Radiation and Network Breaking in Polynesian Linguistics

David Bryant*

February 1, 2005

1 Introduction

The exploration and settlement of Polynesia was surely one of the greatest navigational feats in all human history. At a time when Europeans were tentatively edging out into the Mediterranean, Austronesians had colonised half the globe [7]. Even more extraordinary is that, at least according to Sharp [11], all of this exploration was carried out in primitive sailing vessels capable only of coastal navigation. Sharp claimed that the Polynesians had neither the technology nor the skills required for intentional voyages of longer than 200 miles from a coastline. Hawaii, Easter Island and New Zealand were discovered by accident, perhaps when the early sea-farers were blown off course in a storm.

For many, this process of accidental settlement followed by isolation made Polynesia a convenient laboratory for studying cultural, genetic, and linguistic evolution [5]. Polynesian societies were assumed to be more or less isolated from external influences and from each other. Consequently, Polynesia was one place where phyletic (tree-based) models of linguistic and cultural evolution could be applied without hesitation.

However the 'myth of primitive isolation' [14] has been deconstructed. Sharp's theory of accidental discovery has given way to theories of systematic and intentional exploration. Polynesian sailing, navigational, and exploratory techniques were far more sophisticated than had been at first thought [8]. Hawaii, Easter Island, and New Zealand were discovered not by accident, but through planned and systematic exploration. Indeed the initial discoveries were followed by multiple return voyages. Eastern and western Polynesia were linked by extensive trading networks. The Polynesian islands were far from isolated.

Pawley and Green [9] describe the different theories for Polynesian exploration using two *modes* of settlement:

> *Radiation*, where islands are settled and then, for the most part, isolated. This corresponds to *allopatric* or *peripatric* speciation in evolutionary biology, as is well described by a tree based model.

*McGill Centre for Bioinformatics, 3775 University, Montréal, Québec H3A 2B4. bryant@mcb.mcgill.ca

Network breaking, where groups of islands are settled and at first in close contact, but gradually divergences between different islands appear. The initial period of interaction creates a *dialect chain*. This corresponds to *sympatric* speciation in evolutionary biology.

Kirch and Green [8] stress that analyses must incorporate *both* modes. The network breaking model is more suited for closely clustered islands, while the radiation model is more appropriate for remote islands.

The possibility of frequent and widespread linguistic exchanges between different societies poses a difficult challenge for phylogenetic linguistics. If we limit ourselves to tree building methods, parsimony *or* model based, we run the risk of being severely misled. Recent exchanges may make some languages appear to have diverged far more recently than they actually did, and can lead to incorrectly derived histories. Terrell et al. [14] went further to claim that repeated exchanges have effectively wiped out any remaining historical phylogenetic signal in the data.

If the amount of borrowing between languages is minimal then network based methods (e.g. [3]) may be able to reconstruct the history of exchanges. However if the interactions were continuous and widespread it may not actually be *possible*, for any method, to reconstruct the complete history.

This situation is not unique to linguistics. In population genetics, there are generally not enough data to reconstruct the exact phylogenetic history of the genomes being studied, recombination or not (e.g. [12]). However powerful statistical techniques can be employed to make estimates, or test hypotheses, by integrating over *all* possible histories. The fact that we cannot reconstruct the exact history does not prevent us from making inferences about parameters or features of the data.

In this chapter, we do not even try to reconstruct an explicit history for Polynesian languages. Instead we look at a more general question: do the patterns of language similarity support a *radiation* model or a *network breaking* model? We develop a model for language evolution in Polynesia that incorporates accessibility of islands, language innovations, exchanges between islands, and differing island settlement dates. We derive exact and analytic formulae for the language differences expected under this model, and use these results to estimate the rate of language change and inter-island exchange that has occurred. Our analysis demonstrates (with many caveats) the existence of substantial and continual exchanges between islands.

2 The data

2.1 The accessibility matrix

The islands of Polynesia are not equidistant from each other. Some are separated by just a few days' sailing; others, like New Zealand, by several weeks. Distance is not the only factor affecting how easy it is to get from one island to another, though it is the

most important. A large island with high mountains is easier to find (and return to) than a small atoll that is barely above sea level. Irwin [6] incorporated these factors into an *accessibility matrix*. We modify Irwin's matrix by restricting the islands to the 11 for which we have linguistic data, and by normalising the entries so that they sum to one. The accessibility matrix **A** that we use is presented in Table 1.

