*(*D*|*S*). Therefore, *P*(*D*|*S*) serves as the cost function for the simulated annealing algorithm. Without a quantitative metric such as *P*(*D*|*S*), simulated annealing, or any other machine learning algorithm, is not possible.

For simplicity, this case study considers only sound change and lexical replacement. Furthermore, sound change is limited to the minimal sound change template introduced in Section 3.2: sound changes are exceptionless segment-to-segment mappings that are unconditioned or conditioned either by the preceding or the following environment (not both). The simplifications were made to reduce the amount of computation required by the algorithm. However, both the *P*(*D*|*S*) evaluation metric and simulated annealing are compatible with other diachronic transformations, such as semantic change, morphological change, analogical change, as well as phonological features.

Additionally, the simulated annealing cost function evaluated all sound changes as mergers. In other words, while all types of sound change were considered by the algorithm, all were entered as mergers into the *P*(*D*|*S*) calculation. Recall that Chapter 3 argues that mergers generally incur the highest increase to *P*(*D*|*S*) among sound changes. As such, this decision likely led to an overestimate but not an underestimate of *P*(*D*|*S*), i.e. type II but not type I error. The decision allowed the algorithm to ignore the distinction between mergers, shifts, and links, facilitating a computational implementation at the cost of overestimating *P*(*D*|*S*).

The initial state of the algorithm is one without any sound changes posited. Sometimes, but not always, this corresponds to a *P*(*D*|*S*) of 1. Occasionally, a form in the daughter wordlist matches its analogue in the mother wordlist exactly. Such word-pairs require no sound changes or replacement to be explained and bring *P*(*D*|*S*) down below 1 before any changes are proposed. At every point in the algorithm, i.e. in every reconstruction, every word-pair must either undergo replacement or be perfectly matching after all sound changes have applied. Adjacency in the space of reconstruction is defined by the sound changes posited. Reconstructions differing in one sound change are considered adjacent to each other. At every step, the algorithm moves between adjacent reconstructions, which is the same as either proposing a new sound change or removing an existing one.

Determining which sound changes to propose is a difficult task in and of itself. Even with the restricted sound change structure, the set of conceivable sound changes is intractably large.[3] However, most sound changes are not relevant to the data at hand. As such, this dissertation makes the decision to limit the options explored to those that can derive at least one form in the daughter wordlist from its analogue in the most recent mother wordlist, i.e. the mother wordlist with all other changes applied. To clarify, a sound change that derives word-pairs affected by other sound changes is still considered, but only if it can also independently derive at least one word-pair from the most recent ancestor on its own.

The benefit of limiting the number of proposed sound changes in this way is that the search can focus on immediately beneficial steps and drastically reduce runtime. The downside of this decision is that the entire set of reconstructions is not accessible. It is likely that a situation arises where two sound changes, if proposed at once, would serve to decrease *P*(*D*|*S*), but neither sound change can decrease *P*(*D*|*S*) on its own. For example, observe the Proto-Oceanic and Fijian wordlists in (3). To derive Fijian from Proto-Oceanic, suggesting both *p > v and *R > ∅ reduces *P*(*D*|*S*), as this removes four instances of lexical replacement at the cost of two sound changes, a trade that is always beneficial (see Section 3.2.3).

[3] For conditioned sound changes, given *n* possible segments, there are approximately $2n^3$ possible changes, since the input, output and environment can be one of *n* segments, and the environment is either preceding or following.

(3) *Proto-Oceanic to Fijian*

gloss	*Proto-Oceanic*	*Fijian*
tree sp.	*paRu	vau
stingray	*paRi	vai
fan palm	*piRu	viu
fish sp.	*piRa	via

The simulated annealing algorithm, as employed in this case study, would not be able to find these changes because the reconstruction with neither change is not neighbors with a reconstruction with either change. In other words, there exists no single word-pair such that, after all the other changes have been applied, the daughter can be derived from the mother through either of the two changes but not both. It should be stressed that the issue is not with the simulated annealing algorithm nor with Wordlist Distortion Theory. Simulated annealing is in principle capable of finding global minima by first suffering an increase to its cost function; after all, this is the utility of temperature. Therefore, if the simulated annealing algorithm suggests either *p > v or *R > ∅, despite the fact that each change on its own serves only to increase *P*(*D*|*S*), there is a chance that the lone change will be accepted, and the partner change will be tried at a later stage. The issue is that, due to the definition of configuration neighborhood employed for computational convenience here, neither change will even be suggested.

Defining the reconstruction neighborhood in this way greatly limits the effectiveness of the simulated annealing algorithm. However, without any limitation, an implementation of simulated annealing on the set of reconstructions is not practical. Other ways of reducing the neighborhood size may be tried in the future. As is shown later in this chapter, even with its effectiveness diminished in its way, simulated annealing yields promising results.

The *P*(*D*|*S*) evaluation metric is an indispensable requirement for a simulated annealing implementation, or any optimization algorithm. Without a strict definition of what makes a good reconstruction, candidate reconstructions cannot be compared. The *P*(*D*|*S*) evaluation metric can also be used when defining transition probabilities. In the simulated annealing algorithm, the likelihood of proposing a candidate change was set to be inversely proportional to its rank in the list of all possible candidate changes sorted by estimated decrease to *P*(*D*|*S*). Thus, the algorithm was more likely to consider candidate changes which had a larger decrease on *P*(*D*|*S*).

The simulated annealing algorithm was run for 200 steps. In terms of temperature, a simple geometric cooling schedule was selected. After some trial and error, a factor of .99 was chosen, as this appeared to give promising results. This relatively high factor ensured that the algorithm was adventurous in its reconstruction suggestions. The temperature gradient for a single simulated annealing run can be seen in Figure 25. At the beginning of the search, almost any proposed change was accepted, though changes which were estimated to decrease *P*(*D*|*S*) were still more likely to be proposed. The search became more conservative as time went on. At the one hundredth step, halfway through the search, changes that increased *P*(*D*|*S*) were accepted only about 35% of the time, whereas changes that decreased *P*(*D*|*S*) were (still) accepted 100% of the time. At the final step of the search, a change that increased *P*(*D*|*S*) was accepted only approximately 13% of the time.

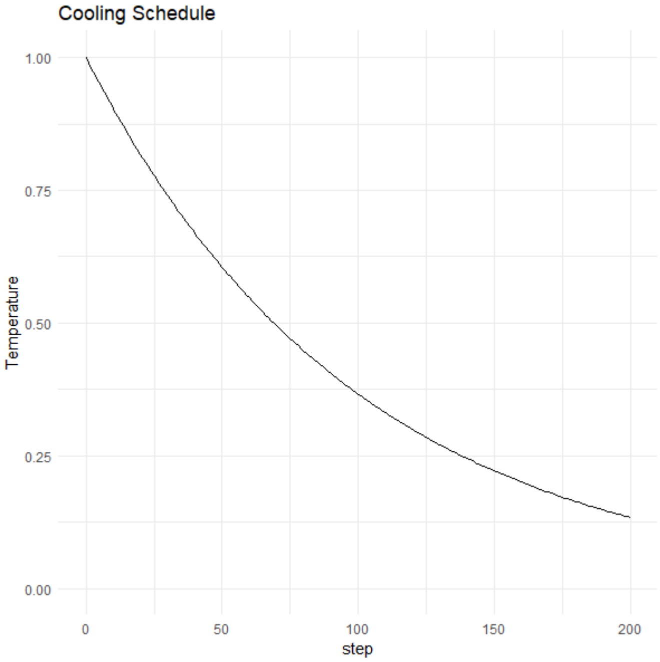


Figure 25: *Cooling schedule*. The simulated annealing cooling schedule employed for the current case study. Temperature, the willingness to accept a subpar change, is on the y-axis. Time, in steps out of 200 total, is on the x-axis.

The simulated annealing algorithm employed in this case study can be summarized in pseudo-code as in (4). Explanatory comments are in red.

(4)

```
T = .99 #variable: temperature, i.e. tolerance for suboptimal changes
changes_current = [] #list: sound changes in the current reconstruction
pds_current = evaluate(changes_current) # variable: $P(D|S)$ of the current reconstruction
changes_best = [] #list: sound changes in the optimal reconstruction
pds_best = pds_current #variable: $P(D|S)$ of the optimal reconstruction

repeat(200): #200 steps of the algorithm
    changes_new = changes + propose_change() #create new reconstruction by adding sound change
    pds_new = evaluate(changes_new) #evaluate new reconstruction

    #accept new reconstruction if $P(D|S)$ decreases or with either way depending on temperature
    if new_pds ≤ current_pds or if random_number() > T:
            changes_current = changes_new
            pds_current = pds_new
    else:
        changes_new = remove_random(current_changes) #create new reconstruction by removing sound change
        pds_new = evaluate(changes_new) #evaluate new reconstruction

        #accept new reconstruction if $P(D|S)$ decreases or with either way depending on temperature
        if new_pds ≤ pds_current or if random_number() > T:
            changes_current = changes_new
            pds_current = pds_new

    #replace optimal reconstruction if new reconstruction is better
    if pds_best < pds_current:
        changes_best = changes_current
        pds_best = pds_current

    T = T * .99 #cooling
```

At the beginning of the run, an initial $P(D|S)$ is computed based on the number of perfect matches between the mother and the daughter wordlists. Temperature is set to .99. At each step thereafter, a new change is proposed. The new change is added to the end of the existing changes. If the reconstruction with the new change does not increase $P(D|S)$, the change is kept.

If the proposed change does increase $P(D|S)$, a random number is generated between 0 and 1. If the random number is lower than the current temperature, the new change is kept anyway.

If the proposed change is rejected, the algorithm selects one of the existing changes at random (if there are any). The selected change is removed from the reconstruction. If the reconstruction without the selected change does not increase $P(D|S)$, the selected change is removed entirely. If it does, a random number is generated; if that number is lower than the temperature, the selected change is removed regardless.

At the end of the step, whether changes were added, removed or neither, cooling occurs: the current temperature is multiplied by .99. The simulated annealing cycle, which is illustrated in Figure 26, repeats thereafter.

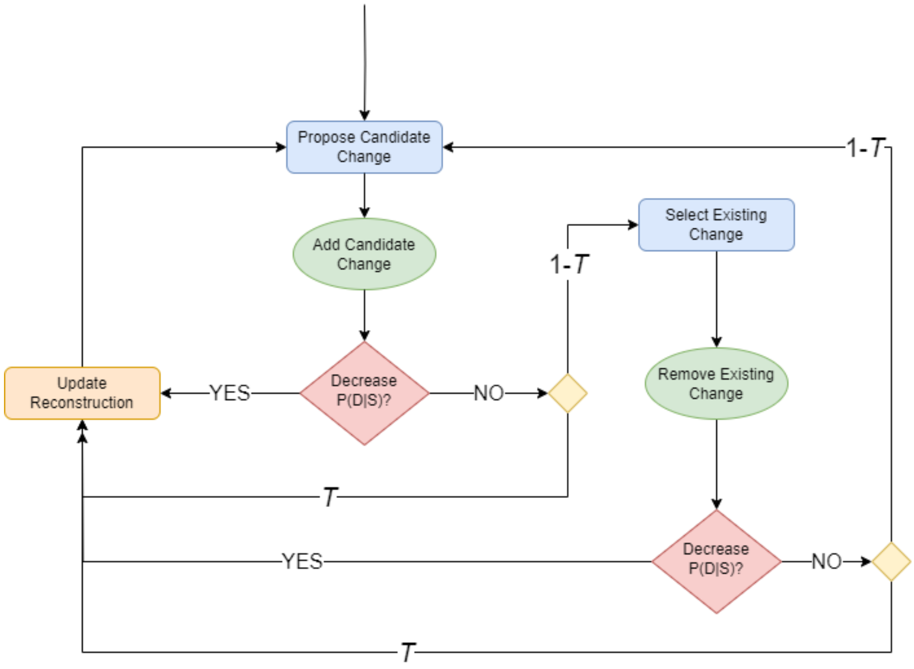


Figure 26: *Simulated annealing loop*. A schematic of the simulated annealing code in (4). *T* is the current temperature in the algorithm.

Effectively, this implementation searches two reconstruction neighborhoods: one for reconstructions that posit one more sound change than the current reconstruction, one for reconstructions that have one fewer sound changes than the current reconstruction. The former neighborhood is explored first; the latter is tried only if no suitable candidate in the former is found. In practice, this prioritizes adding a change to the current reconstruction over removing one. This was deemed to be preferable on intuitive grounds. However, future implementations may choose to combine the two neighborhoods into one.

The reconstruction with the lowest $P(D|S)$ is saved during a simulation run. If at any point, the reconstruction currently explored exhibits a $P(D|S)$ lower than that of the saved reconstruction, the saved reconstruction is simply replaced. Note that the exploration of the reconstruction space happens by altering the current reconstruction and not the one with the lowest $P(D|S)$. In other words, once the reconstruction with the lowest $P(D|S)$ is saved, it has no further effect on the behavior of the algorithm.

To ensure that spurious errors do not affect the analysis, the entire simulated annealing algorithm of 200 steps was run 25 times for every mother-daughter pair. There is no inherent reason for the algorithm to be run 25 times; in fact, it is also possible to run the algorithm only once for each mother-daughter pair. Since simulated annealing is stochastic in nature, each iteration yielded slightly different results. In total, the algorithm searched through 5000 (not necessarily unique) reconstructions for each mother-daughter pair.

4.3 Predictions

Since the concept of $P(D|S)$ is introduced in this dissertation, there is no body of literature to determine its standard of reliability. In the case study, as a significance cut-off, a value of 0.00001 or 10^{-5} was chosen for the reconstruction with the lowest $P(D|S)$. Every combination of proto-language and attested language was compared in the case study, for a total of $7 \cdot 77 = 539$ comparisons (including Ongan languages and English controls). Due to the high number of comparisons, a relatively conservative significance cut-off was chosen. Thus, a connection between the mother language and the daughter language is deemed significant if a reconstruction from the mother wordlist to the daughter wordlist can found such that a randomly generated daughter wordlist merits a reconstruction from the mother wordlist of equal size once in 100,000 attempts or less.

The automated reconstruction with the lowest $P(D|S)$ can be thought of as a higher bound for the ‘true’ $P(D|S)$, i.e. the $P(D|S)$ of the optimal reconstruction. Almost certainly, the reconstructions generated through simulated annealing are not optimal and it is possible to find those that merit a lower *P*(*D*|*S*) value, by positing fewer changes or accounting for a greater proportion of the wordlist. It is impossible to know the *P*(*D*|*S*) for the optimal reconstruction without also knowing the changes required in the optimal reconstruction. However, for the purposes of establishing the higher bound, it makes sense to use the reconstruction with the lowest *P*(*D*|*S*) found rather than the mean.

The *Lord of The Rings* English control wordlists are predicted to exhibit *P*(*D*|*S*) close to or at 1 for simulated reconstructions from all proto-languages tested. The control wordlists are effectively random. While an accidental resemblance or near-resemblance in a few of the word-pairs may produce a reconstruction below 1, reconstructions exhibiting *P*(*D*|*S*) values below the significance cut-off are not expected.

Because *P*(*D*|*S*) quantifies the likelihood of a chance resemblance for a single mother-daughter pair, it is even less clear how to evaluate an entire proto-language with its many descendants, or a group of languages with its many proto-languages. As such, predictions and analyses regarding Austronesian subgrouping reflect numerous *P*(*D*|*S*) values and their total evaluation is by necessity qualitative. The same is true for the Ongan-Austronesian hypothesis.

4.3.1 Austronesian Predictions

All of the languages and proto-languages in the Austronesian Comparative Dictionary are known, or at least widely accepted, to be related. This is true even in the case of a proto-language and a non-descendant. For example, Maranao, a Philippine language spoken on the island of Mindanao, is not a descendant of Proto-Oceanic. Nevertheless, the Maranao wordlist and the Proto-Oceanic wordlist are not independent, as both are ultimately derived from Proto-Austronesian (and, more recently, Proto-Malayo-Polynesian). Therefore, it would not be surprising if a probabilistically non-arbitrary reconstruction from Proto-Oceanic to Maranao can be constructed, even if such a reconstruction does not reflect the history of Maranao.

Nevertheless, it can be assumed that reconstructions from direct ancestors of a daughter language should be easier to construct, i.e. should exhibit lower *P*(*D*|*S*) on average, than reconstructions from proto-languages elsewhere in the family tree. In a reconstruction to a non-descendent daughter wordlist, whatever changes occurred between the latest common ancestor and the proto-language in question must be undone if possible. For example, the changes that occurred between Proto-Malayo-Polynesian and Proto-Oceanic must be applied in reverse when deriving Maranao from Proto-Oceanic before the changes that derive Maranao from Proto-Malayo-Polynesian can be applied. Moreover, only shifts and links are completely reversible; mergers are not. Any merger that occurred between Proto-Malayo-Polynesian and Proto-Oceanic potentially renders word-pairs in Maranao unreconstructible.

Provided that the proto-language wordlists are comparable in size, it should also be the case that reconstructions from recent ancestors of a daughter language should exhibit lower $P(D|S)$ than reconstructions from distant ancestors. Reconstructions from more distant ancestors simply necessitate more changes than reconstructions from more recent ancestors.

These two principles, when applied to the language groups and proto-languages in the case study yield the predictions in Table 11. For the genetic relationship between the language groups and proto-languages, consult Figure 15 at the beginning of the chapter.

Table 11: Case study Austronesian predictions

group	*predicted P(D|S)*
Philippine	p-Philippine < PWMP < PMP < p-Austronesian < p-Oceanic
Oceanic	p-Oceanic < PMP < p-Austronesian < PWMP, p-Philippine
Western	PWMP < PMP < p-Austronesian < p-Philippine, p-Oceanic
Central	PMP < p-Austronesian < PWMP, p-Philippine, p-Oceanic
Formosan	p-Austronesian < PMP, PWMP, p-Philippine, p-Oceanic

For the Philippine languages, wordlists for four direct ancestors are available in the case study. Reconstruction $P(D|S)$ from the four direct ancestors is expected to be correlated with temporal separation: Proto-Philippine eliciting the lowest $P(D|S)$ values, followed by Proto-Western-Malayo-Polynesian, then by Proto-Malayo-Polynesian, and finally by Proto-Austronesian. Reconstruction from Proto-Oceanic (not ancestral) to Philippine languages are expected to yield the highest $P(D|S)$.

Oceanic languages are not descendent from Proto-Western-Malayo-Polynesian. Reconstruction $P(D|S)$ from the three available direct ancestors is expected to also be correlated with temporal separation: Proto-Oceanic eliciting the lowest $P(D|S)$, followed by Proto-Malayo-Polynesian, and then by Proto-Austronesian. Reconstructions from Proto-Philippine and Proto-Western-Malayo-Polynesian (both not ancestral) to Oceanic languages are expected to yield higher $P(D|S)$.

Languages in the so-called Western group are direct descendants of Proto-Western-Malayo-Polynesian. Here too reconstruction *P*(*D*|*S*) from the three available ancestors is expected to be correlated with temporal separation: Proto-Western-Malayo-Polynesian eliciting the lowest *P*(*D*|*S*), followed by Proto-Malayo-Polynesian, and then by Proto-Austronesian. Reconstructions from Proto-Philippine and Proto-Oceanic (both not ancestral) to languages in the Western group are expected to yield higher *P*(*D*|*S*).

Languages in the so-called Central group are direct descendants of Proto-Malayo-Polynesian. *P*(*D*|*S*) of reconstructions from Proto-Malayo-Polynesian is expected to be low, with *P*(*D*|*S*) of reconstruction from Proto-Austronesian higher. Reconstructions from the non-ancestral proto-languages to the languages in the Central group are expected to yield a higher *P*(*D*|*S*) still.

Finally, the Formosan languages are directly descendent from Proto-Austronesian. Reconstruction *P*(*D*|*S*) from this proto-language to the Formosan languages is expected to be low. Reconstructions from the non-ancestral proto-languages to the Formosan languages are expected to yield a high *P*(*D*|*S*).

4.3.2 Ongan-Austronesian Predictions

Predictions regarding the Proto-Ongan-Asutronesian and Proto-Ongan wordlists are not as clear. Unlike Proto-Austronesian, Proto-Ongan-Austronesian is not widely accepted proto-language in the field. Therefore, a lack of probabilistically non-arbitrary reconstruction from Proto-Ongan-Austronesian to any of the Austronesian languages in the dataset could be indicative of either a lack of statistical power in the methodology or a lack of a common ancestor between the two language families. The former becomes more relevant when one considers that the Ongan and Ongan-Austronesian wordlists contain only 83 entries. Recall that the least represented proto-language in the Autronesian dataset is Oceanic, with 1746 entries. Even if the Ongan-Austronesian hypothesis were justified, it is expected that the *P*(*D*|*S*) values for a reconstruction from Proto-Ongan-Austronesian to its descendants to be much higher than for a reconstruction from Proto-Austronesian, both because the former proto-language is further removed in time and exhibits a greater number of diachronic transformations, but also due to the scarcity of data.

The presence of probabilistically non-arbitrary reconstructions from Proto-Ongan-Austronesian to any of the Austronesian languages would indicate that the reconstruction from the hypothetical Proto-Ongan-Austronesian to the given Austronesian wordlists requires fewer transformations than would have been expected through chance. Such a finding would lend credence to the Ongan-Austronesian hypothesis. However, for satisfactory evidence, probabilistically non-arbitrary reconstructions from Proto-Ongan-Austronesian must be found for both of the hypothesized primary branches, Austronesian and Ongan.

4.4 Results

The following sections present results from comparisons between every proto-language wordlist and every attested language wordlist in the dataset, for a total of 539 comparisons. The results of simulated reconstructions to the English controls are presented first, followed by results from the Austronesian dataset, and the Ongan dataset. As in the language background discussion, the Austronesian results are split by language group as defined in the Comparative Austronesian Dictionary (Trussel & Blust, 2010). Results for reconstructions from Proto-Ongan and Proto-Ongan-Austronesian are presented separately. In the results, every mother-daughter comparison is represented by 25 runs of the simulated annealing algorithm.

4.4.1 Control Results

There were 25 simulated annealing runs from each of the 7 proto-languages in the dataset to the *Lord of the Rings* English control wordlists. For each of the proto-languages, Table 12 presents the results from the simulated reconstructions with the lowest *P*(*D*|*S*) along with the corresponding wordlist length *t*.

Table 12: Control results

mother	*P(D|S)*	*t*
Proto-Ongan-Austronesian	.181	83
Proto-Ongan	.181	83
Proto-Austronesian	1	1754
Proto-Malayo-Polynesian	1	2772
Proto-West-Malayo-Polynesian	1	3633
Proto-Philippine	1	1943
Proto-Oceanic	1	1731

For the comparisons between the Austronesian proto-languages and the control, the algorithm was not able to find a single reconstruction with a *P*(*D*|*S*) below 1. For the comparison between Proto-Ongan-Austronesian and the control as well as Proto-Ongan and the control, the algorithm was able to find reconstructions with a *P*(*D*|*S*) of 0.181. Recall that the two mother wordlists were extracted from the original Ongan-Austronesian proposal (Blevins, 2007) and featured the same entries. Therefore, it is not surprising that the algorithm performed similarly for the two. In both reconstructions, the algorithm derived [a], the 420th word in *Lord of the Rings*, from Proto-Austronesian-Ongan *an and Proto-Ongan *an, a locative suffix, through the sound change n > ∅. This single word-pair on its own was enough to yield a *P*(*D*|*S*) below 1, though still far above the threshold of reliability of 0.00001 set in this section. Because for all comparisons in the control dataset, the lowest *P*(*D*|*S*) was far above the threshold of reliability, it can be said that the control wordlists behaved as expected.

4.4.2 Austronesian Results

4.4.2.1 Philippine Group Results

There were a total of 725 simulated annealing runs per proto-language for the 29 languages in the Philippine group. The reconstruction *P*(*D*|*S*) values by mother language are shown in Figure 27. The boxplot conflates results for all Philippine languages. As such, each box represents 725 *P*(*D*|*S*) values.

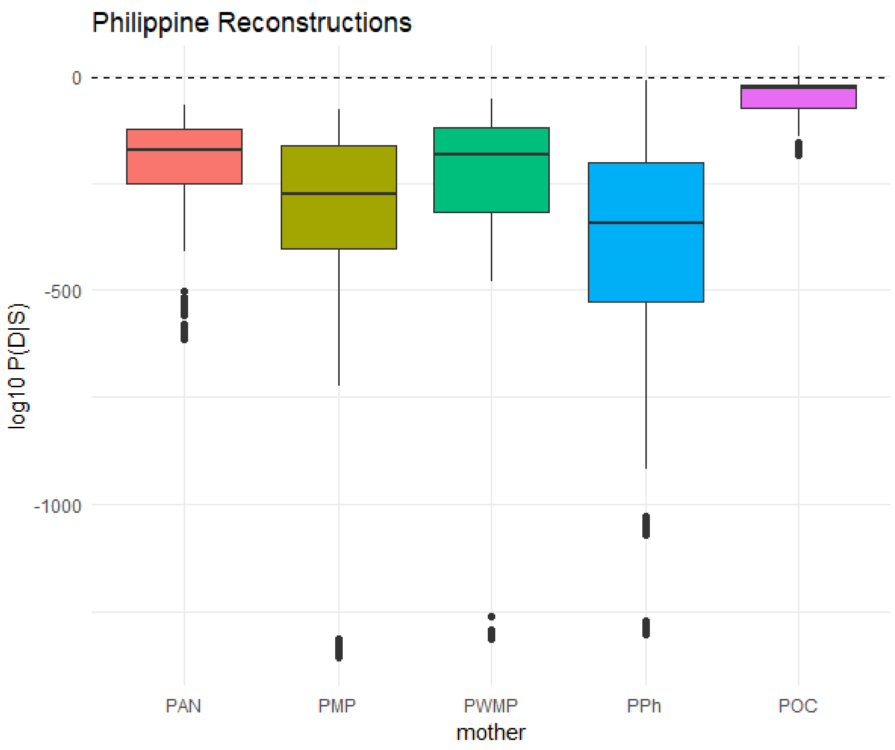


Figure 27: *Philippine group results*. Boxplot with $P(D|S)$ values for reconstructions from 5 Austronesian proto-languages to the 29 languages in the Philippine group. Each box represents $P(D|S)$ values for 25 reconstructions to each Philippine language. The dashed horizontal line corresponds to the chosen cut-off of reliability at .00001 or 10^{-5}. PAN = Proto-Austronesian, PMP = Proto-Malayo-Polynesian, PWMP = Proto-Western-Malayo-Polynesian, PPh = Proto-Philippine, POC = Proto-Oceanic.

Reliable reconstructions with a $P(D|S) < .00001$ were found from all proto-languages known to be ancestral to the Philippine group, i.e. Proto-Austronesian, Proto-Malayo-Polynesian, Proto-West-Malayo-Polynesian, and Proto-Philippine. Reliable reconstructions were also found from Proto-Oceanic to all Philippine languages except for Ayta-Abellan, although this is not a direct ancestor of the group.

It was predicted that $P(D|S)$ in reconstructions to the Philippine group would on average be correlated with proto-language recency. Indeed, reconstructions from Proto-Philippine, the most recent ancestor, elicited the lowest $P(D|S)$ among the Philippine languages. However, although Proto-Western-Malayo-Polynesian is a more recent ancestor than Proto-Malayo-Polynesian, reconstructions from Proto-Western-Malayo-Polynesian on average produced a higher $P(D|S)$ than those from Proto-Malayo-Polynesian. As expected, reconstructions from both Proto-Malayo-Polynesian and Proto-Western-Malayo-Polynesian yielded a $P(D|S)$ lower than reconstructions from Proto-Austronesian, the most distant accepted ancestor. Moreover, as was predicted, reconstructions from Proto-Oceanic, not a direct ancestor of the Philippine languages, produced the highest *P*(*D*|*S*). Therefore, the results are mostly in line with the hypotheses laid out in Section 4.3.1.

The results for the individual daughter languages paint a more nuanced picture. While *P*(*D*|*S*) in reconstructions from Proto-Philippine was the lowest on average, there were a few languages which elicited a high *P*(*D*|*S*) in reconstructions from Proto-Philippine and a comparatively low *P*(*D*|*S*) from the other proto-languages. To illustrate, Figure 28 presents the difference between the *P*(*D*|*S*) minima in reconstructions from Proto-Philippine subtracted from the minima in reconstructions from Proto-Austronesian for each Philippine wordlist in the dataset. The languages with a positive difference, on the left in Figure 28, behave as expected, in that they are more easily derived from Proto-Philippine, the more recent ancestor. The languages with a negative difference, on the right in Figure 28, behave contrary to the predictions, in that they appear to be more easily derived from Proto-Austronesian, the more distant ancestor.

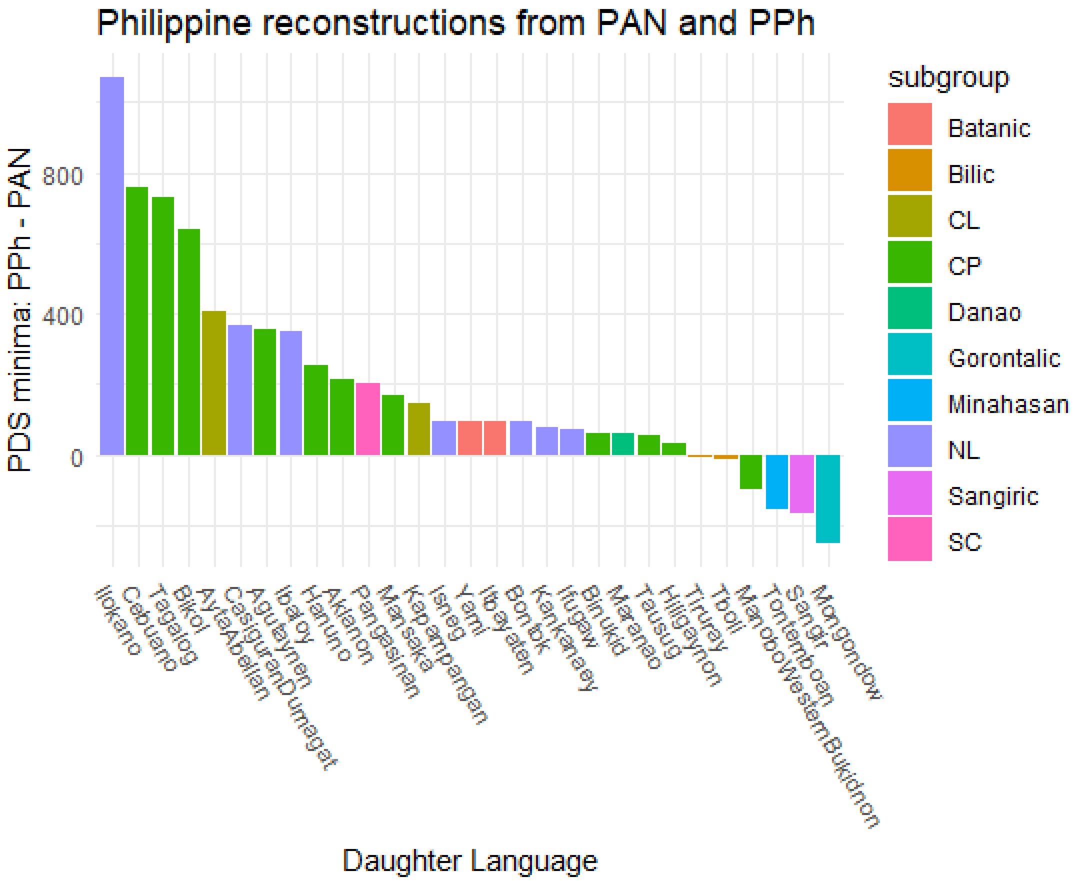


Figure 28: *PPh vs PAN for Philippine*. Bar graph of difference in *P*(*D*|*S*) minima for reconstructions from two proto-languages, Proto-Philippine and Proto-Austronesian, to the languages in the Philippine group. Bars are colored by language branch. CL = Central Luzon, CP = Central Philippine, NL = Northern Luzon, SC = Southern Cordilleran.

There are 6 aberrant languages of this kind. Table 13 presents *P*(*D*|*S*) minima for these 6 languages from all ancestors tested. For all of these languages, reconstructions from Proto-Philippine yielded *P*(*D*|*S*) values far higher than from any other ancestral wordlist, which is contrary to the prediction. In fact, with the exception of Manobo and Tboli, even reconstructions from Proto-Oceanic resulted in lower *P*(*D*|*S*).

Table 13:Log10 of P(D|S) minima for reconstructions to 6 Philippine languages

language	*branch*	*PAN*	*PMP*	*PWMP*	*PPh*
Mongondow	Gorontalic	-332	-596	-413	-80
Sangir	Sangiric	-191	-409	-173	-26
Tontemboan	Minahasan	-209	-326	-141	-53
Manobo	Central Philippine	-410	-724	-469	-311
Tboli	Bilic	-82	-143	-99	-69
Tiruray	Bilic	-126	-286	-313	-119

This pattern is noteworthy particularly because the languages in question are not randomly distributed geographically. As was shown in Figure 18, the Gorontalic, Sangiric, Minahasan branches, represented here by one language each, are spoken at the southernmost edge of the Philippine domain, not in the Philippines but on the Indonesian island of Sulawesi and the Sangir islands immediately north of it. No other Philippine languages in the dataset are spoken in Indonesia. The island of Sulawesi is otherwise home to many non-Philippine Western Malayo-Polynesian languages, such as Wolio, Bare'e, Ta'e, and Makassarese. The degree of contact between the Philippine languages and non-Philippine languages of Sulawesi is unclear. However, most languages of Sulawesi, Philippine or otherwise, exhibit properties not found in closely related languages outside of the island. Most notably, there are several processes of deletion and/or merger in the coda position, rarely found in non-Oceanic Austronesian languages but found in most languages of Sulawesi (Sneddon, 1993; Blust, 2013:83). It is possible that the case study results also indicate contact between the Gorontalic, Sangiric, and Minahasan branches and other local languages. In other words, the contact with non-Philippine Western-Malayo-Polynesian languages may have eroded evidence of Philippine descent while preserving evidence of the shared Austronesian and Malayo-Polynesian descent.

The Bilic languages, which also exhibit an unusually high $P(D|S)$ in reconstructions from Proto-Philippine, are spoken on the south-western coast of Mindanao (Philippines), where contact with Sulawesi is attested (Snodden, 1984). It may be the case that this contact also served to erode the membership of the Bilic languages in the Philippine group, though it should be noted that, in either case, this branch is known to be particularly divergent within the family (Reid, 1982). In summary, even though reliable reconstructions from Proto-Philippine were found for all Philippine languages, the reconstructions from Proto-Philippine were less reliable for the Philippine languages spoken in Indonesia or, to a lesser extent, those in contact with languages spoken in Indonesia.

In contrast to the Gorontalic, Sangiric, Minahasan, and Bilic languages, the Batanic languages of the Batanes islands on the very north of the Philippines, represented in this sample by Yami and Itabayten, do not exhibit any unusual behavior. Because the Batanes islands were likely the first step in the Malayo-Polynesian expansion out of Taiwan, it was previously suggested that these languages are a primary branch of Malayo-Polynesian (Ross, 2005). However, this view is no longer supported in the literature (Blust, 2013:747; Gallego, 2014; Ross, 2020). Instead, it is usually assumed that the Batanic languages are a result of a subsequent northward expansion out of Luzon. The results of the case study confirm this position and do not bring the membership of Batanic into question, as wordlists from both languages were reliably reconstructed from the Proto-Philippine wordlist by the simulated annealing algorithm. For both Batanic languages, reconstructions from Proto-Philippine produced a lower $P(D|S)$ than reconstructions from any of the more distant ancestors.

Because the Malayo-Polynesian expansion out of Taiwan into Melanesia is assumed to have taken place through the Philippines in general, it is sometimes argued that the Philippine languages actually comprise several primary branches of Malayo-Polynesian or a product of an original dialect chain in the region (Reid, 1982; Ross, 2020; Reid, 2020). Additionally, although there are numerous syntactic and lexical correspondences, the only phonological evidence in favor of the subgrouping is the merger of Proto-Malayo-Polynesian *d and *z, which is neither exclusive to the Philippine family nor exhibited by all of its members (Reid, 2020; Liao, 2020). In response, it has been argued that syntactic and lexical evidence in favor of Proto-Philippine is overwhelming and that, in general, the linguistic diversity in the Philippine family is lower than elsewhere in the Malayo-Polynesian family (Blust, 2019; Blust, 2020; Zorc, 2020), usually taken as evidence of a migration into the Philippines, perhaps out of Indonesia (Blust, 2019), after the initial migration out of Taiwan.

The results of the case study do not bring the existence of Proto-Philippine into question. The simulated annealing algorithm was able to find reconstruction from the given Proto-Philippine wordlist to all of its suggested descendants. Although the reconstructions for the Philippine languages in Sulawesi were less reliable in general, even reconstructions to the Sangir, Minahasan, and Gorontalic branches from Proto-Philippine were far below the chosen threshold of reliability.

4.4.2.2 Oceanic Group Results

The simulated annealing algorithm was run 25 times for each mother-daughter pair, resulting in a total of 375 simulated annealing runs per proto-language for the 15 languages in the Oceanic group. The reconstruction $P(D|S)$ values by mother language are shown in Figure 29.

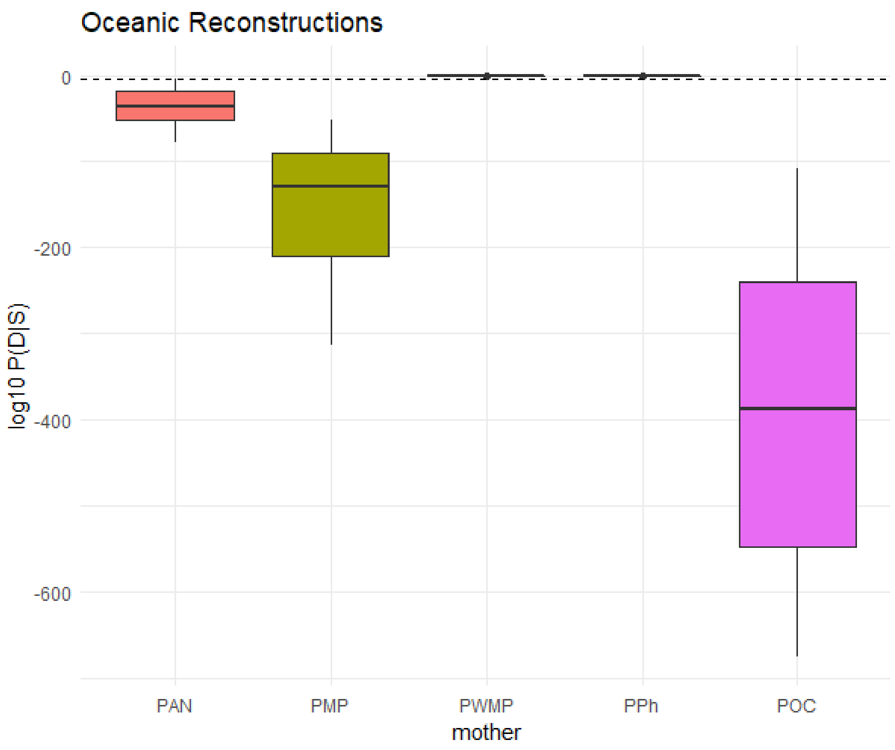


Figure 29: *Oceanic group results*. Boxplot with $P(D|S)$ values for reconstructions from 5 Austronesian proto-languages to the 15 languages in the Oceanic group. Each box represents $P(D|S)$ values for 25 reconstructions to each Oceanic language. The dashed horizontal line corresponds to the chosen cut-off of reliability at .00001 or 10^{-5}. PAN = Proto-Austronesian, PMP = Proto-Malayo-Polynesian, PWMP = Proto-Western-Malayo-Polynesian, PPh = Proto-Philippine, POC = Proto-Oceanic.

Reliable reconstructions with a $P(D|S) < .00001$ were found from all proto-languages known to be ancestral to the Oceanic group, i.e. Proto-Austronesian, Proto-Malayo-Polynesian, and Proto-Oceanic. No reliable reconstruction was found from Proto-Philippine or Proto-West-Malayo-Polynesian (non-ancestral) to the Oceanic languages.

The prediction that $P(D|S)$ in reconstructions to the Oceanic group would on average be correlated with proto-language recency was confirmed in the results. $P(D|S)$ values in reconstructions from Proto-Oceanic were the lowest, followed by reconstructions from Proto-Malayo-Polynesian, followed by reconstructions from Proto-Austronesian. This pattern holds true for individual Oceanic languages as well. For no Oceanic language, were $P(D|S)$ minima in reconstructions from a more recent ancestor higher than $P(D|S)$ in reconstructions from a more distant ancestor. Moreover, as was predicted, reconstructions from Proto-Philippine and Proto-

Western-Malayo-Polynesian, not direct ancestors of Philippine languages, produced the highest $P(D|S)$. Therefore, the results are exactly in line with the hypotheses laid out in Section 4.3.1.

It should be noted that the Oceanic languages exhibit the highest $P(D|S)$ values among the Austronesian languages for reconstructions from Proto-Austronesian, as well as the highest $P(D|S)$ values among the Malayo-Polynesian languages for reconstructions from Proto-Malayo-Polynesian. Compare the average $P(D|S)$ minimum for Oceanic languages in reconstructions from Proto-Austronesian at $10^{-40.8}$ with the average $P(D|S)$ minimum for Central languages, the second highest, at $10^{-95.2}$, a difference of almost 45 orders of magnitude.

The difficulty of establishing a reliable reconstruction to the Oceanic languages from distant ancestors can be explained in two ways. Synchronically, Oceanic languages exhibit relatively low mean homotopy values. It is generally the case for Austronesian languages that phonological inventories decrease with distance from Taiwan. Oceanic languages, inhabiting the easternmost regions of the Austronesian domain, therefore, tend to exhibit the smallest phonological inventories (Lynch et al., 2001; Blust, 2013:171). As discussed in Section 4.1.1.2, Oceanic languages also exhibit lower-than-average mean world-length in the dataset. The low phonological inventory and word-length ensure that Oceanic languages have a relatively small word complexity. As such, a word-pair reconstructed for an Oceanic language on average decreases $P(D|S)$ by a smaller factor than a word-pair in a non-Oceanic Austronesian language. In effect, it can be said that Oceanic languages preserve less information about Proto-Austronesian and other ancestral languages simply by virtue of their phonological properties.

Diachronically, Oceanic languages appear to have undergone a peculiar process of word-initial sporadic nasalization, not found in other Austronesian groups. The computational implementation of the framework, not equipped to tackle sporadic change, treats all changes as perfectly regular. On one hand, incorporating sporadic change into the reconstruction as regular sound change invalidates the entries which do not exhibit the sporadic change, increasing the number of lexical replacements in the reconstruction. On the other hand, excluding a sporadic change invalidates the entries which do exhibit the sporadic change, likewise increasing replacements in the reconstruction. The optimal solution may be to include the change with the phonological conditioning environment that results in the fewest exceptions.[4] No matter the path chosen, there is a high chance that some Oceanic entries widely considered to be Austronesian in origin could be not derived from Proto-Austronesian through the simulated annealing algorithm simply due to the unexpected lack or presence of initial nasalization.

The causes of the sporadic prenasalization of initial obstruents in Proto-Oceanic is apparently unknown (Lynch, et al., 2001). The original Austronesian voice contrast was lost on the way to Proto-Oceanic. However, at some point between proto-East-Malayo-Polynesian and Proto-Oceanic a number of obstruent-initial Oceanic forms became prenasalized. Because the prenasalized obstruents subsequently became voiced, a new voice contrast was introduced into the language. In the literature, these two series of Oceanic forms are known as *nasal grade* and *oral grade*. Both grades are well-represented in the dataset for this case study. Crucially, the nasal/oral grade contrast (or the subsequent voicing contrast derived from it) cannot be predicted based on voicing in Proto-Austronesian (or any other phonological property). Compare the Proto-Austronesian and Proto-Oceanic forms in (5) to confirm that this is the case (Lynch et al., 2001:65). Notice that initial Proto-Austronesian *p corresponds to both Proto-Oceanic *p and *b, while Proto-Austronesian *b also corresponds to both Proto-Oceanic *p and *b.

(5) *Oceanic sporadic initial voicing*

gloss	*Proto-Austronesian*	*Proto-Oceanic*
hot	*panas	*panas
wild pigeon	*punay	*bune
new	*baqeRuh	*paqoru
pig	*beRek	*boRok

Taking the case of initial labials as a proxy for all obstruents, several patterns can be observed in the results. The algorithm appears to have given preference to the oral grade series. Thus, the algorithm successfully derived the oral grade labials but failed to derive the nasal grade in 6 of the 15 Oceanic languages: Fijian, Samoan, Nggela, Renellese, Tongan, Lau, Niue, Saa. The algorithm only partially derived the oral grade labials and failed to derive the nasal grade

[4] As discussed in Section 3.2, it is possible to capture complex patterns, such as sporadic change, with regular sound change, so long as no homophones need to be disambiguated in the process. However, this requires positing multiple regular sound changes, most of which do not serve to decrease $P(D|S)$. As discussed in Section 4.2.2.2, because the simulated annealing algorithm, as implemented here, is blind to certain types of changes, there is a chance that word-pairs requiring more than one change are missed entirely.

altogether in 2 Oceanic languages: Motu and Mota. The algorithm successfully derived the oral grade labials and partially derived the nasal grade in 4 Oceanic languages: Arosi, Maori, Hawaiian, Areare. Finally, only for reconstructions to Tolai did the algorithm successfully derive the nasal grade labials, while also partially deriving the oral grade.

The combination of low word complexity and sporadic voicing seems to have drastically affected the $P(D|S)$ values for reconstructions to Oceanic languages from proto-languages earlier than Proto-Oceanic. Nevertheless, the relationship between the Oceanic group and its two higher level ancestors (Proto-Austronesian and Proto-Malayo-Polynesian) was successfully established by the simulated annealing algorithm.

4.4.2.3 Western Group Results

The simulated annealing algorithm was run 25 times for each mother-daughter pair, resulting in a total of 500 simulated annealing runs per proto-language for the 20 languages in the Western group. The reconstruction $P(D|S)$ values by mother language are shown in Figure 30.

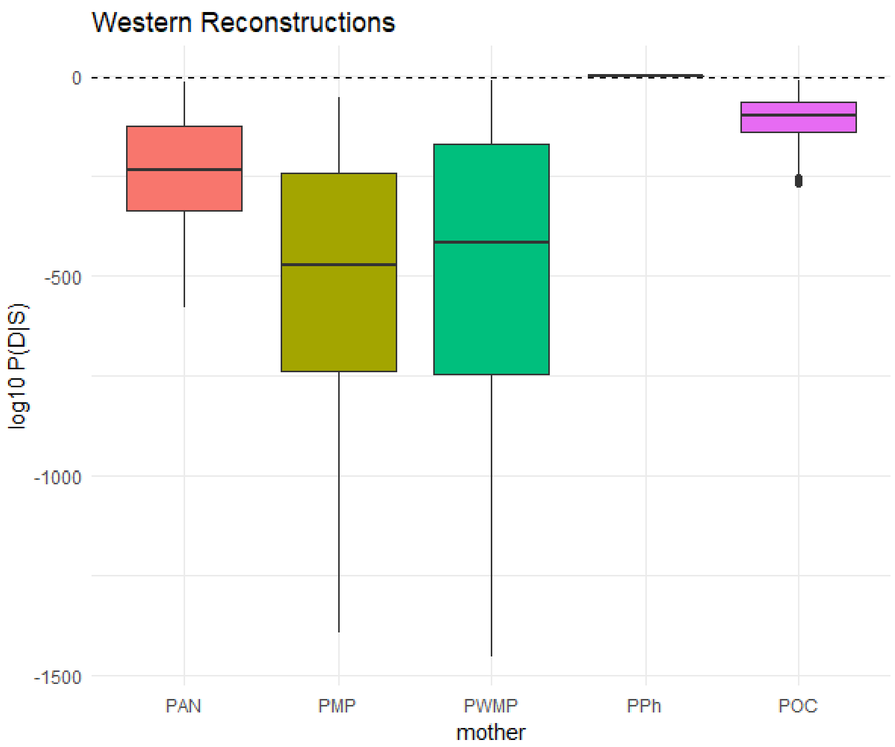


Figure 30: *Western group results*. Boxplot with $P(D|S)$ values for reconstructions from 5 Austronesian proto-languages to the 20 languages in the Western group. Each box represents $P(D|S)$ values for 25 reconstructions to each Western language. The dashed horizontal line corresponds to the chosen cut-off of reliability at .00001 or 10^{-5}. PAN = Proto-Austronesian, PMP = Proto-Malayo-Polynesian, PWMP = Proto-Western-Malayo-Polynesian, PPh = Proto-Philippine, POC = Proto-Oceanic.

Reliable reconstructions with a $P(D|S) < .00001$ were found from all proto-languages known to be ancestral to the Western group, i.e. Proto-Austronesian, Proto-Malayo-Polynesian, and Proto-West-Malayo-Polynesian. Reliable reconstructions were also found from Proto-Oceanic to all Western languages, although this is not a direct ancestor of the group. No reliable reconstructions were found from Proto-Philippine (non-ancestral) to any of the Western languages.

The prediction that $P(D|S)$ in reconstructions to the Western group would on average be correlated with proto-language recency was partially confirmed in the results. The $P(D|S)$ values in reconstructions from Proto-Austronesian were lower than the $P(D|S)$ in reconstructions from Proto-Malayo-Polynesian, both on average and for each individual Western language. However, as was the case for the Philippine group, reconstructions from Proto-Western-Malayo-Polynesian often resulted in a higher $P(D|S)$ than reconstructions from Proto-Malayo-Polynesian, which is the reverse of the predictions. In accordance with expectation, reconstructions from proto-languages not directly ancestral to the Western group, Proto-Oceanic and Proto-Philippine, merited the highest $P(D|S)$. Therefore, the results are mostly in line with the hypotheses laid out in Section 4.3.1.