	1	2	3	4	5	6	7	8	9	10	11
1. Tonga	0.0	.020	.014	.0080	.0098	.0080	.0049	.011	.0040	.0013	.010
2. Samoa	.020	0.0	.012	.010	.0076	.0067	.0085	.0076	.0045	.0022	.011
3. S. Cooks	.014	.012	0.0	.010	.024	.013	.0067	.020	.0062	.0036	.011
4. N. Cooks	.0080	.010	.010	0.0	.010	.0067	.0049	.0094	.0049	.0022	.0045
5. Tahiti	.0098	.0076	.024	.010	0.0	.018	.0089	.044	.0098	.0040	.0089
6. Marquesas	.0080	.0067	.013	.0067	.018	0.0	.0058	.024	.0089	.0045	.0053
7. Hawaii	.0049	.0085	.0067	.0049	.0089	.0058	0.0	.0094	.0036	.0013	.0040
8. Tuamotu	.011	.0076	.020	.0094	.044	.024	.0094	0.0	.015	.0049	.011
9. Mangareva	.0040	.0045	.0062	.0049	.0098	.0089	.0036	.015	0.0	.0062	.0036
10. Easter	.0013	.0022	.0036	.0022	.0040	.0045	.0013	.0049	.0062	0.0	0.0
11. NZ	.010	.011	.011	.0045	.0089	.0053	.0040	.011	.0036	0.0	0.0

Table 1: The accessibility matrix **A**

2.2 Linguistic data

Our Polynesian linguistic data, kindly supplied by Russell Gray, was extracted from the enormous POLLEX database [1]. The data was coded as multi-state *lexical characters*, following [10, 2]. Each character corresponds to a *semantic slot*, or word meaning. We used the 200 semantic slots from the Swadesh list of core vocabulary [13]. The character coding for a particular semantic slot is an assignment of states to each language, with two languages assigned the same state if and only if they possess words that are true cognates for that semantic slot. See [10] and [2] for examples and discussion of this coding.

For each pair of islands we computed the proportion of semantic slots for which the two islands had words that are not cognate. The resulting distance matrix is presented in Table 2. Multi-dimensional scaling (MDS) (using XGVIS[4]) and NeighborNet [3] reveal similar patterns (Figure 1). In both analyses we see a clear east-west division, with Tonga and Samoa well separated from the remaining languages. There is little or no indication of further subdivisions within the eastern Polynesian islands in the MDS, however the NeighborNet supports the division into Tahitic (Tahiti, Tuamotu, NZ, S. Cooks) and Marquesic languages, with Hawaii and Northern Cooks somewhat ambiguous.

	1	2	3	4	5	6	7	8	9	10	11
1. Tonga	0.0	0.2159	0.4395	0.4737	0.4648	0.4523	0.4434	0.4660	0.4208	0.4633	0.4544
2. Samoa	0.2159	0.0	0.3966	0.3848	0.4023	0.3984	0.3954	0.4131	0.3793	0.3943	0.4312
3. S. Cooks	0.4395	0.3966	0.0	0.2132	0.1727	0.2583	0.2298	0.1992	0.2263	0.3163	0.2212
4. N. Cooks	0.4737	0.3848	0.2132	0.0	0.2042	0.2318	0.2533	0.2406	0.2017	0.2439	0.3438
5. Tahiti	0.4648	0.4023	0.1727	0.2042	0.0	0.2381	0.2315	0.1588	0.2316	0.2660	0.2472
6. Marquesas	0.4523	0.3984	0.2583	0.2318	0.2381	0.0	0.2209	0.2298	0.1814	0.2319	0.2843
7. Hawaii	0.4434	0.3954	0.2298	0.2533	0.2315	0.2209	0.0	0.2368	0.2215	0.2657	0.2609
8. Tuamotu	0.4660	0.4131	0.1992	0.2406	0.1588	0.2298	0.2368	0.0	0.2471	0.3017	0.2061
9. Mangareva	0.4208	0.3793	0.2263	0.2017	0.2316	0.1814	0.2215	0.2471	0.0	0.2067	0.3234
10. Easter Island	0.4633	0.3943	0.3163	0.2439	0.2660	0.2319	0.2657	0.3017	0.2067	0.0	0.3891
11. NZ	0.4544	0.4312	0.2212	0.3438	0.2472	0.2843	0.2609	0.2061	0.3234	0.3891	0.

Table 2: The proportion of semantic slots for which the languages possess words that are not cognate

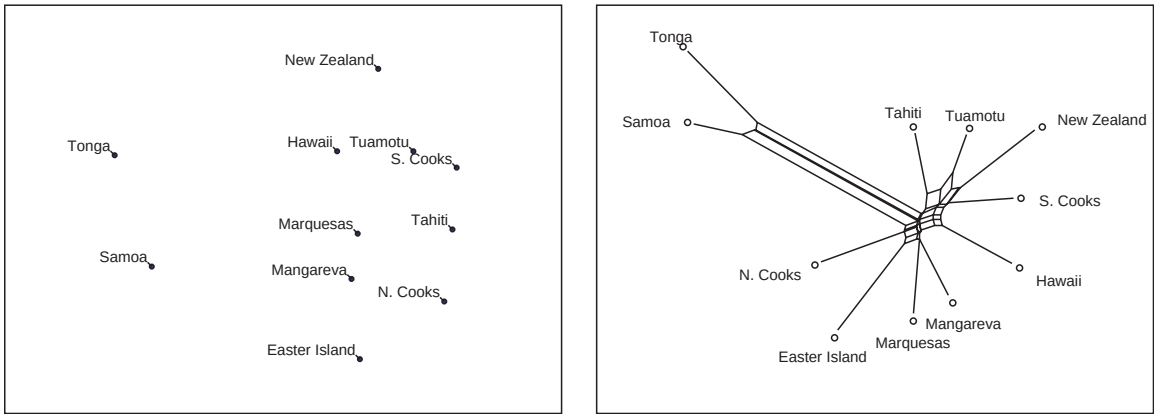

Figure 1: Two representations of the distances between Polynesian languages. Left: Output from a two dimensional multi-dimensional scaling analysis. Right: Neighbor-Net.