As was the case for the Philippine group, some Western languages exhibited unusual behavior in reconstructions from different proto-languages. Figure 31 presents the difference between the $P(D|S)$ minima in reconstructions from Proto-Western-Malayo-Polynesian subtracted from the minima in reconstructions from Proto-Austronesian for each Western wordlist in the dataset. The languages with a positive difference, on the left in Figure 31, behave as expected, in that they are more easily derived from the more recent ancestor. The languages with a negative difference, on the right in Figure 31, behave contrary to the predictions, in that they appear to be more easily derived from the more distant ancestor. No aberrations were found in the comparison of Proto-Malayo-Polynesian and Proto-Austronesian with respect to the Western group.

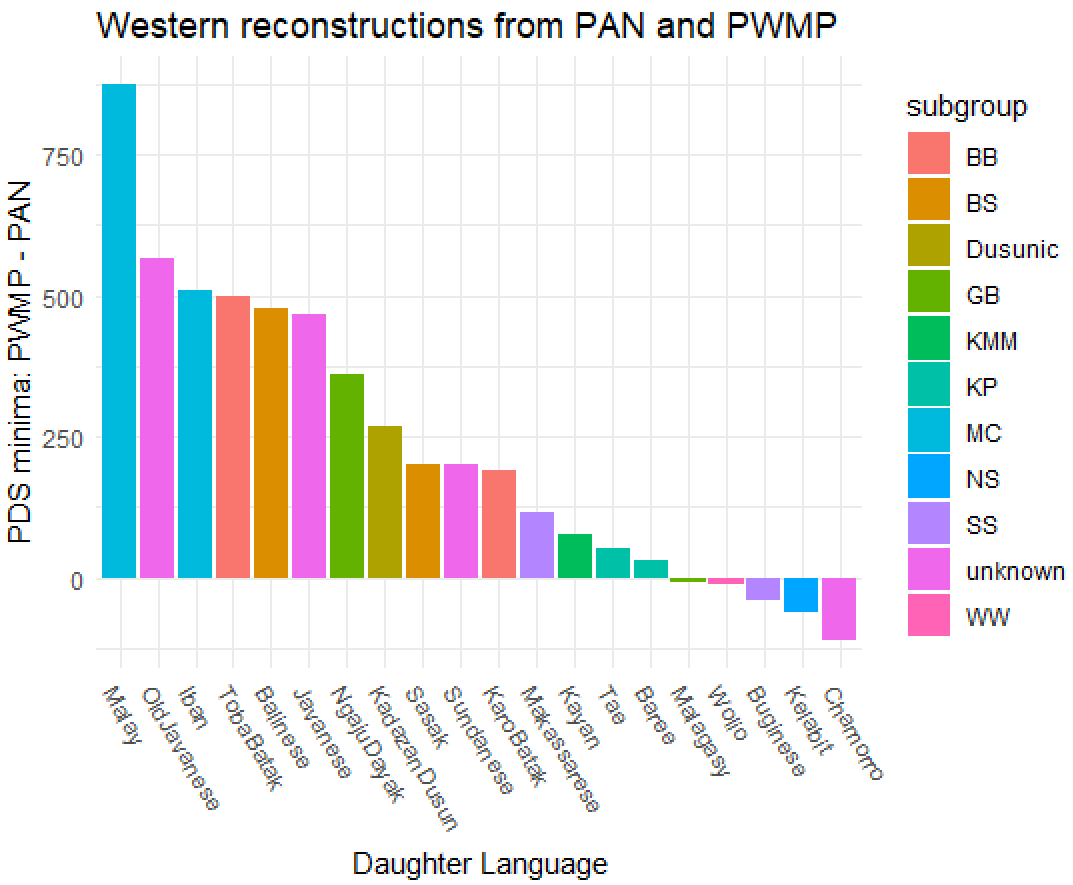


Figure 31: *PWMP vs PAN for Western*. Bar graph of difference in $P(D|S)$ minima for reconstructions from two proto-languages, Proto-Western-Malayo-Polynesian and Proto-Austronesian, to the languages in the Western group. Bars are colored by language branch. BB = Barrier Island-Batak, BS = Bali-Sasak, GB = Greater Barito, KMM = Kayan-Murik-Mudang, KP = Kaili-Pomona, MC = Malayo-Chamic, NS = North Sarawak, SS = Southern Sulawesi, WW = Wotu-Wolio.

Chamorro displays the largest negative difference in $P(D|S)$ minima between reconstructions from Proto-Austronesian (10^{-135}) and Proto-Western-Malayo-Polynesian (10^{-24}). In other words, the reconstruction to Chamorro from the most distant uncontroversial ancestor, Proto-Austronesian, was far more reliable (111 orders of magnitude) than the reconstruction from the alleged recent ancestor, Proto-Western-Malayo-Polynesian. In the case of Chamorro, this pattern can be easily explained. In the long-standing debate surrounding the language, Chamorro is classified in a number of different ways: as a sister to Malayo-Polynesian (Starosta, 1995), as an immediate daughter of it (Reid, 2002; Blust, 2000:104), as a daughter of Proto-West-Malayo-Polynesian (Blust, 2013:31), or even as a Philippine language (Topping, 1980). The results of the case study seem to point to Chamorro as decidedly Austronesian and Malayo-Polynesian but perhaps not Western-Malayo-Polynesian, and certainly not Philippine.

Recall that West-Malayo-Polynesian as a subgrouping is the subject of some debate. An alternative hypothesis is that the grouping is paraphyletic and that the primary branches of West-Malayo-Polynesian are actually primary branches of Malayo-Polynesian itself (Ross, 2005: Klamer, 2019). In fact, the evidence in favor of a West-Malayo-Polynesian family is scant and relies solely on a shared process of morphological nasal substitution, as illustrated in (6) (Blust, 2013).

(6) *Western-Malayo-Polynesian Nasal Substitution in Verbs*

language	*branch*	*gloss*	*base form*	*active form*
Ifugao	Northern Luzon	cover	hukáp	ma-nukáp
Ngaju Dayak	Greater Barito	wrap	buŋkus	ma-muŋkus
Malay	Malayo-Chamic	hit	pukul	mə-mukul
Chamorro	unclear	stay	saga	ma-ñaga

Historically, the nasal coda of Proto-Malayo-Polynesian prefixes *maŋ- ‘active verb’ and *paŋ- ‘agent’ assimilated in place of articulation to the following consonant. The initial consonant of the root was eventually deleted. The nasal was reanalyzed as part of the root, as its place of articulation varied depending on the original (now deleted) initial consonant. The nasal substitution process is extremely common in Western-Malayo-Polynesian languages, both in the ‘Western’ group and the Philippine group (Blust, 2013).

In the absence of a mechanism for morphological change in the framework, it is not surprising that, for both Philippine and Western languages, reconstructions from Proto-West-Malayo-Polynesian do not offer an appreciable increase to $P(D|S)$ over reconstructions from Proto-Malayo-Polynesian. It appears that the nasal substitution process taking place between Proto-Malayo-Polynesian and Proto-West-Malayo-Polynesian had little bearing on the results of the case study.

4.4.2.4 Central Group Results

The simulated annealing algorithm was run 25 times for each mother-daughter pair, resulting in a total of 150 simulated annealing runs per proto-language for the 6 languages in the Central group. The reconstruction $P(D|S)$ values by mother language are shown in Figure 32.

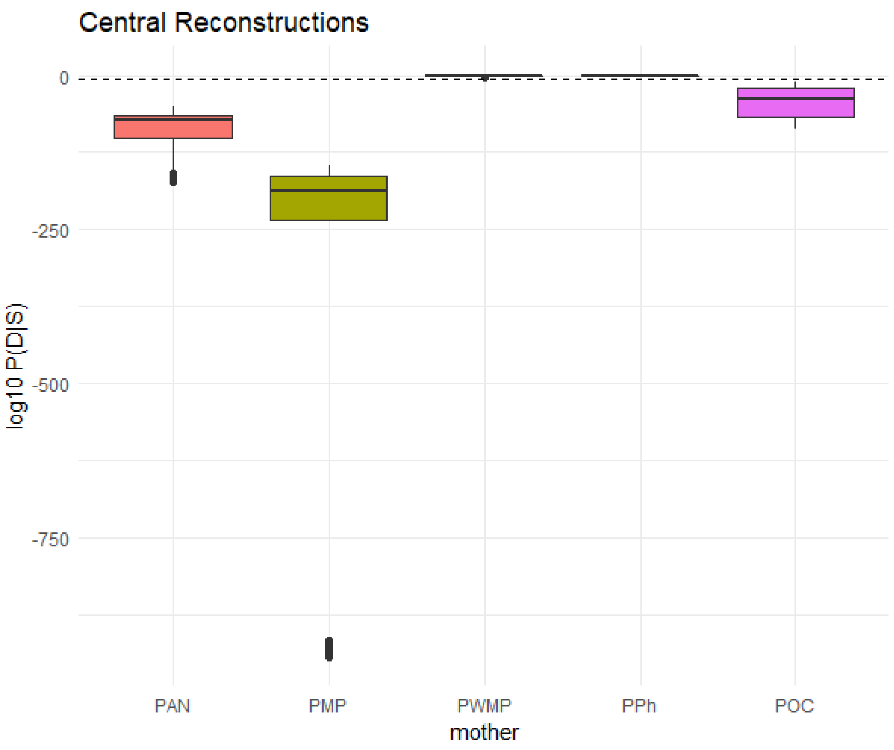


Figure 32: *Central group results*. Boxplot with $P(D|S)$ values for reconstructions from 5 Austronesian proto-languages to the 6 languages in the Central group. Each box represents $P(D|S)$ values for 25 reconstructions to each Central language. The dashed horizontal line corresponds to the chosen cut-off of reliability at .00001 or 10^{-5}. PAN = Proto-Austronesian, PMP = Proto-Malayo-Polynesian, PWMP = Proto-Western-Malayo-Polynesian, PPh = Proto-Philippine, POC = Proto-Oceanic.

Reliable reconstructions with a $P(D|S) < .00001$ were found to the Central languages from the two direct ancestors of the group, Proto-Austronesian and Proto-Malayo-Polynesian. Reliable reconstructions were also found from Proto-Oceanic to all Central languages, although this is not a direct ancestor of the group. No reliable reconstructions were found from Proto-Philippine or Proto-Western-Malayo-Polynesian (both non-ancestral) to any of the Central languages.

As predicted, for the Central languages, $P(D|S)$ values in reconstructions from Proto-Malayo-Polynesian, the more recent common ancestor, were the lowest, followed by reconstructions from Proto-Austronesian, the more distant common ancestor. This pattern was observed both in the group average and also in each individual Central language. Moreover, as was predicted, reconstructions from Proto-Oceanic, Proto-Philippine and Proto-Western-Malayo-Polynesian, not direct ancestors of Central languages, produced the highest $P(D|S)$. Therefore, the results are exactly in line with the hypotheses laid out in Section 4.3.1.

4.4.2.5 Formosan Group Results

The simulated annealing algorithm was run 25 times for each mother-daughter pair, resulting in a total of 100 simulated annealing runs per proto-language for the 4 languages in the Formosan group. The reconstruction *P*(*D*|*S*) values by mother language are shown in Figure 33.

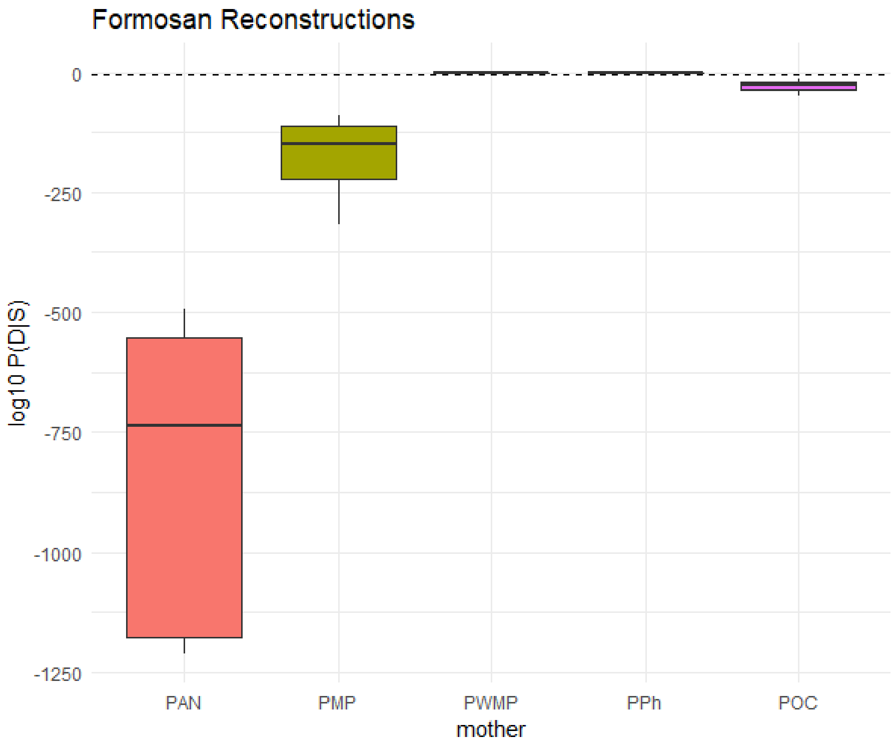


Figure 33: *Formosan group results*. Boxplot with *P*(*D*|*S*) values for reconstructions from 5 Austronesian proto-languages to the 4 languages in the Formosan group. Each box represents *P*(*D*|*S*) values for 25 reconstructions to each Formosan language. The dashed horizontal line corresponds to the chosen cut-off of reliability at .00001 or 10^{-5}. PAN = Proto-Austronesian, PMP = Proto-Malayo-Polynesian, PWMP = Proto-Western-Malayo-Polynesian, PPh = Proto-Philippine, POC = Proto-Oceanic.

Reliable reconstructions with a *P*(*D*|*S*) < .00001 were found to the Formosan languages from Proto-Austronesian, the only direct ancestor. Reliable reconstructions were also found from Proto-Malayo-Polynesian and Proto-Oceanic to all Formosan languages. No reliable reconstructions were found from Proto-Philippine or Proto-Western-Malayo-Polynesian to any of the Formosan languages.

It was predicted that *P*(*D*|*S*) in reconstructions to the Formosan group would be lower in reconstructions from Proto-Austronesian than in reconstructions from non-ancestral proto-languages. This prediction is borne out, and the results are exactly in line with the hypotheses laid out in Section 4.3.1. It should be highlighted that, because Formosan is the only Austronesian group not descendent from Proto-Malayo-Polynesian, this is the only group for which *P*(*D*|*S*) for reconstruction from Proto-Austronesian is predicted to be lower than from Proto-Malayo-Polynesian, as is borne out in the results.

4.4.3 Ongan-Austronesian Results

4.4.3.1 Reconstructions from Proto-Ongan-Austronesian

As was done for the 5 Austronesian proto-languages, the simulated annealing algorithm was used to search for reconstruction to the Austronesian wordlists from the Proto-Ongan-Austronesian and Proto-Ongan wordlists (Blevins, 2007). For these two proto-language wordlists, the predictions are not split by individual Austronesian groups, as all Austronesian languages are hypothesized to be descendent from Proto-Ongan-Austronesian to the same degree.

The *P*(*D*|*S*) for simulated reconstructions from Proto-Ongan-Austronesian to the Austronesian wordlists in the case study is presented in a log plot in Figure 34. No reconstruction with a *P*(*D*|*S*) below the reliability threshold of .00001 was found from Proto-Ongan-Austronesian to any of the Oceanic or Central languages. However, reliable reconstructions were found from Proto-Ongan-Austronesian to all 4 Formosan languages, to 10 of the 29 Philippine languages, and to 12 of the 20 languages in the Western group.

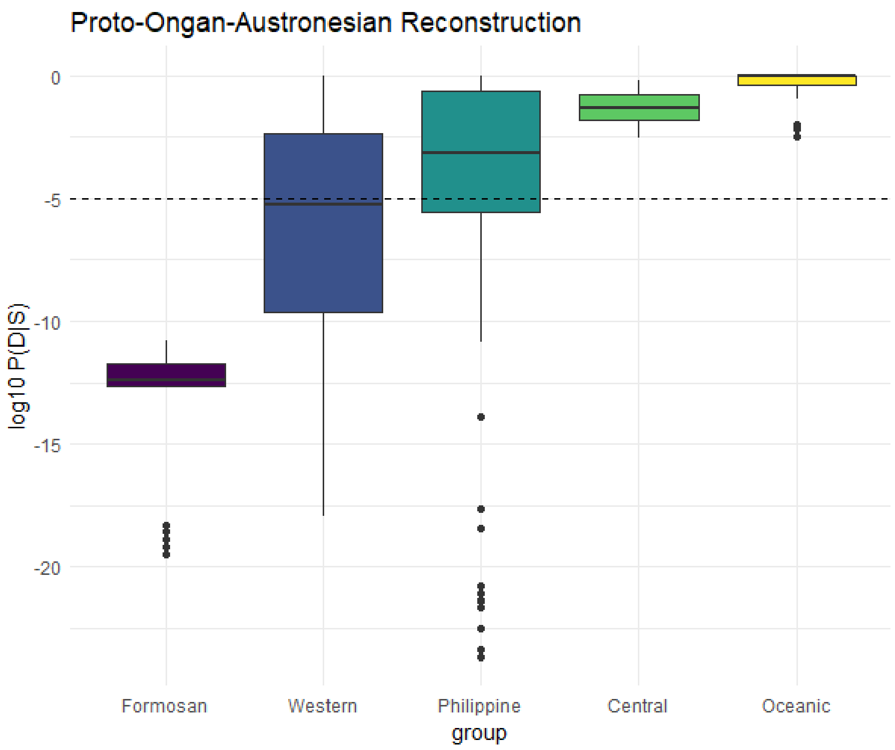


Figure 34: *POA results*. Boxplot of $P(D|S)$ values for reconstructions from the Proto-Ongan-Austronesian wordlist (Blevins, 2007) to all languages in the 5 Austronesian groups. Each box represents $P(D|S)$ values for 25 reconstructions from proto-Ongan-Austronesian to each language in the group. The dashed horizontal line corresponds to the chosen cut-off of reliability at .00001 or 10^{-5}.

As discussed in Section 4.3, there is no body of literature establishing the standard of reliability for a proto-language evidenced by multiple reconstructions. As such, it could be argued that the failure to derive reliable reconstructions from Proto-Ongan-Austronesian to most of its purported descendants is counterevidence to the Ongan-Austronesian hypothesis. Contrast the results in Figure 34 and the ones presented in the previous section for reconstructions from of the proto-languages in the Comparative Austronesian Dictionary. Contrariwise, it could be argued that the presence of even a few reliable reconstructions from Proto-Ongan-Austronesian to attested Austronesian languages is evidence in favor of the Ongan-Austronesian hypothesis. Perhaps, once WDT is tested on a greater collection of genetically and typologically distinct language families, the exact position of Proto-Ongan-Austronesian among these can be ascertained.

For now, it is safe to say that evidence in favor of Ongan-Austronesian in the case study is not conclusive but also not damning. WDT shows that the similarities between the Proto-Ongan-Wordlist and the wordlists of some of its purported Austronesian descendants, as viewed through the lens of diachronic change, cannot be reasonably attributed to chance. For a comparison of Proto-Ongan-Austronesian and a purported descendant, refer to the Maranao data in (7). The $P(D|S)$ minimum of the simulated reconstruction from Proto-Ongan-Austronesian to Maranao is $10^{-23.6}$, the lowest among the Among the Austronesian languages tested.

(7) *Proto-Ongan-Austronesian and Maranao*

gloss	*POA*	*PAN*	*Maranao*
self	*aku	*aku	ako
locative suffix	*an	*an	an
child	*aNak	*aNak	anak
give	*beRay	*beRay	begay
fruit	*buaq	*buaq	boaʔ
bamboo node	*buqu	--	boko
hard exterior	*Cenek	*Cenek	tenek
eat	*kan	*kaen	kan
housefly	*laŋaw	*laŋaw	laŋaw
stative	*ma	*ma	ma
liver	*aCay	*qacay	atay
skin	*aNic	*qaNic	anit
excretion	*pegu	*qapejux	pedo
smoke	*bel	*qebel	bel
rain	*uzaN	*quzaN	oran
celestial light	*cilaq	--	silaʔ

In the reconstruction for (7), 1478 of the 1494 Maranao forms undergo lexical replacement, and the remaining 16 forms undergo 9 sound changes. The Maranao forms that were derived from their Proto-Ongan-Austronesian analogues through lexical replacement are omitted in (7). Proto-Austronesian forms are given for comparison as they appear in the Austronesian Comparative Dictionary (Trussel & Blust, 2010), though note that these were not used in the reconstruction of Maranao from Proto-Ongan-Austronesian.

The 16 Maranao forms are similar enough to their Proto-Ongan-Austronesian analogues that, even with 1478 replacements, the framework gives an overwhelmingly low $P(D|S)$ for the reconstruction. As a reminder, $P(D|S) = 10^{-23.6}$ means that the Maranao wordlist of 1494 forms, if generated at random in accordance with the segmental inventory and phonotactics of the language, would merit a reconstruction requiring fewer replacements and sound changes approximately once out of every $10^{23.6}$ attempts. Therefore, it is not reasonable to attribute the similarities between the Maranao and Proto-Ongan-Austronesian wordlists to chance resemblance.

The interpretation of the $P(D|S)$ values for reconstructions from Proto-Ongan-Austronesian to the Austronesian languages is further complicated by two factors. The first complicating factor is that the Proto-Ongan-Austronesian and Proto-Ongan wordlists were taken from the original proposal at face value (Blevins, 2007). Every proto-form in the proposal was used in the case study as is, provided that the Austronesian analogue featured in the dataset. However, this decision largely ignores the existing criticism of the proposal in the literature (Blust, 2014).

A portion of the criticism levied at Ongan-Austronesian is extra-linguistic and cannot be evaluated using Wordlist Distortion Theory. For example, linguistic anthropology demonstrates that speakers of Proto-Austronesian were sedentary agriculturalists, whereas all of the native populations on the Andaman Islands are hunter-gatherers, making the structure of a Proto-Ongan-Austronesian society unclear. Additionally, the physical characteristics of the Andaman peoples are rather unique: extremely low stature, jet black skin, and occasional steatopygy among females. Some of these characteristics have been observed among speakers of Austronesian, particularly in the Philippines, where people with these qualities are referred to as ‘Negritos’. However, it has long been argued that these Philippine populations are remnants of pre-Austronesian inhabitants of the region rather than descendants of speakers of Proto-Austronesian. As such, the speakers of Ongan and Austronesian are not considered to be genetically similar. Finally, while the western edge of the Austronesian domain is almost adjacent to the Andaman Islands (the South Andaman Island is approximately 650km north-west of the northern tip of Indonesian Sumatra), more than 3000km separate the Andaman Islands and Taiwan, the widely accepted homeland of Austronesian. Therefore, it is difficult to imagine a migration scenario into both regions from a common homeland (Blust, 2014).

The linguistic criticism of Ongan-Austronesian is varied. The morphological boundaries in both the proto-languages and the descendants are argued to be arbitrary, the proto-forms cannot always derive descendants through the suggested sound changes, and some of what is known about lower-level reconstruction is ignored. Furthermore, many of the reconstructions require a radical change in semantics, though recall that this may not be a large issue according to WDT (see Section 2.3).

As an example, Proto-Ongan-Austronesian *bel ‘smoke’ is suggested as the ancestor of both Proto-Ongan *bel and Proto-Austronesian *qe/bel (note that ‘/’ represents a morpheme boundary) (Blevins, 2007:169), cf. Maranao *bel* in (7). However, no reflex of the form appears in Jarawa, and the Onge form *bel/eme* must posit an insufficiently supported morpheme boundary. The morpheme boundary in the Proto-Austronesian form *qe/bel is also not adequately justified (Blust, 2014:324). Issues of this sort are brought up for 25 of the 101 proto-forms in the Onga-Austronesian proposal (Blust, 2014).

Some of these issues are automatically resolved in the current case study. Recall, for example, that the learning algorithm ignored morpheme boundaries altogether. As such, faulty morphology, if it is indeed found in the proto-forms, cannot result in a false positive. Additionally, any attested form not derivable from its proto-form analogue through regular sound change was listed as lexical replacement. For instance, while Maranao *bel* 'smoke' was successfully derived from Proto-Ongan-Austronesian *bel, the algorithm was unable to derive Jarawa *beleme* 'smoke' and it was listed as a replacement instead.

Nevertheless, it could be argued that proto-forms motivated by misunderstood or misrepresented attested forms should not feature in the analysis at all. As such, there may be cause in the future to remove some of the proto-forms from the Proto-Ongan-Austronesian wordlist and repeat the analysis. In this case, the simulated annealing algorithm can simply be rerun on a modified wordlist, with problematic forms removed. In fact, because the formulae behind the evaluation metric are presented in this dissertation, the $P(D|S)$ can even be adjusted theoretically. However, this dissertation chooses to evaluate the original proposal, rather than one adjusted based on critique in the literature, and this is what is reflected in the results.

The second complicating factor is the fact that Ongan-Austronesian is a hypothesis relating two existing language families. As such, reliable reconstructions to members of one the two families are insufficient evidence of the hypothesis. The hypothesis can only be evaluated with respect to reconstructions from Proto-Ongan-Austronesian to both language families. The question of whether the Ongan languages can be reconstructed from the Proto-Ongan-Austronesian wordlist is left for the next section.

So as to be thorough, reconstructions were also run to the Austronesian languages from the Proto-Ongan wordlist included in the original Ongan-Austronesian hypothesis (Blevins, 2007). Proto-Ongan is not a direct ancestor to the Austronesian languages, even if the Ongan-Austronesian connection is correct. However, just as was the case within the Austronesian family, it could be the case that probabilistically non-arbitrary reconstructions could be found from a proto-language and a related non-descendant (see Section 4.3.1).

The $P(D|S)$ for simulated reconstructions from Proto-Ongan to the Austronesian wordlists in the case study is presented in a log plot in Figure 35. As with Proto-Ongan-Austronesian, no reconstruction with a $P(D|S)$ below the reliability threshold of .00001 was found from Proto-Ongan to the Oceanic or Central languages. Reliable reconstructions were found from Proto-Ongan-Austronesian to 2 of the 4 Formosan languages, to 3 of the 29 Philippine languages, and to 3 of the 20 languages in the Western group.

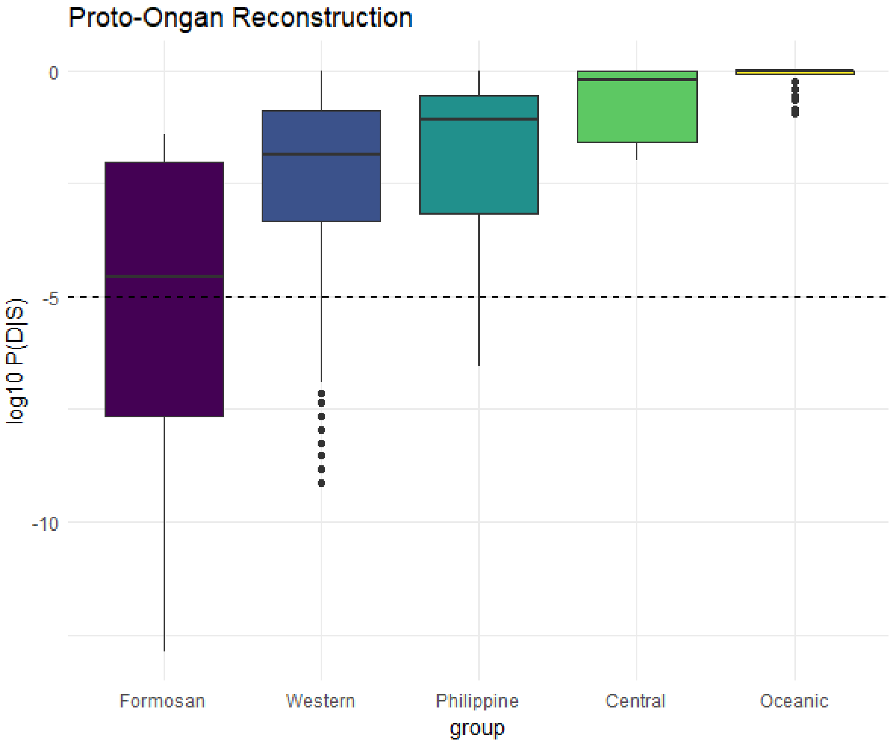


Figure 35: *PON results*. boxplot of $P(D|S)$ values for reconstructions from the Proto-Ongan wordlist (Blevins, 2007) to all languages in the 5 Austronesian groups. Each box represents $P(D|S)$ values for 25 reconstructions from proto-Ongan to each language in the group. The horizontal line within the boxes corresponds to the median. The dashed horizontal line corresponds to the chosen cut-off of reliability at .00001 or 10^{-5}.

As would be expected if the Ongan-Austronesian hypothesis was correct, $P(D|S)$ values in reconstructions from Proto-Ongan to the Austronesian languages are higher than values in reconstructions from Proto-Ongan-Austronesian. The 8 Austronesian languages with reliable reconstructions from Proto-Ongan also merited reliable reconstructions from Proto-Ongan-Austronesian.

In general, the automated reconstructions from Proto-Ongan-Austronesian to the Austronesian languages are compatible with the Ongan-Austronesian hypothesis. Though, since reliable reconstructions were found for less than half of the Austronesian languages in the dataset, the results are likely not strong evidence in favor of the proposal. To properly assess the Ongan-Austronesian hypothesis, reconstructions must be evaluated from Proto-Ongan-Austronesian to the other purported primary branch of Ongan-Austronesian, the two Ongan languages.

4.4.3.2 Reconstruction to Ongan Languages

Automatic reconstructions were also generated from all of the proto-languages to the two Ongan wordlists from the original Ongan-Austronesian proposal (Blevins, 2007). No reconstruction from an Austronesian proto-language to either of the Ongan languages was found to be reliable. Additionally, the reliability of reconstructions from Proto-Ongan-Austronesian and Proto-Ongan to the Ongan languages is difficult to evaluate for practical reasons.

Jarawa and Ongan forms were included in the proposal only if they supported the Proto-Ongan-Austronesian reconstruction. In other words, forms which failed to support the hypothesis are left out. Evaluating the reconstruction from Proto-Ongan-Austronesian as is would constitute a form of cherry-picking, as discussed in Section 2.2. In this case study, rather than calculating the $P(D|S)$ of the reconstructions from Proto-Ongan-Austronesian to Jarawa and Onge based on wordlist length t and lexical replacement l, a range of $P(D|S)$ values was calculated where t and l were allowed to vary. Thereafter, for each reconstruction, the highest possible t and l which maintain a $P(D|S)$ below the reliability cut-off of 0.00001 was calculated.

Let n be the number of forms successfully reconstructed by the simulated annealing algorithm using sound change. The total number of words consulted, i.e. wordlist length t, must be equal to $n + l$, the number of words derived by the algorithm as well as the number of words the algorithm failed to derive. Each hypothetical form consulted but not included in the proposal increases both t and l by 1.

To illustrate, the reconstruction from Proto-Ongan-Austronesian to Jarawa returned 16 forms that did not undergo replacement, out of a total of 61 suggested cognates in the proposal, requiring 11 sound changes (3 conditioned). We know that t must be greater than or equal to 61 and l must be $t - 16$. It could be that $t = 61$ and $l = 45$, or $t = 62$ and $l = 46$, or $t = 63$ and $l = 47$, etc.

For comparison, the reconstruction from Proto-Ongan-Austronesian to Onge returned 11 forms that did not undergo lexical replacement, out of a total of 85 suggested cognates in the proposal. These Onge forms required 14 sound changes, 2 of which were conditioned. The reconstruction from Proto-Ongan to Jarawa returned 31 forms that did not undergo lexical replacement, requiring 8 sound changes, 2 of which are conditioned. The reconstruction from Proto-Ongan to Onge returned 33 forms that did not undergo lexical replacement, requiring 15 sound changes, 6 of which were conditioned.

Figure 36 is a log-log plot of $P(D|S)$ and total wordlist length t for all four reconstructions in question. $P(D|S)$ for the reconstructions was calculated in accordance with the relationship discussed thus far, where t and l were allowed to vary but not independently so. As can be seen in the figure, all four reconstructions start well below the threshold of reliability. In other words, if the Jarawa and Onge wordlists in the original proposal represent the entire corpus of the languages consulted for the reconstruction, then the similarities between the two wordlists and their purported ancestors cannot be reasonably attributed to chance. If more words were consulted to construct the Jarawa and Onge wordlists, but not included in the proposal, the resulting reconstruction are less convincing. For instance, the 16 Jarawa forms successfully derived from their Proto-Ongan-Austronesian analogues are certainly convincing if they were picked from a list of 61, less so if they were picked from a list of 1,000, and not at all if they were picked from a list of 10,000. In general, increasing the number of words consulted has a roughly logarithmic effect on $P(D|S)$. As rule of thumb, log of $P(D|S)$ increases by about the same amount for every order of magnitude increase to wordlist size t, e.g. going from 100 to 1,000 words and going from 1,000 to 10,000 words.

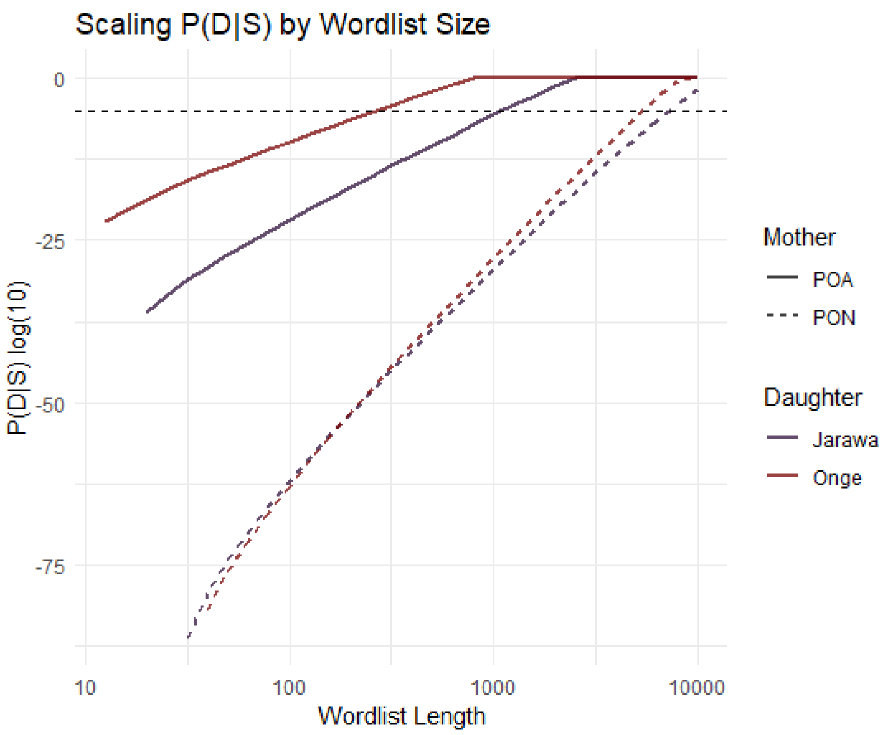


Figure 36: *P(D|S) reliability cut-off for Ongan.* $P(D|S)$ values for reconstructions from Proto-Ongan-Austronesian and Proto-Ongan to Jarawa and Onge by wordlist length t. The dashed horizontal line corresponds to the chosen cut-off of reliability at .00001 or 10^{-5}.

For each reconstruction, it is possible to calculate an upper bound for wordlist length t, such that reconstruction $P(D|S) < .00001$. Table 14 presents the upper bounds for the reconstructions from the two mother languages to the two daughter languages in question. If the number of words consulted during the construction of the daughter wordlists exceeds the corresponding upper-bound in Table 14, the reconstruction cannot have a $P(D|S)$ below the reliability threshold of 0.00001. Even though the upper bounds for reconstructions form the two proto-languages are listed separately, presumably only one Jarawa wordlist and one Onge wordlist were consulted in the original proposal. Therefore, if t is lower than the upper bound for the reconstruction from Proto-Ongan-Austronesian to Jarawa it is also lower than the upper bound for the reconstruction from Proto-Ongan to Jarawa; the same is true for the Onge wordlist.

Table 14: Wordlist length cut-off for Ongan languages

mother	*daughter*	*t upper bound*
POA	Jarawa	1100
POA	Onge	266
PON	Jarawa	7201
PON	Onge	5392

The upper bounds for the reconstructions from Proto-Ongan to both Ongan languages is relatively high. It seems likely that the total number of forms consulted in Jarawa and Onge was lower than 7201 and 5392 respectively. Recall that the wordlist for Malay in this case study, a far better attested language, boasted only 2332 entries. As such, although the true values for t and l in reconstructions from Proto-Ongan to the two Ongan languages are unknown, the $P(D|S)$ for these reconstructions is plausibly below the threshold of reliability at 0.00001.

The upper bound for the reconstruction from Proto-Ongan-Austronesian to Onge at 266 is certainly too low. In fact, just one of the sources for the Onge data listed in the original Ongan-Austronesian proposal (Cogoy, 2004) contains 527 entries. This places the $P(D|S)$ for reconstruction from Proto-Ongan-Austronesian to Onge well over the threshold of reliability set in this dissertation.

The situation for the reconstruction from Proto-Ongan-Austronesian to Jarawa at 1100 is not as clear. I was not able to obtain all of the Jarawa sources listed in the original proposal. None of the sources I was able to obtain contained more than 1100 entries. It is conceivable that the sum of all sources consulted contains more than 1100 Jarawa entries. However, it is also conceivable that this is not the case, especially since forms in different sources are likely to overlap.

In summary, the case study found some evidence in favor of the Ongan-Austronesian hypothesis. There can be no doubt that a sizable genetically diverse subset of Austronesian languages is similar in probabilistically unlikely ways to the Proto-Ongan-Austronesian wordlist from the original proposal. A smaller subset of these can even be derived from the Proto-Ongan wordlist. It appears likely that the two Ongan languages can also be derived from the Proto-Ongan wordlist. However, it is not clear if even one of the Ongan wordlists can be derived from the Proto-Ongan-Austronesian wordlist. Without confirmation from both of the major branches, no strong conclusion can be drawn with respect to Ongan-Austronesian. Stated informally, the Ongan-Austronesian hypothesis is either not reliable or its reliability is hovering just at the edge of what is detectable with the current methodology. It is possible that a future implementation of the current framework – one equipped with more nuanced theoretical machinery or a more powerful learning algorithm or even one with more suitable wordlists – could come to a stronger conclusion.

4.4.4 General Results

4.4.4.1 Review of Predictions

The case study included wordlists from 74 attested Austronesian languages and 5 Austronesian proto-languages, as well as Proto-Ongan and the hypothetical Proto-Ongan-Austronesian. In total, there were 539 mother-daughter comparisons, 370 of which took place within the Austronesian language family. There were 237 comparisons between an Austronesian proto-language and a direct Austronesian descendant; all (100%) of these yielded reconstructions with a $P(D|S)$ well below the reliability threshold of 0.00001. In fact, the highest $P(D|S)$ minimum between an Austronesian proto-language and a known descendant was $10^{-14.9}$, between Proto-Austronesian and the Oceanic (Southeast Solomonic) language Areare.

The 133 comparisons between an attested Austronesian language and an Austronesian proto-language elsewhere in the family tree, yielded 58 reliable reconstructions (43.6%). These should not be thought of as false positives. All wordlists in the Austronesian dataset are descendants of Proto-Austronesian and are, therefore, not independent. For almost half of these indirect comparisons, the algorithm was still able to derive the daughter from the mother in probabilistically non-arbitrary ways.

The primary purpose of the case study is to showcase the utility of Wordlist Distortion Theory in computational approaches to historical linguistics, rather than prove or disprove the existence of an Austronesian language family. As such, methods were kept at a bare minimum. The wordlists, taken from the Austronesian Comparative Dictionary (Trussel & Blust, 2010), were not developed for the purpose of reconstruction; they were not filtered, curated, or converted into phonetic transcription. The $P(D|S)$ evaluation metric was not fitted with a mechanism for dealing with morphological change, analogical change, or sporadic change. The evaluation metric did have a provision for sound change. However, the sound change template was also kept at a minimum; only sound changes that are conditioned by either the preceding or following environment or unconditioned were permitted. Nevertheless, in all instances of a known genetic descent, the simulated annealing algorithm, armed with $P(D|S)$ as a cost function, succeeded in finding a probabilistically reliable reconstruction.

At the extreme, in the case of Proto-Austronesian, the relationship between proto-language and descendant is obscured by approximately 5,500 years of language change (Blust, 2019). It seems likely that future implementations of the current framework can succeed in confirming deeper relationships still, through improvements in data selection and preparation, fine tuning the $P(D|S)$ cost function or employing a different machine learning algorithm. The Ongan-Austronesian connection explored in the previous section may prove to be such a relationship in the future.

Table 15 reviews the predictions from Section 4.3.1 with respect to the relative $P(D|S)$ values in reconstructions from different ancestors. Results not in line with the predictions are bolded. Relative $P(D|S)$ values in reconstructions from Proto-Austronesian, Proto-Malayo-Polynesian, Proto-Philippine, and Proto-Oceanic for all 5 language groups are in line with the predictions. In other words, for these proto-languages, reconstructions from more distant ancestors consistently elicit a lower $P(D|S)$ than reconstructions from more recent ancestors and reconstruction from non-ancestral proto-languages.

Table 15: Case study predictions reviewed

group	*predicted P(D|S)*	*observed P(D|S)*
Philippine	PPh < PWMP < PMP < PAN < POC	PPh < **PMP** < **PWMP** < PAN < POC
Oceanic	POC < PMP < PAN < PWMP, PPh	POC < PMP < PAN < PWMP, PPh
Western	PWMP < PMP < PAN < PPh, POC	**PMP** < **PWMP** < PAN < PPh, POC
Central	PMP < PAN < PWMP, PPh, POC	PMP < PAN < PWMP, PPh, POC
Formosan	PAN < PMP, PWMP, PPh, POC	PAN < PMP, PWMP, PPh, POC

Only the relative $P(D|S)$ values between reconstructions from Proto-West-Malayo-Polynesian and Proto-Malayo-Polynesian fell outside expectation. Where this comparison is relevant, i.e. for the Western and Philippine groups, reconstructions from Proto-West-Malayo-Polynesian yielded higher $P(D|S)$ values than its direct ancestor. This aberrant result is certainly not a result of a lack of data in Proto-West-Malayo-Polynesian, since, at 3681 entries, it is the most attested proto-language in the case study. It is simply the case that descendants of Proto-Western-Malayo-Polynesian could usually be more easily derived from Proto-Malayo-Polynesian.

The results with respect to Proto-Western-Malayo-Polynesian are perhaps not surprising. Due to the rapid expansion of the Austronesian peoples through Indonesia and Melanesia, the time frame between Proto-Malayo-Polynesian and Proto-West-Malayo-Polynesian was likely extremely short. Contrast the well-defined difference in the case study results between Proto-Austronesian and Proto-Malayo-Polynesian, separated by approximately 1500 years, or between Proto-Malayo-Polynesian and Proto-Oceanic, separated by approximately 500 years (Blust, 2019). By comparison, Proto-West-Malayo-Polynesian does not exhibit any phonological innovations from Proto-Malayo-Polynesian. Rather, the proto-language is evidenced solely by lexical innovations and a single morphophonological process of nasal substitution in prefixes (Blust, 2013:31). Many have considered this to be insufficient and skepticism around a single Proto-West-Malayo-Polynesian is common (Ross, 1995; Klamer, 2019). In some ways, the current case study reflects the uncertainty surrounding the Proto-West-Malayo-Polynesian language. Whatever changes did occur on the way to Proto-West-Malayo-Polynesian, the wordlist provided in the Comparative Austronesian Dictionary (Trussel & Blust, 2010) does not lend itself to a reconstruction to the Austronesian languages as easily as its immediate ancestor.

Nevertheless, the results of the case study provide some support for a Western Malayo-Polynesian grouping in general. Recall that the Central group did not merit a single reliable reconstruction from Proto-West-Malayo-Polynesian. There must be some explanation for the difference in behavior between the Western and the Central groups, as both are descendent from Proto-Malayo-Polynesian but not proto-Eastern-Malayo-Polynesian. A single proto-language ancestral to the Western and Philippine languages but not the Central languages, such as Proto-Western-Malayo-Polynesian, is the simplest explanation. The same explanation can be given for the somewhat aberrant behavior of Chamorro, which is easily reconstructible from Proto-Malayo-Polynesian, but not as easily from Proto-West-Malayo-Polynesian. As such, while the relative $P(D|S)$ values in reconstructions from Proto-Malayo-Polynesian and Proto-West-Malayo-Polynesian defied expectation, the results still exhibit some evidence in favor of a genetic grouping between the Western and Philippine languages to the exclusion of the Central, Oceanic, and Formosan languages.

4.4.4.2 Effect of Lexical Replacement

As explored in Chapters 2 and 3, the effect of lexical replacement on $P(D|S)$ is expected to be substantially larger than the effect of other diachronic transformations on theoretical grounds. After all, each replacement increases $P(D|S)$ by roughly a factor of c, the number of phonologically possible words in the language. In comparison, phonological mergers increase $P(D|S)$ by roughly a factor of $\hbar$, the mean number of segments which can occur in a given position in the language. Section 3.2.3 addresses this question more closely.

The results of the case study are in line with the theory presented in Chapter 2. Figure 37 displays $P(D|S)$ values of reconstructions in the case study against the proportion of the wordlist accounted for through regular sound change, i.e. the proportion not undergoing replacement. As can be seen in the figure, lexical replacement and *P*(*D*|*S*) are correlated in reconstructions from all five Austronesian proto-languages. As the number of replacements posited by the reconstruction increases, *P*(*D*|*S*) also increases. Moreover, as discussed in section 2.2.1, the effect of replacement on *P*(*D*|*S*) is mostly a function of the word complexity of the daughter language. The effect of replacement is larger in languages with higher word complexity and smaller in languages with lower word complexity.

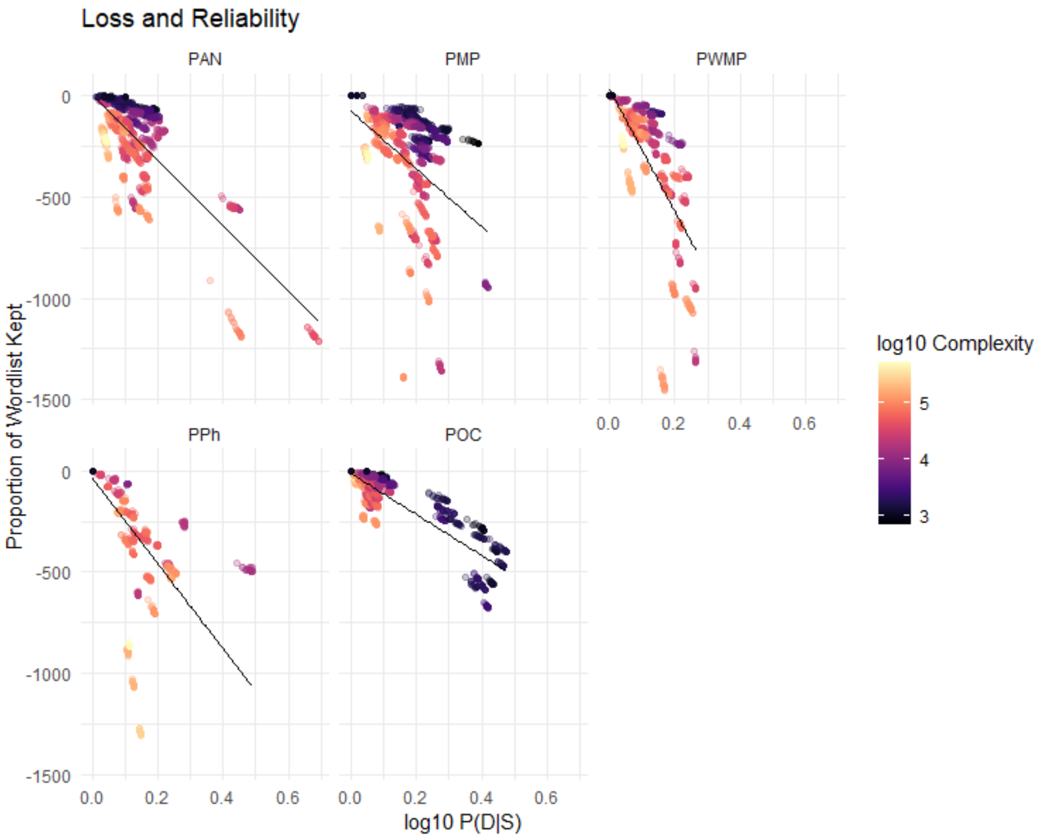


Figure 37: *Replacement and reliability. P*(*D*|*S*) values for reconstructions from the five Austronesian proto-languages to all attested languages in the case study. The proportion of the wordlist not undergoing lexical replacement is on the y-axis. Log base 10 of *P*(*D*|*S*) is on the x axis. Each point corresponds to a reconstruction. Points are colored by word complexity *c* in the daughter language. Results are split by proto-language. A linear line of best fit was added using the stat_smooth() function in *R* (R Core Team, 2019).

A strong negative correlation between lexical replacement and genetic relatedness is the basis of other quantitative frameworks in historical linguistics. Most notably this is the case for lexicostatistics (Swadesh, 1955) as well as Bayesian phylogenetic methods conducted on cognate tables (Greenhill et al., 2017; Robeets et al., 2021; Greenhill et al., 2023). The robust effect of lexical replacement and reconstruction reliability, i.e. *P*(*D*|*S*), observed both in the case study results and predicted by WDT, serves to justify such approaches. In many cases, the number of lexical replacements, as evidenced by the presence or absence of shared cognates, is an adequate proxy for genetic relatedness.

One of the advantages of WDT over previous approaches is the ability to factor in different kinds of diachronic transformations, such as sound change, as well as synchronic factors, such as word complexity and mean number of homotopes. Although lexical replacement plays a large role, other diachronic transformations also affect $P(D|S)$. This effect can be observed in a WDT evaluation but not in a lexicostatistic or phylogenetic Bayesian analysis. Presumably, in applications of WDT that incorporate a more diverse set of diachronic transformations (e.g. semantic change, morphological change), the effect of lexical replacement on $P(D|S)$ would be less pronounced and the difference in results from WDT evaluation and lexicostatistic analysis would increase.

In addition, a WDT evaluation also takes into account synchronic factors of the language analyzed, namely segmental inventory and phonotactics. Figure 37 demonstrates that word complexity in particular plays a large role in determining $P(D|S)$. In other words, because languages exhibiting certain phonological properties (e.g. small segmental inventory, rigid phonotactics) are easier to derive from a random wordlist, they also require more evidence in a reconstruction. Note that, because segmental information is omitted from lexicostatistics analyses in general, there is no sense in which phonological properties of the language are taken into account in these approaches. Nevertheless, it has been observed previously that synchronic properties of the languages in question should play a role in hypothesis testing of genetic relatedness (Ringe, 1999). For this reason, studies employing multilateral comparison often evaluate observed phonetic distance against random permutations of the wordlist in what is known as a Monte Carlo simulation (Kessler & Lehtonen, 2006; Kassian et al., 2015; Ceolin, 2019). The results of this case study give further credence to this practice.

4.4.4.3 Sound Change Naturalness

Recall that reconstructions differing in one sound change were defined as neighbors for the purposes of the simulated annealing algorithm. Adding or removing sound changes was equivalent to searching through possible reconstructions. Therefore, to ensure that the algorithm adequately explored the space of possible reconstructions, one need only ensure that it suggested a reasonable number of sound changes and that these sound changes are in line with our understanding of phonology and the history of Austronesian.

Figure 38 displays the number of changes suggested by the reconstruction for different Austronesian proto-languages and different groups. The mean number of sound changes found per reconstruction in the case study was 21.2. However, this was noticeably higher for reconstructions between an Austronesian proto-language and a direct descendant at 39.2. Overall, the algorithm seems to have successfully explored the space of possible reconstructions.

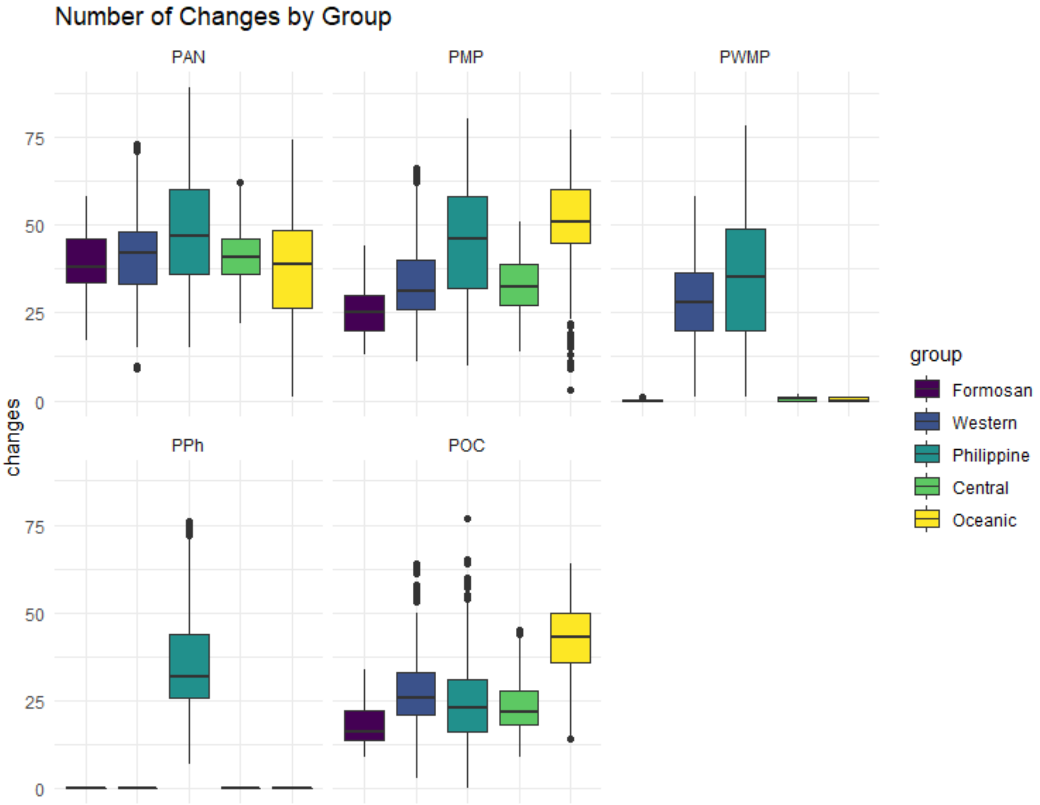


Figure 38: *Sound changes*. Boxplot for the number of sound changes proposed in reconstructions on the y-axis and the Austronesian group on the x-axis, split by proto-language.

Although it was common for the simulated annealing algorithm to suggest more than 50 sound changes in a single reconstruction, most sound changes applied to only a subset of the wordlist. As a result, each entry in the wordlist underwent only a subset of the posited sound changes. For instance, the algorithm required only 5 sound change to transform Proto-Oceanic *raqankayu to Hawaiian *laaʔau* ‘tree, plant, wood’, 6 sound changes to transform Proto-Austronesian *baqeRuh to Aklanon *bagó* ‘before’, and 7 sound changes to transform Proto-Austronesian *busuR to Bikol *buság* ‘bow’. No entry undergoing more than 7 changes was found in the case study.

Figure 39 presents the confusion matrix for the consonant segments across all reconstructions in the case study. The sound change environment is omitted in the chart, as are rare and unusual segments found in individual wordlists. Recall that the Austronesian Comparative Dictionary uses a combination of native orthography, standard Austronesian orthography, and phonetic transcription; refer to Table 10 for details. As such, Figure 39 should be used to identify general trends only. Nevertheless, the figure can be used to confirm some known facts about Austronesian historical phonology.

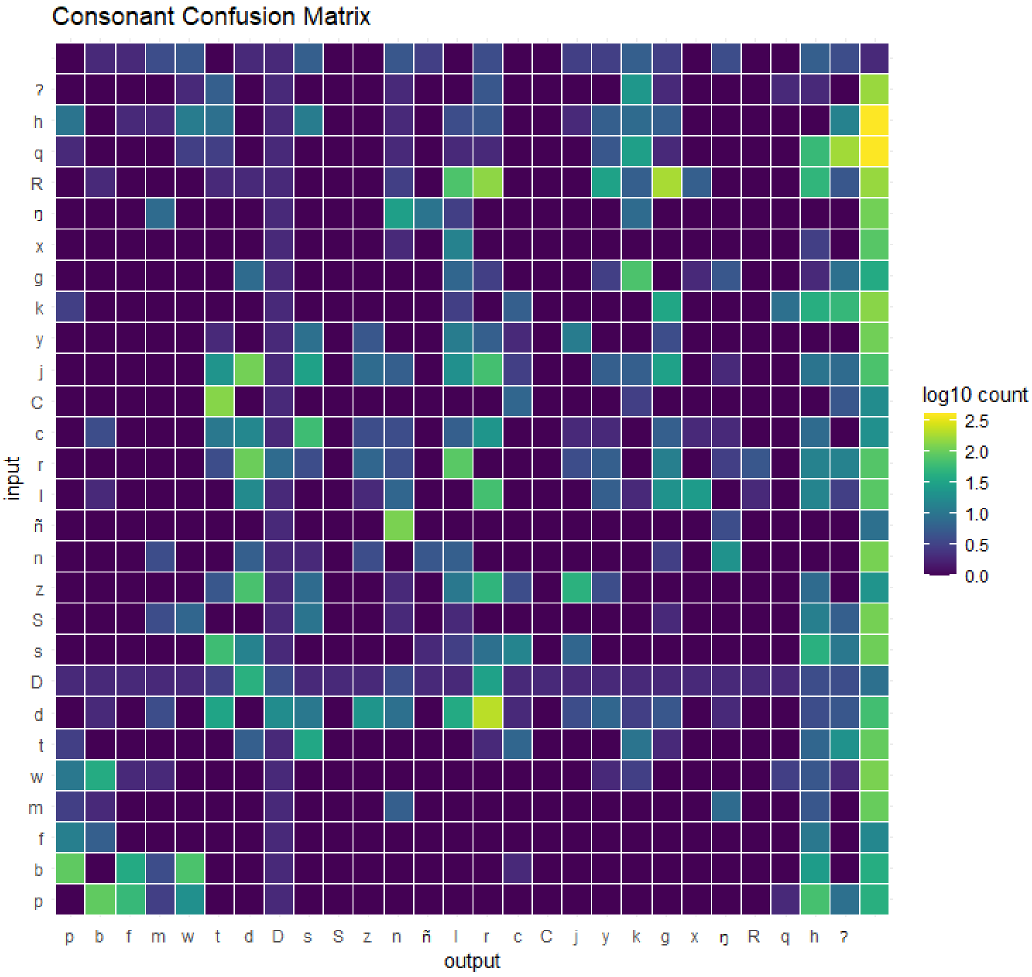


Figure 39: *Consonant confusion matrix*. Sound change inputs are along the y-axis; sound change outputs are along the x-axis. Color represents the number of times the simulated annealing algorithm proposed a mapping in the case study. Sound change environments are not included. The final row and column are null (∅), for epenthesis and deletion.

Although the simulated annealing algorithm was not equipped with any notion of sound change naturalness or phonological features, the suggested changes are cross-linguistically common ones and often directly supported by the Austronesian literature. The 5 most common consonantal changes in reconstructions found by the simulated annealing algorithm are given in (8).

(8) *Common consonantal changes in the case study*

input		*output*	*count*
*h	>	∅	401
*q	>	∅	395
*d	>	r	218
*R	>	g	184
*q	>	ʔ	174

Deletion, as can be seen both in Figure 39, is commonly suggested in the case study for all consonants, but especially for *h and *q. The deletion processes suggested in (8) are in line with the literature on Austronesian. It is known that one of the most common reflexes of the Proto-Austronesian uvular stop *q is cross-linguistically null. As can be seen in (9), other common reflexes of Proto-Austronesian *q are [ʔ], as in (8), as well as [h] and [k]; these can be confirmed in Figure 39.

(9) *Reflexes of Proto-Austronesian *q*

gloss	*PAN*	*Paiwan*	*Amis*	*Tboli*	*Malay*	*Tagalog*
drift	*qañud	qal^{y}udj	ʔalol	konul	hañut	ánod
liver	*qaCay	qatsay	ʔatay	katay	hati	atáy
rain	*quzaN	qudjal^y	ʔorad	kulon	hujan	ulán

The deletion of *h, as shown in (8), is known to be extremely common in Austronesian, particularly as the final stage in a lenition chain (Blust, 2013:612). Compare, for example, Proto-Austronesian *susu 'breast' and Hawaiian [uu] 'breast'. An intermediate stage following *s > h is evidenced in some languages where *h is maintained, such as Kambera [huhu] 'breast'.

The lenition *d > r, as shown in (8), is well-attested in Austronesian (Blust, 2013:617). Proto-Austronesian *d commonly lenited to [r] in intervocalic position. This change occurred independently in several distantly related branches and is evidenced in languages as diverse as Maranao (Philippine), Amis (Formosan), Malagasy (Western), and Ngadha (Central). Other common reflexes of Proto-Austronesian *d are [t, z, s, n, h, l] and ∅, as is confirmed both in the literature (Blust, 2013:617) and in Figure 39 of this case study.

Finally, the *R > g change, as shown in (8), also merits some discussion. The phonetic properties of Proto-Austronesian *R are debated. However, it is usually assumed that the segment was a trill at either the uvular or alveolar place of articulation. The reflexes of the consonant in modern Austronesian languages vary widely, and almost any lingual consonant can be found as a descendant. The change *R > g is evidenced in several Philippine branches as well as Chamorro (Blust, 2013:595).

Figure 40 presents a confusion matrix for the vowel segments. Note that the algorithm did not suggest any instances of vowel to consonant or consonant to vowel changes. Because vowel accents are part of the orthography of many Austronesian languages, particularly those spoken in the Philippines, many of the sound changes involving vowels were concerned with removing or adding the accent.[5] Since vowels in the Proto-Austronesian and Proto-Malayo-Polynesian wordlists are devoid of accent marks, the algorithm was effectively tasked with deriving a system with contrastive stress from a system without contrastive stress through regular sound change.[6] Successful reconstructions were found for most languages and proto-languages, even when vowel accents were involved. As such, it appears that the algorithm was at least partially successful in this task.

[5] In addition to the acute accent, some languages also employ a grave or, more rarely, circumflex accent. For simplicity, all three types of accented vowels have been conflated in the analysis.

[6] Whether Austronesian languages with contrastive stress inherited it from Proto-Austronesian is a matter of some debate (Blust, 2013:563). The Proto-Austronesian and Proto-Malayo-Polynesian forms in the Austronesian Comparative Dictionary are not marked for stress.

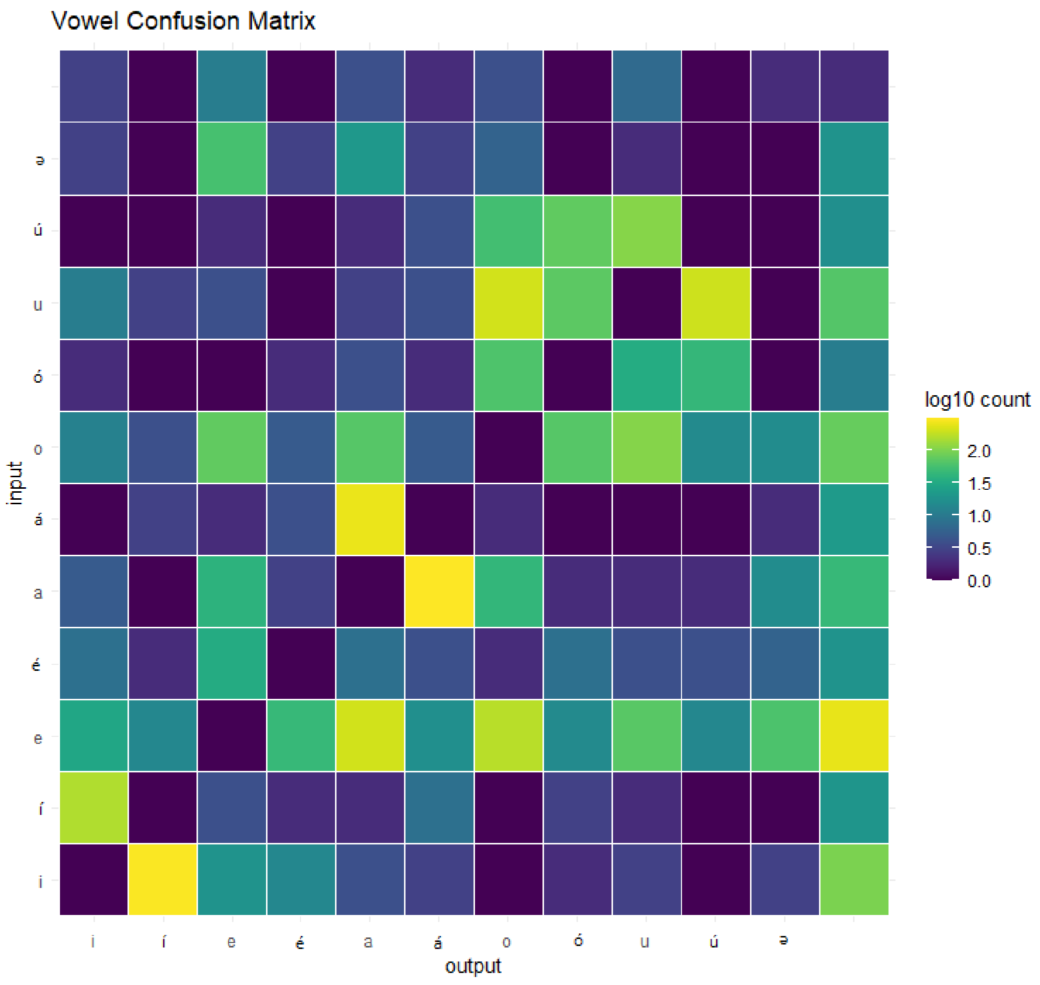


Figure 40: *Vowel confusion matrix*. Sound change inputs are along the y-axis; sound change outputs are along the x-axis. Color represents the number of times the simulated annealing algorithm proposed a mapping in the case study. Sound change environments are not included. The final row and column are null (∅), for epenthesis and deletion.

In general, the vowel segments seem to have been less stable than consonants in the history of Austronesian. Mappings from most vowels to most other vowels can be found in the case study, though the most likely outputs of changes are vowels of the same backness or, in the case of mid vowels, of the same height. Changes between the round back vowels *o*, *u*, *ó*, *ú* appeared particularly frequently. The vowel *e* is the most unstable, commonly undergoing lowering to *a*, backing to *o*, or undergoing deletion. Because in standard Austronesian orthography the symbol *e* can also designate *ə* (Blust, 2013), the volatility of this vowel is not surprising.

All in all, the simulated annealing algorithm adequately explored the space of possible reconstructions by suggesting sound changes. Although the algorithm was not equipped with knowledge of features or common sound change patters, the sound changes proposed appear phonologically natural. Individual sound changes proposed by the algorithm can be identified in the Austronesian literature.

The simulated annealing algorithm was run 25 times for each mother-daughter comparison. Each run consisted of 200 steps. The reconstruction with the $P(D|S)$ minimum was on average identified at step 88. However, this was noticeably higher for reconstructions between an Austronesian proto-language and a direct descendant at 143. On 82 occasions, the reconstruction with the $P(D|S)$ minimum was found on the very last step of the algorithm. Therefore, it seems likely that lower $P(D|S)$ values for some of the reconstructions in the dataset can be found by increasing the simulated annealing runtime.

There was some variance in the performance of the 25 simulated annealing runs for each mother-daughter comparison. The standard deviation of the $P(D|S)$ minima for each comparison in the case study was on average 4.18 log units. The highest standard deviation in reconstruction $P(D|S)$ was 61.1 log units, found for reconstructions from Proto-Austronesian to Paiwan. Due to this sizeable difference in performance, it may be a good idea to execute multiple simulated annealing runs in future implementations as well.

4.5 Multilateral Comparison

The following section presents a multilateral comparison conducted on the same Austronesian dataset. The purpose of the case study is to compare the multilateral comparison results to the results of WDT evaluation of the reconstructions generated through simulated annealing. As discussed previously, no methodology for the quantitative evaluation of comparative reconstructions existed before Wordlist Distortion Theory. Thus, a straightforward comparison between WDT and previous quantitative applications in historical linguistics is not possible. However, some comparison can be drawn between the simulated annealing implementation of WDT and other quantitative measures.

The section will begin by outlining the specific methodology employed (Kessler & Lehtonen, 2006; Ceolin, 2019), as well as any modifications introduced here. The results of the multilateral comparison will follow. Finally, the section concludes with a comparison of two methodologies more broadly.

It should be stressed that the two methodologies in question, multilateral comparison and simulated annealing equipped with WDT evaluation, differ both in their purpose and implementation. As such, any comparison is by necessity informal. Rather than arguing for the use of one methodology of the other, this section is primarily intended to demonstrate that the product of the simulated annealing algorithm is largely in line with estimates given by existing quantitative measures and to illustrate and understand differences in outcomes between the approaches where they arise.

4.5.1 Multilateral Comparison Method

For this section, the phonetic similarity within the Austronesian and Ongan language families was evaluated using the algorithm introduced in Kessler & Lehtonen (2006). In the original paper, the algorithm was used to test the Indo-Uralic hypothesis, demonstrating some statistically significant phonetic similarity within the two families but not between them. A subsequent study (Ceolin, 2019) used the same methodology to test the phonetic similarity between Mongolian, Manchu, and Turkic, three branches of the hypothetical Altaic family. The study found some statistically significant similarity between geographically proximate language groups, even across families, but not between geographically distal ones. As such, it is not clear if the results suggest a genetic relationship or an areal one.

To evaluate phonetic similarity, segments are first grouped into major class by place of articulation. Each major place of articulation is assigned a value based on the approximate position of the tongue in the oral tract during articulation, as in Table 16. Thus, labial segments are assigned a value of 0, dental and alveolar segments a value of 4, postalveolar and palatal segments a value of 6, velar segments a value of 9, and, finally, segments articulated at the uvula and glottis are assigned a value of 10. In this methodology, front vowels are treated as alveolar, while back vowels are treated as velar.

Table 16: Major place of articulation

Major Place	*Position*
Labial	0
Coronal anterior	4
Coronal posterior	6
Velar	9
Uvular and glottal	10

Phonetic distance between two segments is evaluated based on the approximate articulatory distance between them. Thus, labial and alveolar segments are a distance of 4 apart, while alveolar and velar segments are a distance of 5 apart. The distance between two non-identical segments at the same major place of articulation – e.g. [t] vs [s] or [k] vs [u] – is defined to be 0.5.

The intuition behind this phonetic distance metric is that sound change is less likely to affect place of articulation than other phonological features, e.g. voicing or manner. Additionally, sound changes that alter place of articulation typically alter the place of articulation by a minimal amount, e.g. [s] > [ʃ] or [q] > [k]. Less typical are changes that result in a segment originally articulated at the back of the oral tract to be articulated at the front and *vice versa*, though cf. Lat. [okto] 'eight' and Rom. [opt] 'eight'.

Many other metrics for estimating phonetic distance are possible. A common idea is to split the segmental inventory into groups according to both place and manner of articulation and ignore everything except for the first consonant (Dolgopolsky, 1986; Baxter & Ramer, 2000; Kassian et al., 2015). Some studies propose a more nuanced approach based on articulatory features (Kondrak, 2003), though other studies utilize edit distance and ignore segment articulation altogether (Brown et al., 2009; Holman et al., 2011).

Following Ceolin (2019), this section defines phonetic distance between two words to be the average of the distances between each pair of segments. This means that the relative positions of the segments in the words do not matter for this type of calculation. In other words, the distance of a form relative to another is the same as the distance of all of its permutations, e.g. [pin], [nip], [ipn], etc. For a concrete example, the calculation of phonetic distance between the words [pin] and [ki] is illustrated in Figure 41.

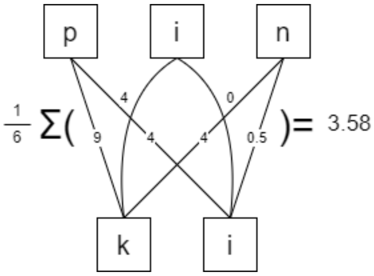


Figure 41: *Phonetic distance example*. The phonetic distance between the strings [pin] and [ki] as calculated using a particular phonetic distance measure (Kessler & Lehtonen, 2006; Ceolin, 2019). The word distance score is equal to the mean distance of all pairs of segments between the two words.

Other metrics of phonetic distance developed for multilateral comparison make use of segment ordering by introducing algorithms for alignment (Kondrak, 2003; List, 2014). In such implementations, the strict ordering of segments in comparanda is preserved. For each word-pair, a specific alignment is chosen, where some of the segments in one word can remain without analogue in the shorter. For example, given the comparison of [pin] and [ki], the alignment that leaves [n] without analogue compares [p] to [k] and [i] to [i] resulting in a relatively low phonetic distance score, cf. the alignment where [p] is without analogue, where [i] is compared to [k] and [n] to [i]. Choosing an alignment that minimizes distance is the primary purpose of alignment algorithms.

While alignment algorithms are commonly employed in the field, there exist measures of phonetic similarity that ignore this aspect of comparison altogether, such as the one employed here (Kessler & Lehtonen, 2006; Ceolin, 2019). With the possible exception of detecting metathesis, one would expect methods employing some sort of alignment to outperform methods that do not, as they can make use of the information preserved in the ordering of the segments. However, as we will see in the following section, for the data at hand, even methods without any sort of alignment can give results in line with the current understanding in the field.

Calculating the distance between two wordlists is as simple as averaging the distances across all word-pairs. Only entries that are found in both wordlists are included in this calculation, since there is no sense in calculating the similarity of a word-pair if only one of the wordlists exhibits a word. In this way, the multilateral comparison methodology differs from any WDT implementation, as the latter has an explicit provision for lexical replacement.

Mean pairwise distance is typically not interpreted on its own. Recall that synchronic factors, such as phonological inventory and phonotactics, can affect the likelihood of substantiating a reconstruction and that this is the motivation behind computing synchronically compatible wordlists in WDT. The same is true for multilateral comparison, where genetically unrelated but typologically similar languages are expected to exhibit a low phonetic distance for any collection of words. As such, it is common in such approaches to simulate randomness by sampling from random permutations of the wordlist, in what is known as a Monte Carlo Simulation (Baxter, 1996; Kessler & Lehtonen, 2006; Kassian et al., 2015; Ceolin, 2019). By calculating the mean pairwise distance of each permutation, one can get a sense of how unlikely the distance score of the original score is. In effect, this is similar to a t-test calculation. For example, if the mean pairwise distance between two wordlists is smaller than 95% of the permutations, then *p* is approximately 0.05.

Previous studies have sampled between 10,000 (Ceolin, 2019) and 100,000 (Kessler & Lehtonen, 2006) permutations per comparison. In general, the accuracy of the *p* value estimate depends on the number of permutations as well as the expected size of the effect, which for the purposes of multilateral comparison corresponds to the phonetic distance. However, as a rule of thumb, it is impossible to estimate *p* values below $\frac{1}{n}$ with fewer than *n* permutations sampled, though accurate estimates require more samples still (Li et al., 2013; Helwig, 2019).

Because accuracy in Monte Carlo simulations is determined by the number of permutation, this methodology is susceptible to a floor effect, where the probability of extremely rare events cannot be estimated probed. In contrast, recall that WDT estimates the likelihood of a randomly generated match theoretically and, therefore, does not suffer from the same issue. For instance, in the Austronesian case study, the probability of a randomly generated Malay wordlist substantiating a reconstruction from Proto-West-Malayo-Polynesian wordlist using the same type and number of diachronic transformations (or fewer) as the attested Malay wordlist is approximately $\frac{1}{10^{1454}}$, far lower than anything that could be simulated using modern hardware.

4.5.2 Multilateral Comparison Implementation

For the current section, the multilateral comparison algorithm (Kessler & Lehtonen, 2006) was tested on every mother-daughter in the combined Austronesian and Ongan datasets. This is the same dataset as was used in the simulated annealing study earlier in the chapter. In total, there were 518 comparisons.

It should be stressed that multilateral comparison algorithms are designed to compare sister wordlists rather than mother-daughter pairs (Greenberg, 1984; Kessler, 2015). In contrast, WDT can only evaluate reconstruction from a mother to a daughter wordlist. As such, for a comparison between the two methodologies to even be possible, it is necessary to use multilateral comparison to evaluate the phonetic distance between an attested wordlist and its reconstructed ancestor, i.e. not as intended. Comparing mother and daughter wordlists rather than sister wordlists can be expected to reduce the amount of diachronic transformations between comparanda by roughly a factor of two, since the number of changes in the independent history of two languages is roughly double the number of changes in the history of each. Therefore, the results of this section are expected to come out as ‘more significant’ than comparisons for similar languages in previous studies, i.e. those that use the same methodology as intended.

The only change to the similarity metric introduced in this section is the treatment of doubly articulated segments, i.e. the treatment of [w], as no other doubly-articulated segments are found in the dataset. In some implementations (Ceolin, 2019), doubly articulated segments are treated as exhibiting both articulations, with preference given to the articulation exhibiting the lowest phonetic distance in the analysis. Thus, for example, the distance between labio-velar [w] and alveolar [d] was 4 (as if [w] were simply labial), whereas the distance between [w] and uvular [ɢ] was 1 (as if [w] were simply velar). In this section, for the sake of simplicity, the position of major place of articulation for doubly articulated segments was simply the average of the positions of the two places in question. In other words, [w] was treated as having a major place of articulation value of 4.5 according to Table 16.

The number of permutations generated for each mother-daughter comparison depended on the length of the wordlist. If the number of possible permutations of the wordlist exceeded 1,000,000, then 1,000,000 permutations were generated at random and evaluated. If the number of possible permutations was less than 1,000,000, then every permutation was generated once.

This number of permutations per comparison is likely superfluous and unnecessary for establishing statistically significant phonetic similarity. However, in order to get the most accurate *p* value estimates from the methodology, the maximum number of permutations was tested given the available hardware.

Finally, recall that the Austronesian dataset combines several different transcription standards. Attested languages are mostly presented in their original orthography, with a few cross-linguistically common segments replaced with their IPA equivalents, whereas proto-languages are transcribed using standard Austronesian transcription. The fact that comparanda are potentially presented in different orthographies is an issue for both WDT approaches (see section 4.2) and for multilateral comparison. However, while a mismatch in orthographic standard can be mitigated in a simulated reconstruction by positing additional sound changes, the same is not true of multilateral comparison approaches. When computing phonetic distance, there is no way of telling if a mismatch is a result of segment dissimilarity or disparate orthography. As such, it is standard practice in multilateral comparison studies to convert all data to the same transcription standard (Brown et al., 2009; Holman et al., 2011).

Nevertheless, in order to keep the results of the multilateral comparison comparable to the simulated annealing case study, the Austronesian dataset was kept as is. For characters that are not part of standard Latin orthography, the major place of articulation was inferred based on a) their approximate pronunciation in Austronesian orthography or b) based on their pronunciation in the most widespread language. As a result, just as in the simulated annealing case study, there are potentially instances where an apparent mismatch in character does not reflect an acoustic or articulatory difference. However, as will be shown in the following section, this did not seem to substantially affect the results.

4.5.3 Multilateral Comparison Results

Just as in previous studies (Kessler & Lehtonen, 2006: Ceolin, 2019), the phonetic distance between two wordlists was deemed statistically significant if a random permutation of one of the wordlists was equally or less distant than the original 5% of the time or less. Of the 518 mother-daughter comparisons, 440 comparisons (84.9%) were statistically significant. In fact, for 336 comparisons (64.9%), no random permutation with a lower or equal phonetic distance was found.

In contrast, recall that the simulated annealing algorithm was able to find a reliable reconstruction for only 333 of the mother-daughter comparisons (64.3%). Compare the number of significant comparisons found through multilateral comparison and through the WDT simulated annealing algorithm, as in Table 17.

Table 17: Number of Significant Comparisons

multilateral comparison	*simulated annealing*	*number of comparisons*
significant	significant	333
significant	not significant	107
not significant	significant	0
not significant	not significant	78

In 333 comparisons, there was a probabilistically reliable reconstruction found through simulated annealing as well as a phonetic similarity beyond any reasonable doubt found through multilateral comparison. For 78 comparisons, neither a reliable reconstruction nor a statistically significant phonetic similarity was found. Thus, for 411 comparisons (79.3%) the two methodologies are in agreement.

There were no comparisons for which a probabilistically reliable reconstruction was found but the wordlists were not found to be phonetically similar. However, there were 107 comparisons (20.7%) for which no probabilistically reliable reconstruction was found but the wordlists were phonetically similar to a statistically significant degree. As such, one can tentatively surmise that the WDT simulated annealing algorithm gives more conservative results than the Kessler & Lehtonen (2006) multilateral comparison algorithm.

Of the 107 comparisons where the two methodologies give conflicting results, 104 were between an Ongan proto-language (Proto-Ongan or Proto-Ongan-Austronesian) and an Austronesian attested language. There were 44 comparisons between Proto-Ongan-Austronesian and an Austronesian wordlist for which no probabilistically reliable reconstruction was found, but the multilateral comparison yielded a significant result. Similarly, there were 60 comparisons between Proto-Ongan and an Austronesian wordlist for which no probabilistically reliable reconstruction was found, but the multilateral comparison yielded a significant result.

The other 3 comparisons for which the two methodologies gave conflicting results were within the Austronesian family. The multilateral comparison methodology found phonetic similarity beyond a reasonable doubt between Proto-Oceanic and Ayta Abellan (Philippine), Proto-West-Malayo-Polynesian and Manggarai (Central), and between Proto-West-Malayo-Polynesian and Paiwan (Formosan). In all three cases, no probabilistically reliable reconstruction was found using simulated annealing.

There was no comparison between a proto-language and a known descendant, for which simulated annealing failed to find a reconstruction, but phonetic similarity was established using multilateral comparison. This is not surprising, since a probabilistically reliable reconstructions were found between all ancestor wordlists and direct descendants. In other words, the comparisons for which multilateral comparison yields a significant phonetic similarity but no reliable reconstruction was found are comparisons where a reliable reconstruction is not actually expected. For example, Ayta Abellan is not a descendant of Proto-Oceanic and it is not necessarily surprising that a probabilistically reliable reconstruction between the two was not found. The same is true for the lack of probabilistically reliable reconstructions from Proto-West-Malayo-Polynesian to Manggarai and Paiwan, as well as from Proto-Ongan to the Austronesian languages.

In contrast to the results from WDT simulated annealing, a multilateral comparison performed on the Ongan and Austronesian datasets arguably gives clear evidence in favor of an Ongan-Austronesian grouping. Phonetic distance was found to be lower than chance in 70 out of 74 comparisons between Proto-Ongan-Austronesian and an Austronesian wordlist. Only 4 Austronesian languages were not found to be sufficiently phonetically similar with respect to Proto-Ongan-Austronesian, those being the three Philippine languages Agutaynen, Ayta Abellan and Tiruray and the Central language Kambera. In contrast, a reliable reconstruction from Proto-Ongan-Austronesian was found to only 26 of the 74 Austronesian languages tested.

Furthermore, all but 6 Austronesian wordlists were found to be phonetically similar to the Proto-Ongan wordlist using multilateral comparison.[7] This is particularly surprising, since, even if the

[7] These are the same 4 wordlists that were not found to be similar to Proto-Ongan-Austronesian (Agutaynen, Ayta Abellan, Tiruray, and Kambera), as well as Wolio (Western) and Areare (Oceanic).

Austronesian-Ognan link is justified, Proto-Ongan is not a direct ancestor of any Austronesian language and is presumably only distantly related.

It should, once again, be stressed that measuring the similarity between attested languages and reconstructed languages is highly atypical in the field, as these are not technically independent. As a result, it is not clear whether the results constitute evidence in favor of an Ongan-Austronesian grouping. In fact, it has been suggested that the use of reconstructed data in multilateral comparison should be avoided altogether to circumvent additional sources of human error (Ceolin, 2019). However, it should be acknowledged that some multilateral comparison studies do compare reconstructed wordlists with other reconstructed wordlists (Kassian et al., 2015).

Because the genetic relationship between Ongan and Austronesian is not widely accepted in the field (Blust, 2013; Blust, 2014), it appears that the results of the WDT simulated annealing algorithm are more closely in line with the current understanding in the field than the results of multilateral comparison performed on the same dataset. In other words, although the Proto-Ongan-Austronesian wordlist is phonetically similar to the Austronesian wordlists, the WDT simulated annealing case study results imply that this similarity has little bearing on the diachrony and does not (easily) evidence a systematic and reliable reconstruction in most comparisons.

Comparing the results of the two analyses more closely reveals that the results of multilateral comparison and simulated annealing are correlated. Figure 42 plots the $P(D|S)$ minimum for reconstructions between a mother and daughter wordlist against the mean phonetic distance. Recall that the multilateral comparison results exhibit a floor effect due to the fact that simulations were limited to 1,000,000 permutations per comparison. In comparison, the $P(D|S)$ values range between 1 and $\frac{1}{10^{1454}}$. As such, although $P(D|S)$ values are better discussed on the logarithmic scale, due to the floor effect, it makes more sense to compare them to multilateral comparison p values on the linear scale.

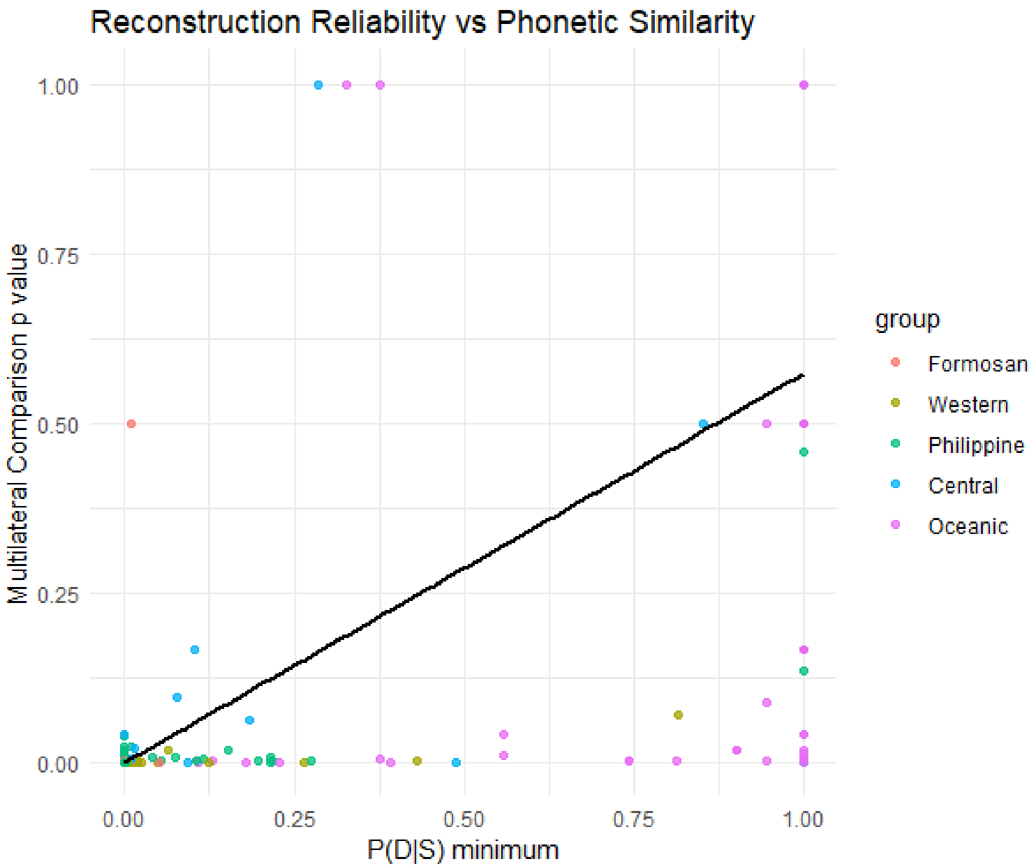


Figure 42: *P(D|S) and phonetic distance*. Plot of phonetic distance *p* values between a mother and daughter wordlist derived through multilateral comparison (y-axis) against *P*(*D*|*S*) minima of simulated reconstructions between the two (x-axis). Phonetic distance was calculated using the values in Table 16 (Ceolin, 2019). A linear line of best fit (black) was calculated using the stat_smooth() function in *R*.

Notice that in Figure 42, most of the data on both the y-axis and the x-axis is concentrated on the edges. It appears that for both multilateral comparison and WDT simulated annealing, most comparisons in the Austronesian dataset yielded extreme results. In other words, a random permutation of a daughter wordlist was either extremely unlikely to be equally or less distant from the mother or (presumably for unrelated or distantly related languages) extremely likely. Likewise, for the dataset tested, a reconstruction from a proto-language to an attested language either required too few changes for a random wordlist to be derivable through the same amount or it required so many that most random wordlists could be derived through the same amount.

Nevertheless, a linear line of best fit (black in Figure 42) shows that the two measures are positively correlated. Pairs of wordlists that are phonetically similar beyond reasonable doubt also tend merit a reconstruction requiring fewer changes. The adjusted R-squared for the correlation is 0.5628. In other words, although phonetic similarity plays no direct role in the computation of *P*(*D*|*S*), this measure to some extent reflects the amount of diachronic change between two wordlists.

As can be seen in Figure 43, there is no correlation between phonetic distance between wordlists and the number of sounds changes posited in the reconstruction from one to the other.

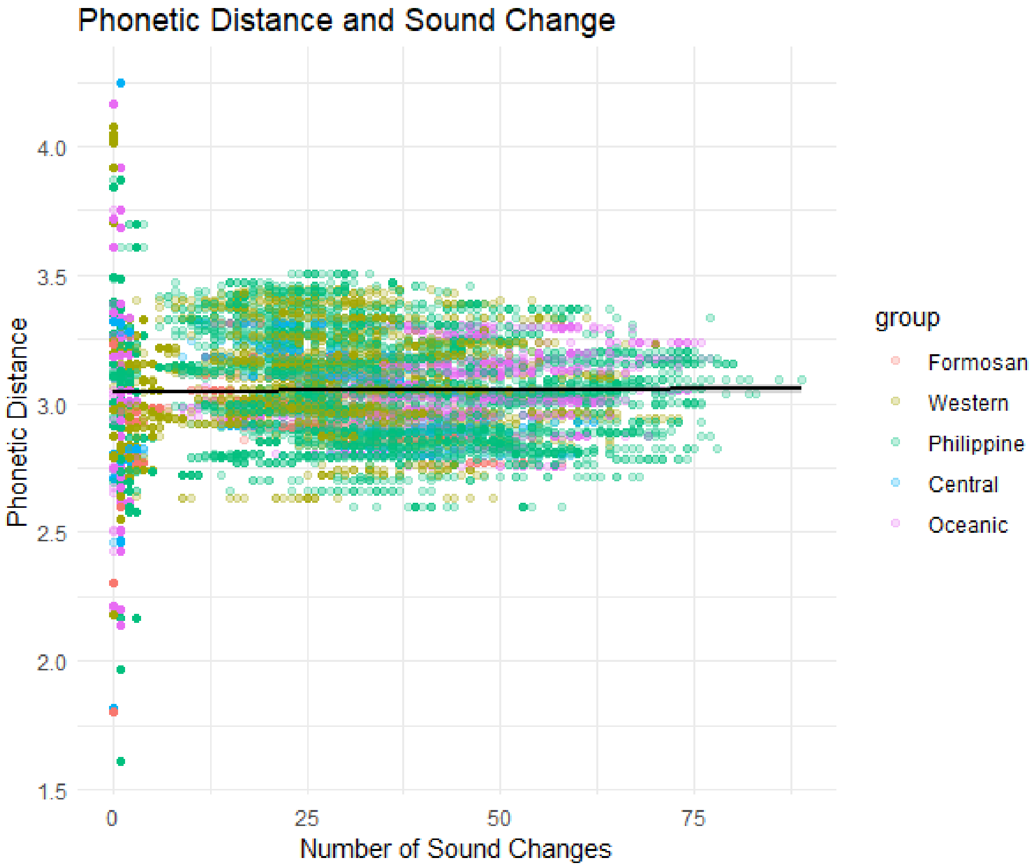


Figure 43: *Phonetic distance and sound change*. Plot of mean pairwise phonetic distance between two wordlists (y-axis) against the number of sound changes required in a simulated reconstruction between the two wordlists (x-axis). Phonetic distance was calculated using the values in Table 16 (Ceolin, 2019). A linear line of best fit (black) was calculated using the stat_smooth() function in *R*.

Wordlists in the dataset were a phonetic distance of 3.05 apart on average. Stated differently, the average distance between the major place of articulation values in Table 1 between two segments in the dataset was 3.05. This value is comparable to the results of previous studies using the same scale (Kessler & Lehtonen, 2006; Ceolin, 2019). Notably, this average distance persisted regardless of the number of sound changes required in the reconstruction. Controlling for proto-language or language group did not improve the correlation.

As such, it was not the case that wordlists that exhibited a greater phonetic mismatch required a greater number of sound changes in the reconstruction. These results suggest that phonetic distance is a poor proxy for the number of sound changes. Likely, this is due to the fact that a single sound change can alter the phonetic distance between two wordlists by a different amount depending on the number of word-pairs that it applies to. As an example, two different sound changes that alter two segments in two different ways affect the phonetic distance in the same way as one sound change that alters two of the same segments in the same way. Therefore, although both phonetic distance and the number of sound changes deal with mismatches in the segment domain, the two measurements appear to be largely independent in this dataset.

Finally, as can be seen in Figure 44, the number of possible words as estimated using WDT is correlated with the likelihood that a random premutation of the daughter wordlist is found to be equally or less phonetically distant from the mother. Wordlists that exhibited a lower word complexity, either due to a smaller phonological inventory or stricter phonotactics, were more likely to yield permutations that were equally phonetically distant to the mother.

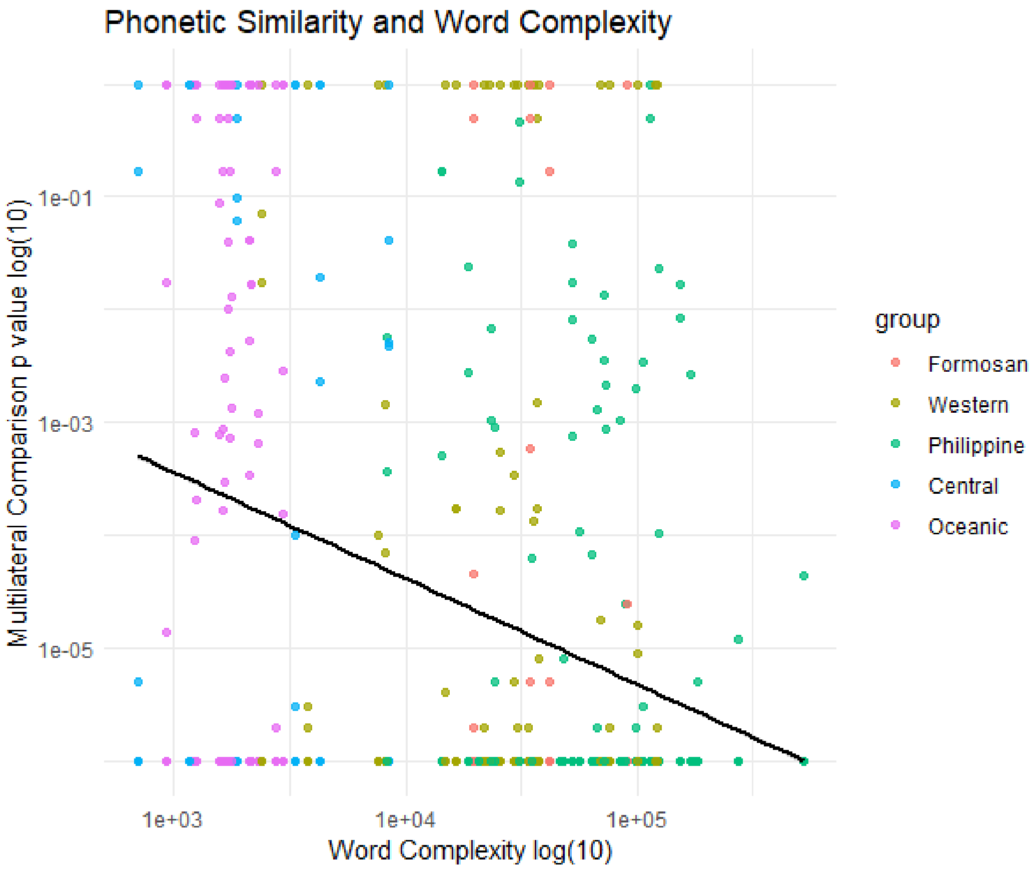


Figure 44: Plot of the multilateral comparison *p* value (y-axis) against the estimated number of possible words in the daughter language (x-axis). Note that both axes are logarithmic. Phonetic distance was calculated using the values in Table 16 (Ceolin, 2019). A linear line of best fit (black) was calculated using the stat_smooth() function in *R*.

Accounting for synchronic differences between languages is a necessary component of any probabilistic approach to language or language genealogy (Ringe, 1995). Recall that, in WDT, the number and type of diachronic transformations required for a reconstruction from a mother wordlist to a daughter wordlist must be interpreted with respect to the phonological properties of the daughter. Specifically, the estimated number of possible words forms part of the denominator of General Formula (see Section 2.1). As such, wordlists that exhibit a smaller number of phonologically possible words are easier to derive from a random wordlist. In fact, the relatively high *P*(*D*|*S*) values found in reconstructions to the Oceanic languages were partially explained by the low word complexity of languages in group (see Section 4.4.2.2).

It is somewhat striking, therefore, that a measure developed to account for synchronic differences between wordlists in WDT is correlated with the results of a multilateral comparison.

Although the two methodologies operate differently, one mitigates against the synchronic properties of the language through simulation, the other does so theoretically, they appear to come to a similar conclusion. Patterns across wordlists, be they phonetic similarity or ease of diachronic derivability, are easier to spot in wordlists that are less phonologically complex. As a result, *ceteris paribus*, wordlists exhibiting fewer phonological structures require more data to establish significance. In a way, the correlation between the result of simulated annealing and a word complexity as calculated using WDT is not surprising and serves to vindicate both methodologies.

4.6 Chapter Summary

The results of the case study are generally in line with expectations. The simulated annealing algorithm was able to make use of $P(D|S)$, as calculated using Wordlist Distortion Theory, as a cost function in the evaluation of stochastically generated comparative reconstructions. Using the simulated annealing algorithm, reconstructions with $P(D|S)$ lower than 0.00001, the threshold of reliability chosen, was found between all Austronesian proto-languages and known descendants. No reliable reconstructions were found between any of the proto-languages and the English controls. The reliability of a simulated reconstruction between a proto-language and descendant is correlated with statistical measures of significance found through previous methodology, namely multilateral comparison (Kessler & Lehtonen, 2006; Ceolin, 2019).

With the exception of Proto-West-Malayo-Polynesian, reconstruction reliability was correlated with proto-language time-depth. In general, reconstructions from a more distant ancestor to a descendant yielded higher $P(D|S)$ than reconstructions from more recent ancestors. There were a few individual languages which elicited lower $P(D|S)$ values from distant ancestors than from recent ancestors. However, some of these cases appear to have an explanation in the Austronesian literature. Most notably, Philippine languages which elicited particularly high $P(D|S)$ values from Proto-Philippine seem to have been influenced by neighboring non-Philippine languages in Indonesian Sulawesi. Additionally, Chamorro may have elicited particularly high $P(D|S)$ values from Proto-Western-Malayo-Polynesian because this classification may not accurately reflect the language's position in the Austronesian family.

The evaluation of the Ongan-Austronesian hypothesis is promising but not definitive. Reliable reconstructions from Proto-Ongan-Austronesian were found to 26 of the 74 Austronesian languages tested (35%). However, it is unclear if a reconstruction was found from Proto-Ongan-Austronesian to either of the two Ongan languages. If such a reconstruction is found in the future, perhaps through a different implementation of Wordlist Distortion Theory, it would constitute strong evidence in favor of an Ongan-Austronesian connection.

The $P(D|S)$ of the reconstructions was mostly determined by the number of lexical replacements required in the reconstruction, as well as the word complexity of the daughter language. Nevertheless, the simulated annealing algorithm successfully explored the space of possible reconstructions by suggesting sound changes. While not equipped with a definition of sound change naturalness, the algorithm found changes that appear phonologically natural and are independently justified in the Austronesian literature. Future computational implementations of Wordlist Distortion Theory can do the same for other types of diachronic transformations, such as semantic change or morphological change.

Chapter 5
Conclusion

This dissertation introduces Wordlist Distortion Theory, as well as $P(D|S)$, its measure of reconstruction reliability. The measure is the first of its kind, as no method for the probabilistic evaluation of comparative reconstruction has been proposed previously. The framework is distinct from computational approaches in historical linguistics developed to estimate genetic proximity based on measures of surface similarity, e.g. lexicostatistics and multilateral comparison. Instead, WDT is a method of evaluating the reconstruction itself as a series of transformations posited by the researcher. Although the dissertation briefly discusses other possible domains where such thinking is applicable, e.g. chess and photo-editing, it is particularly valuable in disciplines such as historical linguistics, where the number of transformations between states is independently notable and of interest to specialists.

The applications of Wordlist Distortion Theory are far-reaching and can be broadly classified into several categories. First and foremost, WDT can act as a platform for objective and accessible debate surrounding reconstructions in historical linguistics. The arguments in WDT are quantitative and interpretable by non-specialists. Moreover, the effect of arguments on reconstruction reliability is made immediately clear by the $P(D|S)$ measure. In particular, this framework is expected to expedite the debate surrounding proto-languages at greater time depths, where agreement from traditional qualitative arguments is not forthcoming. As a result, WDT may shed light on the temporal limits of the comparative method itself. As a start, this dissertation provides nearly unequivocal evidence in favor of a Proto-Austronesian language in the case study in Chapter 4, a language removed from the present by approximately 5,500 years (Blust, 2019).

Second, the framework is perfectly equipped for the comparison of competing analyses for the same dataset, as it evaluates the reconstruction rather than the data. Thus, reconstructions positing clashing diachronic transformations or accounting for different subsets of the data can simply be evaluated independently by the framework. In fact, WDT evaluation can be incorporated into the comparative method itself. Even before the reconstruction is complete, the effect of a proposed diachronic transformation on $P(D|S)$ can be quantified, as demonstrated in Section 3.2.2, and the change can be vetted if the effect decreases reconstruction reliability.

Third, WDT allows for general arguments about the reconstruction process, which in turn serve to assist in the comparative method. Thus, Section 2.3.1 demonstrates that the effect of semantic shift on $P(D|S)$ is relatively minor, implying that reconstructions comparing words of different meanings can still be reliable. In a similar vein, Section 3.3 demonstrates that across-the-board shift has an almost negligible effect on $P(D|S)$, indicating that a complete phonetic mismatch between comparanda has little bearing on reconstruction reliability. Furthermore, Section 3.2.2 presents a convenient rule-of thumb which states that positing one conditioned sound change for every two segments accounted for in the daughter wordlist will never increase $P(D|S)$.

Finally, $P(D|S)$, the quantitative proxy for reconstruction reliability as calculated by Wordlist Distortion Theory, can be employed as a cost function in machine learning approaches to historical linguistics. Because the measure evaluates the reconstruction, rather than the data, $P(D|S)$ allows for the discrimination between less desirable and more desirable reconstructions, paving the way for total or partial automation of the comparative method in the future. Chapter 4 demonstrates the utility of $P(D|S)$ as a component of a simulated annealing machine learning algorithm. The algorithm performed well in a case study of 74 attested languages and 5 proto-languages in the Austronesian family and was able to generate reconstructions from an Austronesian proto-language to a known descendant in all cases with a $P(D|S)$ well below the threshold of reliability. As expected, the algorithm found reconstructions with lower $P(D|S)$ values for recent ancestors and higher $P(D|S)$ values for more distant ones, all while suggesting changes which are phonologically natural and attested in the literature.

At the heart of Wordlist Distortion Theory lies the General Formula, introduced in Section 1.2.3. The formula estimates the likelihood that a randomly generated wordlist would merit the same number and type of changes from the mother wordlist as in a reconstruction to the daughter wordlist. Thus, WDT is a method of estimating the likelihood that evidence of a particular change or combination of changes can appear in random data. The estimate rests on the definition of two sets of wordlists: the synchronically compatible set, the set of wordlists exhibiting the phonological properties of the daughter, and the diachronically local set, the set of wordlists derivable from the mother through the same number and type of diachronic transformations.

The bulk of this dissertation is focused on providing formulae for estimating the number of diachronically local and synchronically compatible wordlists implied by different types of diachronic transformations. Thus, Chapter 2 provides estimates for the number of synchronically compatible and diachronically local wordlists introduced by lexical replacement and semantic shift, while Chapter 3 is concerned with the number of synchronically compatible wordlists introduced by various types of sound change. A working model of diachronic change in Wordlist Distortion Theory requires various assumptions about lexical replacement, semantic change, and sound change. The accuracy of formulae for capturing the effects of these individual types of transformations rests on the assumptions made. It is possible, likely even, that different assumptions in the future will lead to different formulae and different versions of Wordlist Distortion Theory. It is only the General Formula – which defines the relationship between the size of the synchronically compatible and diachronically local sets and reconstruction reliability – that is indispensable to the project.

A few avenues for future work on Wordlist Distortion Theory are clear. Firstly, progress can be made by applying the methodology laid out in this dissertation to other languages and language families. Results from more language families are useful in both the study of the particular language family as well as the growth of WDT more generally. As is the case for the Austronesian language family in this dissertation, WDT can provide probabilistic evidence in favor of subgroupings and reveal insights about the family's history. Additionally, a larger sample of languages would aid in the interpretation of result given by the framework. For instance, it is not currently clear what standard of reliability should be used for reconstruction $P(D|S)$. As this is not a question for WDT itself but rather the scholarly community, a larger body of work on the topic is required before it can be properly addressed.

Secondly, the individual formula introduced throughout this dissertation can be adjusted to incorporate alternate assumptions into the framework. Much of what is known about language and language change is omitted from WDT in its current form. Most notably, the formulae treat all changes of the same type as equally likely. Thus, the sound change *y > [i] is analyzed in the same way as *y > [ʎ]. Section 3.5 discusses the implementation of phonological features, which would address this issue in the domain of sound change. However, the same may be done for semantic change and lexical replacement, where the likelihood of seeing a particular transformation varies by entry. It remains to be seen how the asymmetry in transformation likelihood is to be incorporated into the mathematical backbone of the framework.

Thirdly, additional formulae may be introduced to expand the scope of Wordlist Distortion Theory. The framework can in principle model morphological change, sporadic sound change, analogy, onomatopoeia, as well as any other well-defined and attested transformation. The logic of the formulae used to derive diachronic transformations covered in this dissertation can be extended to these other changes as well. For each new method of distorting the wordlist, one simply needs to estimate the number of wordlists that are synchronically compatible with and diachronically local to the attested daughter given the reconstruction. In addition to increasing the accuracy of WDT evaluations, such developments may yield insights about the individual diachronic transformations.

Finally, future research may choose to focus on the theoretical implications of Wordlist Distortion Theory more broadly. A myriad of aspects, both in synchronic and diachronic linguistics, interact with reconstruction reliability in non-intuitive ways but remain unexplored in this dissertation. For instance, WDT could be used to estimate the number of distortions which would render the wordlist unreconstructible given a particular standard of reliability. The framework can also be used to test competing assumptions about synchronic and diachronic linguistics to see which ones are more conducive to comparative reconstruction evaluation, i.e. which ones yield a lower $P(D|S)$.

Ultimately, Wordlist Distortion Theory is simply a way of thinking about comparative reconstruction and language change. Each change is evaluated based on the evidence that the change leaves behind and the likelihood of observing the evidence at random. As such, WDT is compatible with almost any avenue of inquiry of interest to historical linguistics. The framework should be viewed as an extension of the comparative method rather than as an alternative. It cannot give results contradictory to the comparative method, as its role is to evaluate the product of the comparative method. This project is intended to bridge the gap between traditional and quantitative methods in historical linguistics under the assumption that the two are most effective unison rather than in parallel.

References

Aarts, E. & Laarhoven, P. (1989). Simulated Annealing: an introduction. *Statistica Neerlandica* 43(1), 31-52. doi:10.1111/j.1467-9574.1989.tb01245.x

Abbi, A. (2009). Is Great Andamanese genetically and typologically distinct from Onge and Jarawa? *Language Sciences*, 31, 791-812. doi:10.1016/j.langsci.2008.02.002

Adelaar, K. (1989). Malay influence on Malagasy: Linguistic and culture-historical implications. *Oceanic Linguistics*, 28(1), 1-46. doi:10.2307/3622973

Arlotto, A. (1981). *Introduction to Historical Linguistics*. University Press of America.

Atiqullah, M. (2004). An efficient simple cooling schedule for simulated annealing. *International Conference on Computational Science and Its Applications - ICCSA 2004*, 396-404.

Atkinson, Q., & Gray, R. (2005). Curious parallels and curious connections—Phylogenetic thinking in biology and historical linguistics. *Systematic Biology*, 54(4), 513-526. doi:10.1080/10635150590950317

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium.

Baković, E. (2011). Opacity and ordering. In Goldsmith, J., Riggle & Yu, A. (Eds.) *The Handbook of Phonological Theory*. Blackwell.

Barlow, R (2023). Papuan-Austronesian contact and the spread of numeral systems in Melanesia. *Diachronica*, 40(3), 287-240. doi:10.1075/dia.22005.bar

Baxter, W. (1995). A stronger affinity than could have been produced by accident: a probabilistic comparison of Old Chinese and Tibeto-Burman. *Journal of Chinese Linguistic Monograph Series*, 1-39.

Baxter, W. (1996). [Review of “On calculating the factor of chance in language comparison”, Donald A. Ringe, Jr.]. *Diachronica*, 13(2), 371-384. doi:10.1075/dia.13.2.11bax

Baxter, W. & Ramer, A. (2000). Byond lumping and splitting: probabilistic issues in historical linguistics. In Renfrew, C. & McMahon, A. (Eds.) *Time Depth in Historical Linguistics*. Cambridge.

Beguš, G. (2018). *Unnatural Phonology: A Synchrony-Diachrony Interface Approach* [Doctoral dissertation]. Harvard University.

Benedict, P. (1976). Austro-Thai and Austroasiatic. *Oceanic Linguistics Special Publications*, 13(1), 1-36.

Benjamin, G. (2012). The Aslian languages of Malaysia and Thailand: An assessment. In McGill, S. & Austin, P. (Eds.) *Language Documentation and Description* (pp. 136-230). SOAS.

Berezkin, Y. (2019). Athabaskan-Siberian folklore links: In search of Na-Dene origins. *Folklore*, 130(1), 31-47.

Bertsimas, D. & Tsitsiklis, J. (1993). Simulated annealing. *Statistical Science*, 8(1), 10-15. doi: 10.1214/ss/1177011077

Blaho, S. (2008). *The syntax of Phonology: A Radically Substance-free Approach* [Doctoral dissertation]. University of Tromso.

Blevins, J. (2007). A long lost sister of Proto-Austronesian?: Proto-Ongan, Mother of Jarawa and Onge of the Andaman Islands. *Oceanic Linguistics*, 46(1), 154-198. doi:10.1353/ol.2007.0015

Blevins, J., & Wedel, A. (2009). Inhibited sound change. *Diachronica*, 26(2), 143–183. doi:10.1075/dia.26.2.01ble

Blevins, J. (2020). Linguistic clues to Andamanese prehistory: Understanding the north-south divide. In Guildemann, T., McConvell, P. & Rhodes, R. (Eds.) *The Language of Hunter-Gatherers* (pp. 198-229). Cambrdige University Press.

Blust, R. (2000). Chamorro historical phonology. *Oceanic Linguistics*, 39(1), 83-122. doi:10.2307/3623218

Blust, R.t (2013). *The Austronesian Languages*. Pacific Linguistics Press.

Blust, R. (2014). Some recent proposals concerning the classification of the Austronesian Languages. *Oceanic Linguistics*, 53(2), 300-391. doi:10.1353/ol.2014.0025

Blust, R. (2019). The Austronesian homeland and dispersal. *The Annual Review of Linguistics*, 5, 417-434. doi:10.1146/annurev-linguistics-011718-012440

Bogart, K. (1983). *Introductory Combinatorics*. Brooks Cole.

Bomhard, A. (2008). *A Comprehensive Introduction to Nostratic Comparative Linguistics*. Florence.

Bostoen, K. (2007). Pots, words and the Bantu problem: On lexical reconstruction and early African history. *The Journal of African History,* 48(2), 173-199. doi:10.1017/s002185370700254x

Bowern, C., & Atkinson, Q. (2012). Computational phylogenetics and the internal structure of Pama-Nyungan. *Language*, 88(4), 817-845. doi:10.1353/lan.2012.0081

Boyadzhiev, K. (2018). *Notes on the binomial transform: Theory and table with appendix on Stirling Transform*. World Scientific Publishing Co. Pte Ltd.

Brown, C., Holman, W., Wichmann, S., & Velupillai, V. (2008). Automated classification of the World′s languages: A description of the method and preliminary results. *Language Typology and Universals*, 61(4), 285–308. doi:10.1524/stuf.2008.0026

Brown, C., Beck, D., Kondrak, G., Watters, J. & Wichmann, S. (2011). Totozoquean. *Linguistics*, 77(3), 323-372. doi:10.1086/660972

Brown, C., Wichmann, S. & Beck, D. (2014). Chitimacha: A Mesoamerican language in the lower Mississippi Valley. *International Journal of American Linguistics,* 80(4), 425-474. doi:10.1086/677911

Burridge, K., & Benczes, R. (2018). Taboo as a driver of language change. In Allen K. (Ed.) *The Oxford Handbook of Taboo Words and Language* (pp. 179–198).

Campbell, L. (1998). Nostratic: A personal assessment. In Salmons J. & Joseph B. (Eds.) *Nostratic: Sifting through the Evidence*. John Benjamins. doi:10.1075/cilt.142.08cam

Campbell, L. (2006). Why Sir William Jones got it all wrong, or Jones' role in how to establish language families. *Anuario del Seminario de Filologia Vasca*, 40(1-2), 245-264.

Campbell, L. (2011). The Dene–Yeniseian connection. *International Journal of American Linguistics*, 77(3), 445-451. doi:10.1086/660977

Campbell, L. (2013). *Historical Linguistics* (3rd edition). MIT press.

Campbell, L. (2017). How to show languages are related: Methods for distant genetic relationship. In Joseph, B. & Janda, R. *The Handbook of Historical Linguistics*. Blackwell.

Ceolin, A. (2019). Significance Testing of the Altaic Family. *Diachronica*, 36(3), 299-336. doi:10.1075/dia.17007.ceo

Chang, A. (2006). *A Reference Grammar of Paiwan* [Doctoral dissertation]. Australian National University.

Chang, W., Cathcart, C., Hall, D., & Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1), 194-244. doi:10.1353/lan.2015.0005

Clements, G. & Ridouane, R. (Eds.) (2011). *Where Do Features Come From? Cognitive, Physical and Developmental Bases of Distinctive Speech Categories*. John Benjamins.

Cogoy, E. (2004). *Onge Dictionary*. www.freelang.net/dictionary/onge.html.

Collinder, B. (1965). *An Introduction to the Uralic Languages*. University of California Press.

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

Cowper, E. & Hall, D. (2015). Reductio ad discirimen: Where features come from. *Nordlyd*, 41(2), 145-164. doi:10.7557/12.3411

Croft, W. (2008). Evolutionary Linguistics. *Annual Review of Anthropology*, 37(1), 219-234. doi:10.1146/annurev.anthro.37.081407.085156

Degaetano-Ortlieb, S. & Teich, E. (2018). Using relative entropy fore detection and analysis of perdios of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workship on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pg. 22-33).

Diakonov, I. & Starostin, S. (1986). *Hurro-Urartian as an Eastern Caucasian Language*. R. Kitzinger.

Dixon, R., & Aikhenvald, A. (2006). *Complementation: A Cross-linguistic Typology*. Oxford University Press.

Doerfer, G. (1967) *Türkische und Mongolische Elemente in Neupersichen*. Wiesbaden.

Dolgopolsky, A. (1986). A probabilistic hypothesis concerning the oldest relationships among the language families in Northern Eurasia. In Shevoroshkin, V. & Markey, L. (eds.), *Typology, relationship and time*, 27–50. Karoma.

Dolgopolsky, A. (2012). *Nostratic Dictionary*. Cambridge: McDonald Institute for Archaeological Research.

Downey, S., Hallmark, B., Cox, M., Norquest, P., & Lansing, J. (2008). Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction. *Journal of Quantitative Linguistics*, 15(4), 340-369. doi:10.1080/09296170802326681

Dresher, E. (2009). *The Contrastive Hierarchy in Phonology*. Cambridge University Press.

Dunn, M. (2012). [Review of *Anthropological Papers of the University of Alaska*, Karl, J. & Potter, B. (Eds.)]. Language, 88(2), 429-432.

Eberhard, D., Simons G. & Fennig C. (2023). *Ethnologue: Languages of the World*. SIL International.

Embleton, S. (1986). *Statistics in Historical Linguistics* [doctoral dissertation]. University of Toronto.

Fahnrich, H. (2007). *Kartwelisches etymologisches Wörterbuch*. Brill.

Finley, S. (2011). The privileged status of locality in consonant harmony. *Journal of Memory and Language*, 65(1), 74-83. doi: 10.1016/j.jml.2011.02.006

Fortson, B. (2003). An approach to semantic change. In Joseph, D. & Janda, R. (Eds.) *The Encyclopedic Dictionary of Applied Linguistics: A Handbook for Language Teaching* (pg. 648 – 666). Blackwell.

Fox, A (1995). *Linguistic reconstruction. An introduction to theory and method*. Oxford University Press, Oxford.

Gafos, A. (2013). *Articulatory Basis for Strict Locality in Phonology*. Routledge.

Gallagher, G. (2019). Phonotactic knowledge and phonetically unnatural classes: The plain uvular in Cochabamba Quechua. *Phonology*, 36, 37-60. doi:10.1017/S0952675719000034

Gallego, M. (2014). *Tracing ancestry and descent: A reconstruction of the proto-Batanic language* [Master's thesis]. University of the Philippines.

Georg, S., Michalove, P., Ramer, A., & Sidwell, P. (1999). Telling general linguists about Altaic. *Journal of Linguistics*, 35(1), 65-98. doi:10.1017/s0022226798007312

Goddard, I. (1986). Pre-Cheyenne *y. In Shipley, W. (Ed.) *In honor of Mary Haas: Haas festival Conference on Native American Linguistics* (pp. 345-360).

Gordon, M. (2011). Methodological and theoretical issues in the study of Chain Shifting. *Language and Linguistics Compass, 5*(11), 784–794. doi:10.1111/j.1749-818x.2011.00310.x

Gray, R., & Atkinson, Q. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965), 435-439. doi:10.1038/nature02029

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

Greenhill, S., Wu C., Hua X., Dunn M., Levinson S. & Gray R. (2017). Evolutionary dynamics of language systems. *Anthropology*, 114(42), E8822-E8829. doi:10.1073/pnas.1700388114

Greenhill, S., Haynie, H., Ross, R., Chira, A., List, J., Campbell, L., Botero, C. & Gray, R. et al. (2023). A recent northern origin for the Uto-Aztecan family. *Language*, 99(1), 81-97. doi:10.1353/lan.2023.0006

Hamilton, W., Leskovec, J. & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pg. 1489-1501). Association for Computational Linguistics.

Hammarström, H., Forkel, R., Haspelmath, M. & Bank S. (2022). *Glottolog 4.7*. Max Planck Institute for Evolutionary Anthropology.

Hansen, M., & Visconti, J. (2009). *Current trends in diachronic semantics and pragmatics*. Emerald.

Harrison, S. (2003). On the limits of the comparative method. In Brain, J. & Janda, R. (Eds.) *The Handbook of Historical Linguistics* (pp. 213-243). Blackwell.

Hasan, M., & Rai, A. (2020). Groundwater quality assessment in the lower Ganga Basin using entropy information theory and GIS. *Journal of Cleaner Production*, 274. doi:10.1016/j.jclepro.2020.123077

Haspelmath, M. (2021). Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics*, *57*(3), 605–633. doi:10.1017/s0022226720000535

Heap, R. (1963). Permutations by interchanges. *The Computer Journal*, 6(3), 293-4. doi:10.1093/comjnl/6.3.293

Heggarty, P., Anderson, C., Scarboroguh, M., King, B., Bouckert, R., Jocz, L., Kümmel, M., Jügel, T., Irslinger, B., Pooth, R., Liljegren, H., Strand, R., Haig, G., Macák, M., Kim, R., Anonby, E., Pronk, T., Belyaev, O., Dewey-Findell, T., Boutilier, M., Freiberg, C., Tegethoff, R., Serangeli, M., Liosis, N., Stroński, K., Schulte, K., Gupta, G., Haak, W., Krause, J., Atkinson Q., Grenhill, S., Kühnert, D., Gray, R. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656), doi: 10.1126/science.abg0818

Heinz, J. (2010). Learning long-distance phonotactics. *Linguistics Inquiry*, 41(4), 623-661. doi:10.1162/LING_a_00015

Helwig, N. (2019). Statistical nonparametric mapping: Multivariate permutation tests for location, correlation, and regression problems in neuroimaging. *WIREs Computational Statistics*, 11(2). doi:10.1002/wics.1457

Herd, J. (2005). Loanword adaptation and the evaluation of similarity. *Toronto Working Papers in Linguistics*, 24.

Heye, J. & Hidalgo, C. (1967). An outline of Southern Ivatan phonology. *General Linguistics*, 7(2), 105.

Hill, N. (2014). Grammatically Conditioned Sound Change. *Language and Linguistic Compass*, 8(6), 211-229. doi:10.1111/lnc3.12073

Hock, H. (2009). *Principles of Historical Linguistics*. De Gruyter Mouton.

Hoenigswald, H. (2019). *Diachronic, Areal, and Typological Linguistics*. De Gruyter.

Holman, W., Brown, H., Wichmann, S., Muller, A., Velupillai, V., Hammarström, Sauppe, S., Jung, H., Bakker, D., Brown, P., Belyaev, O., Urban, M., Mailhammer, R., List, J. & Egorov, D. (2011). Automated dating of the World's language families based on lexical similarity. *Current Anthropology*, 52(6), 841-875. doi:10.1086/662127

Hruschka, D., Branford, S., Smith, E., Wilkins, J., Meade, A., Pagel, M., & Bhattacharya, T. (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, 25(1), 1-9. doi:10.1016/j.cub.2014.10.064

Illich-Svitych, V. (1963). Алтайские дентальные: t, d, ð. *Вопросы Языкознания*, 6, 37-56.

Iosad, P. (2014). The phonology and morphosyntax of mutation in Breton. *Lingue e linguaggio*, 13(1), 23-42. doi:10.1418/76998

Jäger, Gerhard (2019). Computational historical linguistics. *Theoretical Linguistics*, 45(3-4), 151-182. doi:10.1515/tl-2019-0011

Johanson, L. & Csato, E. (1998). *The Turkic Languages*. Routledge.

Jost, L. (2006). Entropy and Diversity. *Oikos*, *113*(2), 363–375. doi:10.1111/j.2006.0030-1299.14714.x

Kalping, G. & Klamer, M. (2018). LexiRumah: An online lexical database of the Lesser Sunda Islands. *PLoS ONE*, 13(10), e0205250. doi:10.1371/journal.pone.0205250

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

Kay, M. (1964). *The logic of cognate recognition in historical linguistics*. RAND Corporation.

Kessler, B. & Lehtonen, A. (2006). Multilateral comparison and significance testing of the Indo-Uralic question. In Forster, P. & Renfrew, C. *Phylogenetic Methods and the Prehistory of Languages* (pp. 33-42). McDonald Institute of Archaeological Research.

Kessler, B. (2008). The mathematical assessment of long-range linguistic relationships. *Language and Linguistics Compass*, 2(5), 821-839. doi:10.1111/j.1749-818x.2008.00083.x

Kessler, B. (2015). Computational and quantitative approaches to historical phonology. *Oxford Handbooks Online*. doi:10.1093/oxfordhb/9780199232819.013.030

Kiparsky, P. (2015). New perspectives in historical linguistics. In Bowern C. & B. Evans (Eds.) *The Routledge Handbook of Historical Linguistics* (pp. 64-102). Routledge. doi:10.4324/9781315794013.ch2

Klamer, M. (2019). The dispersal of Austronesian languages in Island South East Asia: Current findings and debates. *Language and Linguistics Compass*, 13(4). doi:10.1111/lnc3.12325

Kondrak, G. (2002). *Algorithms for Language Reconstruction* [doctoral dissertation]. University of Toronto.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and the Humanities*, 37(3), 273-291. doi:10.1023/a:1025071200644

Kondrak, G. (2009). Identification of Cognates and Recurrent Sound Correspondences in Word Lists. *Traitement Automatique des Langues*, 50(2), 201–235.

Krauss, M. (2009). St. Lawrence Eskimo phonology and orthography. *Linguistics*, 13(152), 39-72. doi:10.1515/ling.1975.13.152.39

Kulikov, L. & Lavidas, N. (2015). *Proto-European Syntax and its Development*. John Benjamins.

Kuper, R. (2020). Preference and restorative potential for landscape models that depict diverse arrangements of defoliated, foliated, and Evergreen plants. *Urban Forestry & Urban Greening*, *48*. doi:10.1016/j.ufug.2019.126570

Lenstra, J. & Kan, A. (1975). Some simple applications of the travelling salesman problem. *Journal of the Operational Research Society*, 26(4), 717-733. doi:10.1057/jors.1975.1

Li, R., Wang, M., Jin, L. & He, Y. (2013). A Monte Carlo permutation test for random mating using genome sequences. *PLoS ONE*, 8(8). doi:10.1371/journal.pone.0071496

Li, Z., Luo, Z., Wang, Y., Fan, G., & Zhang, J. (2022). Suitability evaluation system for the shallow geothermal energy implementation in region by entropy weight method and Topsis method. *Renewable Energy*, 184, 564–576. doi:10.1016/j.renene.2021.11.112

Lindgren, N (1989). The use of ѣ in manuscripts of the XVII century. *Russian Linguistics*, *13*(1), 15–32. doi:10.1007/bf02527987

List, J. (2014). *Sequence comparison in historical linguistics* [doctoral dissertation]. Dusseldorf University.

List, J. (2019). Automatic Inference of Sound Correspondence Patterns across Multiple Languages. *Computational Linguistics*, 45(1), 137–161. doi:10.1162/coli_a_00344

Longobardi, G, Guardiano, C., Silvestri, G., Boattini, A. & Ceolin, A. Toward a syntactic phylogeny of modern Indo-European languages. *Journal of Historical Linguistics*, 3(1), 122-152. doi:10.1075/jhl.3.1.07lon

Lowe, J. & Mazaudon, M. (1994). The reconstruction engine: A computer implementation of the comparative method. *Computational Linguistics*, 20(3), 381–417.

Lubbe, J. (2018). *Information Theory*. Cambridge University Press.

Lynch, J., Ross, M. & Crowley, T. (2002). *The Oceanic Languages*. Curzon.

Marr, C. & Mortensen, D. (2023). Large-scale computerized forward reconstruction yields new perspectives in French diachronic phonology. *Diachronica*, 40(2), 238-285. doi:10.1075/dia.20027.mar

Matisoff, J. (1990). On megalocomparison. *Language*, 66(1), 106-120. doi:10.2307/415281

McMahon, A., & McMahon, R. (2003). Finding families: quantitative methods in language classification. *Transactions of the Philological Society*, 101(1), 7-55. doi:10.1111/1467-968x.00108

Michalove, P., Georg, S., & Ramer, A. (1998). Current issues in linguistic taxonomy. *Annual Review of Anthropology*, 27(1), 451-472. doi:10.1146/annurev.anthro.27.1.451

Mielke, J. (2008). *The Emergence of Distinctive Features*. Oxford University Press.

Miller, R. (1984). The classification of the Uto-Aztecan Languages based on lexical evidence. *International Journal of American Linguistics*, 50(1), 1-24. doi:10.1086/465813

Mitchell, B. & Robinson, C. (2012). *A Guide to Old English* (8th edition). Wiley-Blackwell.

Miyake, M. (2003). *Old Japanese: a phonetic reconstruction*. RoutledgeCurzon.

Montemurro, M., & Zanette, D. (2011). Universal entropy of word ordering across linguistic families. *PLoS ONE*, *6*(5). doi:10.1371/journal.pone.0019875

Morén, B. (2003). Affricates, palatals, and iotization in Serbian: Representational solutions to longstanding puzzles. *Poljarnyj Vestnik*, 6. 46–70. doi: 10.7557/6.1344

Nichols, J. (1996). The comparative method as heuristic. In Durie, M. & Ross, M. *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press.

Nichols, J., & Peterson, A. (1996). The Amerind personal pronouns. *Language*, 72(2), 336-371. doi:10.2307/416653

Norman, J. (2009). A new look at Altaic. *Journal of the American Oriental Society*, 129(1), 83-89. doi:10.2307/40593870

Nourani, Y. & Andresen, B. (1998). A comparison of simulated annealing cooling strategies. *Journal of Physics A: Mathematical and General*, 31(41), 8373-8385. doi:10.1088/0305-4470/31/41/011

Oakes, M. (2000). Computer Estimation of Vocabulary in a Protolanguage from Word Lists in Four Daughter Languages. *Journal of Quantitative Linguistics*, 7(3), 233-243. doi:10.1076/jqul.7.3.233.4105

Odden, D. (1994). Adjacency parameters in phonology. *Language*, 70(2), 89-330. doi: 10.2307/415830

Ogura, M., & Wang, W. (2018). Evolution of homophones and syntactic categories noun and verb. *Proceedings of the 12th International Conference on the Evolution of Language (Evolang12)*. doi:10.12775/3991-1.086

Oh, B. (2002). *A System for Image-Based Modeling and Photo Editing* [Doctoral dissertation]. Massachusetts Institute of Technology.

Osmond, M., Ross, M., & Pawley, A. (2007). *The Lexicon of Proto Oceanic: The Culture and Environment of Ancestral Oceanic Society: 2 the physical environment*. ANU E Press.

Oswalt, R. (1970). The detection of remote linguistics relationships. *Computer Studies in the Humanities and Verbal Behavior*, 3(3), 117-129.

Oxford, W. (2015). Patterns of contrast in phonological change: Evidence from Algonquian vowel systems. *Language*, 91(2), 308-358. doi:10.1353/lan.2015.0028

Pagel, M., Atkinson, Q., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449(7163), 717–720. doi:10.1038/nature06176

Pagel, M. (2017). Darwinian perspectives on the evolution of human languages. *Psychonomic Bulletin & Review*, 24(1), 151-157. doi:10.3758/s13423-016-1072-z

Pagel, M. & Meade, A. (2018). The deep history of the number words. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1740). doi:10.1098/rstb.2016.0517

Pereltsvaig, A. & Lewis, M. (2015). *The Indo-European Controversy*. Cambridge University Press.

Picard, M. (1990). On the evaluation of competing analyses in historical phonology: Naturalness, minimality and the case of Armenian /erk/. *Language Sciences*, 12(1), 85-99. doi:10.1016/0388-0001(90)90025-C

Pimentel, T., Roark, B., & Cotterell, R. (2020). Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*, 8, 1-18. doi:10.1162/tacl_a_00296

Pompei, S., Loreto, V. & Tria, F. (2011). On the accuracy of language trees. *PLoS ONE*, 6(6). doi:10.1371/journal.pone.0020109

Poppe N. (1960). Vergleichende Grammatik der Altaischen Sprachen. *Porta Linguarum Orientalum*, 4.

Power, J., Guido, G. & List, J. Evolutionary dynamics in the dispersal of sign languages. *Royal Society Open Science*, 7(1). doi:10.1098/rsos.191100

R Core Team (2019). *A language and environment for statistical computing.* http://www.r-project.org/.

Rama, T., List, J., Wahle, J. & Jäger, G. (2017). Fast and unsupervised methods for multilingual cognate clustering. arXiv:1702.04938. doi:10.48550/arXiv.1702.04938

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

Ramstedt, G. (1912). Zur Geschichte des labialen Spiranten in Mongolischen. In *Festchrift Vilhelm Thomson zur Vollendung des siebzigsten Lebenjahren* (pp. 182-187). Harrassowitz.

Rannala, B. & Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, 43, 304-311. doi:10.1007/BF02338839

Ratcliffe, R. (2003). On calculating the reliability of the comparative method at long and medium distances: Afroasiatic comparative lexica as a test case. *Journal of Historical Linguistics*, 2(2), 239-281. doi:10.1075/jhl.2.2.04rat

Reid, L. (1971). Philippine minor languages: Word lists and phonologies. *Oceanic Linguistics Special Publications*, 8.

Reid, L. (1994). Morphological evidence for Austric. *Oceanic Linguistics,* 33(2), 323-344. doi:10.2307/3623132

Reid, L. (1999). New linguistic evidence for the Austric Hypothesis. *Selected Papers from the Eighth International Conference on Austronesian Linguistics*, 5-30.

Reid, L. (2002). Morphosyntactic evidence for the position of Chamorro in the Austronesian language family. *Collected papers on southeast Asian and Pacific languages*, 63-94.

Renfrew, C. McMahon, A. & Trask, R. (2000). *Time depth in historical linguistics*. McDonald Institute for Archaeological Research.

Rexova, K., Frynta, D., & Zrzavy, J. (2003). Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics*, 19(2), 120-127. doi:10.1111/j.1096-0031.2003.tb00299.x

Ringe, D. (1999). How hard is it to match CVC-roots? *Transactions of the Philological Society*, 97(2), 213-244. doi:10.1111/1467-968x.00049

Ringe, D. (2017). *From Proto-Indo-European to Proto-Germanic*. Oxford University Press.

Robbeets, M. (2005). *Is Japanese related to Korean, Tungusic, Mongolic and Turkic?* Harrassowitz.

Robbeets, M., Bouckaert, R., Conte, M., Savelyev, A., Li, T., An, D., Shinoda, K., Cui, Y., Kawashima, T., Kim, G., Uchiyama, J., Dolińska, J., Oskolskaya, S., Yamano, K., Seguchi, N., Tomita, H., Takamiya, H., Kanzawa-Kiriyama, H., Oota, H., Ishida, H., Kimura, R., Sato, T., Kim, J., Deng, B., Bjørn, R., Rhee, S., Ahn, K., Gruntov, I., Mazo, O., Bentley, J., Fernandes, R., Roberts, P., Bausch, I., Gilaizeau, L., Yoneda, M. Kugai, M., Bianco, R., Zhang, F., Himmel, M., Hudson, M. & Ning, C. (2021). Triangulation supports agricultural spread of the Transeurasian languages. *Nature*, 599, 616-621. doi:10.1038/s41586-021-04108-8

Ross, M. & Durie, M (1996). Introduction. In Durie, M. & Ross, M. *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*. Oxford University Press.

Ross, M. (2005). The Batanic Languages in Relation to the Early History of the Malayo-Polynesian Subgroup of Austronesian. *Journal of Austronesian Studies*, 1(2), 1-24.

Ross, M. (2020). Comment on Blust "The resurrection of proto-Philippines". *Oceanic Linguistics*, 1, 366-373.

Roughan, M., Mitchell, L., & South, T. (2020). How the avengers assemble: Ecological modelling of effective cast sizes for movies. *PLoS ONE*, *15*(2). doi:10.1371/journal.pone.0223833

Rubicz, R. Melvin, K & Crawford M. (2002). Genetic evidence for the phylogenetic relationship between Na-Dene and Yeniseian speakers. *Human Biology*, 74(6), 743-760. doi:10.1353/hub.2003.0011

Ruhlen, M. (1994). *On the Origin of Languages*. Stanford University Press.

Rutenbar, R. (1989). Simulated annealing algorithms: An overview. *IEEE Circuits and Devices Magazine*, 5(1), 19-26. doi:10.1109/101.17235

Schwartz, J., Boe, L, Vallee, N. and Abry, C. (1997). Major trends in vowel system inventories. *Journal of Phonetics*, 25(3), 233-253. doi:10.1006/jpho.1997.0044

Shi, Y., & Lei, L. (2020). Lexical richness and text length: An entropy-based perspective. *Journal of Quantitative Linguistics*, *29*(1), 62–79. doi:10.1080/09296174.2020.1766346

Sicoli, A., & Holton, G. (2014). Linguistic phylogenies support back-migration from Beringia to Asia. *PLoS ONE*, 9(3). doi:10.1371/journal.pone.0091722

Sigurd, B., Eeg-Olofsson, M., & Weijer, J. V. (2004). Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica*, 58(1), 37-52. doi:10.1111/j.0039-3193.2004.00109.x

Silverman, D. (2009). Neutralization and anti-homophony in Korean. *Journal of Linguistics*, 46(2), 453–482. doi:10.1017/s0022226709990247

Smith, A. (2017). *The Languages of Borneo: A Comprehensive Classification* [doctoral dissertation]. University of Hawai'i.

Snedden, J. (1984). Proto-Sangiric and the Sangiric Languages. *Pacific Linguistics*, B(91).

Sneddon, J. (1989). The North Sulawesi microgroups: In search of higher level connections. *Studies in Sulawesi linguistics*, 83-107.

Starosta, S. (1995). A grammatical subgrouping of Formosan languages. *Symposium series of the Institute of History and Philology*, 3.

Starostin, G. (2012). Dene-Yeniseian: A critical assessment. *Journal of Language Relationship*, 8, 117-152. doi:10.31826/jlr-2012-080109

Starostin, S., Dybo, A. & Mudrak, O. (2003). *Etymological Dictionary of the Altaic Languages*. Brill.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21(2), 121-137. doi:10.1086/464321

Tahmasebi, N., Borin, L., Jatowt, A. (2021). Survey of computational approaches to lexical semantic change detection. In Tahmasebi, N, Borin, L., Jatowt, A., Xu, Y. & Hengchen, S. (Eds.) *Computational Approaches to Semantic Change* (pg. 1-91). Language Science Press.

Terrill, A. (2011). Languages in contact: An exploration of stability and change in the Solomon Islands. *Oceanic Linguistics*, 50(2), 312-337. doi:10.1353/ol.2011.0021

Theil, R. (2006). Is Omotic Afroasiatic? A critical discussion. *Proceedings from the David Dwyer retirement symposium*.

Topping, D. (1980). *Chamorro Reference Grammar.* University of Hawaii Press.

Traugott, E., & Dasher, R. (2001). *Regularity in semantic change*. Cambridge University Press.

Traugott, E. (2017). Semantic Change. *Oxford Research Encyclopedia of Linguistics*. doi: 10.1093/acrefore/9780199384655.013.323

Triki, E., Colette, Y. & Siarry, P. (2005). A theoretical study on the behaviour of simulated annealing leading to a new cooling schedule. *European Journal of Operational Research*, 166(1), 77-92. doi:10.1016/j.ejor.2004.03.035

Trussel, S. & Blust, R. (2010). *Austronesian Comparative Dictionary*. https://www.trussel2.com/ACD/introduction.htm

Tuck, A. (2006). Singing the rug: Patterned textiles and the origins of Indo-European Metrical Poetry. *American Journal of Archaeology*, 110, 539-550. doi:10.3764/aja.110.4.539

Um, Y. (2003). An ordering paradox as constraint interaction: Alternation of n and l in Korean. *Studies in Phonetics, Phonology, and Morphology*, 9(1), 111-131.

Vajda, E. (2010). A Siberian link with Na-Dene languages. In Karl, J. & Potter, B. (Eds.) *The Dene-Yeniseian Connection*. Anthropological Papers of the University of Alaska.

Vajda, E. (2011). A response to Campbell. *International Journal of American Linguistics*, 77(3), 451-452.

Vajda, E. (2018). Dene-Yeniseian. *Diachronica*, 35(2), 277-295. doi:10.1075/dia.18001.vaj

Vajda, E. Ives, J. & Rice, S. (2010). Dene-Yeniseian and processes of deep change in kin terminologies. *Archaeological Papers of the University of Alaska*, 6, 120-236.

Vejdemo, S., & Hörberg, T. (2016). Semantic factors predict the rate of lexical replacement of content words. *PLoS ONE*, 11(1). doi:10.1371/journal.pone.0147924

Vera, J., Fuentealba, D., Lopez, M., Ponce, H., & Zariquiey, R. (2021). On the von neumann entropy of language networks: Applications to cross-linguistic comparisons. *Europhysics Letters*, 136(6), 68003. https://doi.org/10.1209/0295-5075/ac39ee

Vovin, A. (2005). The end of the Altaic controversy in memory of Gerhard Doerfer. *Central Asiatic Journal*, 49(1), 71-132.

Vylomova, E. & Haslam, N. (2021). Semantic changes in harm-related concepts in English. In Tahmasebi, N, Borin, L., Jatowt, A., Xu, Y. & Hengchen, S. (Eds.) *Computational Approaches to Semantic Change* (pg. 93-121). Language Science Press.

Wichmann, S., Muller, A. & Velupillai, V. (2010). Homelands of the world's language families: A quantitative approach. *Diachronica*, 27(2), 247-276. doi:10.1075/dia.27.2.05wic

Wikander, S. (1967). Maya and Altaic. *Journal of Anthropology*, 32(1-4), 141-148. doi:10.1080/00141844.1967.9980993

Winter, B., & Wedel, A. (2016). The Co- evolution of Speech and the lexicon: The interaction of functional pressures, redundancy, and category variation. *Topics in Cognitive Science*, 8(2), 503–513. doi:10.1111/tops.12202

Wu, M., Schweikhard, N., Bodt, T., Hill, N. & List, J. (2020). Computer-Assisted Language Comparison: State of the ArtW. *Journal of Open Humanities Data*, 6(2), 1-14. doi:10.5334/johd.12

Xu, Y. & Kemp, C. (2015). A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015*. Pasadena.

Yang, W., Li, W., Cao, Y., Luo, Y., & He, L. (2020). Real-time production and logistics self-adaption scheduling based on information entropy theory. *Sensors*, 20(16), 4507. doi:10.3390/s20164507

Yi, K. & S. Ramsey (2011). *A History of the Korean language*. Cambridge University Press.

Yin, S., & White, J. (2018). Neutralization and homophony avoidance in phonological learning. *Cognition*, 179, 89–101. doi:10.1016/j.cognition.2018.05.023

Zhang, M., & Gong, T. (2016). How many is enough?—Statistical principles for lexicostatistics. *Frontiers in Psychology*, 7. doi:10.3389/fpsyg.2016.019161086/660978

Appendix A

This section contains all of the critical formulae required to perform an analysis using Wordlist Distortion Theory. For explanations and justifications, see original text.

(1)

$$P(D|S) = \frac{|D \cap S|}{|S|}$$

General Formula: Formula for calculating the probability that a wordlist generated at random in accordance with the phonological properties of the daughter wordlist evidences the same number and type of transformations as those in the reconstruction. *S* is the set of wordlists synchronically compatible with the daughter. *D* is the set of wordlists diachronically local to the daughter given the proposed reconstruction. See Section 1.2.3.

(2)

$$|S| = c^t$$

Synchronic Formula: Formula for calculating the size of the synchronically compatible wordlist set |*S*|, given the number of possible words *c* and wordlist size *t*. See Section 2.1.

(3)

$$|D \cap S|_l = \sum_{i=0}^{l} (c-1)^i \binom{t}{i}$$

Replacement Formula: Formula for calculating the number of wordlists that are both synchronically compatible with and diachronically local to the daughter in a reconstruction that posits a number of lost lexemes *l*, in a wordlist of length *t*, with word complexity *c*. See Section 2.2.1.

(4)

$$|D \cap S| = c^l \binom{t}{l}$$

Replacement Approximation: Approximation of the number of wordlists that are both synchronically compatible with and diachronically local to the daughter in a reconstruction that posits a number of lost lexemes *l*, in a wordlist of length *t*. See Section 2.2.2.

(5)

$$|D \cap S|_s = \sum_{i=0}^{s} (\min\{c, t\} - 1)^i \binom{t}{i}$$

Semantic Change Formula: Formula for calculating the number of wordlists that are both synchronically compatible with and diachronically local to the daughter in a reconstruction that posits a number of semantic changes s, in a wordlist of length t¸ and possible words c. See Section 2.3.1.

(6)

$$c = \hbar^w$$

Word Complexity Formula: Formula for estimating the number of possible words c, using the mean word-length w and the mean number of homotopes $\hbar$. See Section 3.1.2.

(7)

$$|D \cap S\,|_\varphi = \hbar^\varphi$$

Merger Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by φ mergers with mean number of homotopes $\hbar$. See Section 3.2.1.

(8)

$$|D \cap S\,|_\zeta = 2^\zeta$$

Environment Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by ζ conditioning environments of mergers. See Section 3.2.2.

(9)

$$\varphi < \frac{-\Delta lw}{2}$$

Sound change vs lexical replacement comparison: Rule-of-thumb formula which, if satisfied, guarantees that the suggested sound change(s) decrease $P(D|S)$. See Section 3.2.3.

(10)

$$|D \cap S\,|_\chi = \chi!$$

Shift Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by χ shifts. See Section 3.3.

(11)

$$|D \cap S\,|_{\chi,\varphi} = \sum_{i=0}^{\chi} (-1)^{i} \binom{\chi}{i} (\hbar - i)^{\chi+\varphi}$$

Sound Change Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by χ shifts and φ mergers, with mean number of homotopes $\hbar$. See Section 3.3.1.

(12)

$$|D \cap S\,|_{\lambda} = \binom{\lambda + \varphi + \chi - 1}{\lambda} (\hbar r)^{\lambda}$$

Link Formula: Formula for calculating the number of wordlists diachronically local to and synchronically compatible with the daughter introduced by λ links distributed between φ mergers and χ shifts, with mean number of homotopic segments $\hbar$. The variable $\hbar$ is scaled by r to account for non-unique outputs of chains.
See Section 3.4.

(13)

$$|D \cap S| = \prod_{x \in X} |D \cap S|_{x}$$

Diachronic Formula: Formula for estimating the total number of diachronically local and synchronically compatible wordlists given the number of diachronically local wordlist introduced by *x* in *X*, where *X* is the set containing the number of diachronically local and synchronically compatible wordlists calculated by domain-specific formulas. See Section 3.6.

Appendix B

This section briefly discusses the similarities between the Regular Replacement Formula and the Binomial Formula in probability theory. It is shown that, when *c* = 2, calculating *P*(*D*|*S*) using the Regular Replacement Formula is identical to performing a one-sided binomial test.

Imagine that each entry in the wordlist stores not a phonological string but a simple result of a coin flip, H or T. Thus, in this 'language', word complexity *c* = 2, as each entry is selected from a bank of two possible 'words'. The 'wordlist' in this case is a run of coin flips of length *t*.

For languages where *c* = 2, calculating *P*(*D*|*S*) is equivalent to performing a one-sided binomial test of significance with *l* successful trials out of *t* total and 1/2 probability of success. Recall that the motivation behind *P*(*D*|*S*) is the same as for the traditional *p*-value in statistics. Both measures are calculated using three elements: the probability of the observed outcome, the probability of outcomes that are equally likely, the probability of outcomes that are less likely. Therefore, calculating *P*(*D*|*S*) for a reconstruction with *c* = 2 is the same as adding the probability of getting *l* or fewer heads out of *t*, where heads are expected in all trials (hence, a one-sided test). For these purposes, the General Formula can be simplified such that *c* = 2, as in (1).

(1)

$$\frac{\sum_{i=0}^{l}(2-1)^{l}\binom{t}{l}}{2^{t}} = \frac{\sum_{i=0}^{l}\binom{t}{l}}{2^{t}}$$

We can confirm that the General Formula and the one-sided binomial test of significance give the same result by comparing the two in Table 1. Here, results are given for a run of 4 coin flips. *P*(*D*|*S*) is calculated using Wordlist Distortion Theory, whereas the binomial test was run using the *binom.test* function in *R*. The values in the two columns are identical.

Table 1: Comparison of General Formula and binomial test

replacements	*general formula*	*binomial test*
0/4	$P(D\|S) = \frac{\binom{4}{0}}{2^4} = \frac{1}{16}$	*p*=.0625
1/4	$P(D\|S) = \frac{\binom{4}{0}+\binom{4}{1}}{2^4} = \frac{1+4}{2^4} = \frac{5}{16}$	*p*=.3125
2/4	$P(D\|S) = \frac{\binom{4}{0}+\binom{4}{1}+\binom{4}{2}}{2^4} = \frac{1+4+6}{2^4} = \frac{11}{16}$	*p*=.6875
3/4	$P(D\|S) = \frac{\binom{4}{0}+\binom{4}{1}+\binom{4}{2}+\binom{4}{3}}{2^4} = \frac{1+4+6+4}{2^4} = \frac{15}{16}$	*p*=.9375
4/4	$P(D\|S) = \frac{\binom{4}{0}+\binom{4}{1}+\binom{4}{2}+\binom{4}{3}+\binom{4}{4}}{2^4} = \frac{1+4+6+4+1}{2^4} = \frac{16}{16}$	*p*=1

In Wordlist Distortion Theory, *P*(*D*|*S*) is the estimated probability that a diachronically equivalent wordlist is synchronically equivalent. In the binomial test, *p* is the probability of an event equally or more extreme to the one observed. In this simplified example, with diachronic equivalency treated as the dependent variable in the binomial test, the two are identical.