2.3 Settlement dates

Figure 2 shows a hypothetical time-line for settlement of the Polynesian islands. The dates were estimated (somewhat roughly) from a summary of archaeological evidence made by Kirch and Green [8]. These dates corresponds to the earliest archaeological evidence for human presence. The date for Tuamotu was extrapolated from the date for the Austral Islands.

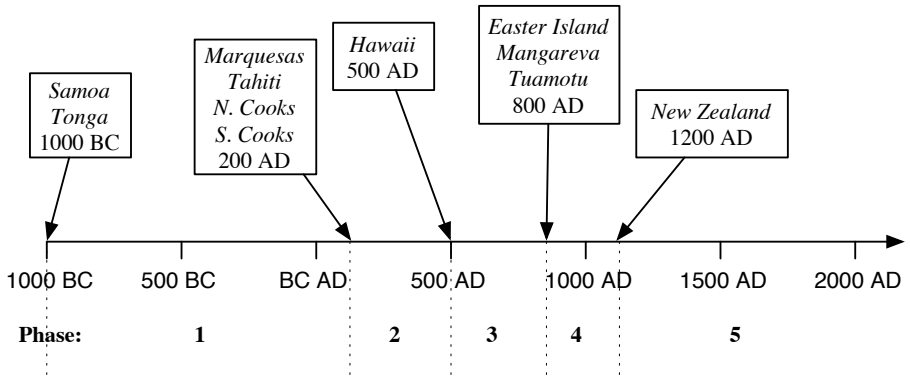

Figure 2: Timeline of Polynesian settlement, giving the date estimates and phases of settlement used in this paper.

3 Constructing a model of Polynesian language evolution

In this section we detail our probabilistic model of language evolution on these eleven Polynesian islands. We are only interested in the patterns of cognate and non-cognate words for the 200 semantic slots. For us, a language history is an assignment of states for every character (semantic slot) for every point of time, for every island settled by that time. At a given time, two islands have the same state for a given slot if and only if, at that time, they possessed words for that slot that were cognate. Time ranges from 0 (1000 BC) to 3004 (the present).

Our model is *Markovian*, in the sense that the probabilities of a specific language assignment at time $t+h$, given the assignment at time t, is independent of the assignments for all times before t. In other words, only the current state of the system affects what happens next.

We model the progressive settlement of different islands by defining five *settlement phases* (see Figure 2). The phases are delimited by the settlement times, and are numbered one to five. The number of settled islands, and therefore the number of islands involved in potential exchanges, changes for every phase.

Our model assumes (implicitly) that settlement of the different islands was a discrete event, with the islands introduced in each phase being settled on the exact dates given

in Figure 2. A more sophisticated model would incorporate variability and estimation error into these settlement times. However this change would make the analysis a lot more complicated, and it is not immediately apparent that we would get a significantly better fit of the model to the data.

Let $S^{(i)}$ denote the set of islands that have been settled by phase i, where i ranges from 1 to 5. Thus $S^{(1)}$ contains Tonga and Samoa only, while $S^{(5)}$ contains all eleven islands.

3.1 Initialisation

At time $t = 0$ Samoa and Tonga are the only settled islands. For each semantic slot we assign the same state to both islands.

3.2 During a phase

There are two kinds of events that can occur during a give phase i: exchanges and innovations. Each occurs according to a *Poisson process*, with a different rate for each different event. We have tried to keep both events as simple as possible. This is appropriate if we are to test null hypotheses regarding the processes of language mutation and borrowing. However in both cases we are clearly modelling highly complex processes with very simple ones. Future versions could apply more sophisticated models.

Exchange events Exchange events from island a to island b in $S^{(i)}$ occur with rate $\tau \mathbf{A}_{ab}$ per semantic slot per unit time with values for $\mathbf{A}_{ab}$ taken from Table 1. When an exchange occurs, a word in a semantic slot for the language on island a is copied into the corresponding slot for island b.

Word innovation events A word innovation occurs at a rate μ per semantic slot per populated island. The word resulting from an innovation is not cognate with any of the words used in any other language.

3.3 Between phases

The borderlines between phases correspond to settlement events. Suppose we are between phase i and phase $i + 1$. Each island in $S^{(i+1)}$ that is not in $S^{(i)}$ is settled from a randomly chosen island in $S^{(i)}$. The probability that $b \in S^{(i+1)}$ is settled from $a \in S^{(i)}$ is

$$\frac{\mathbf{A}_{ab}}{\sum_{x \in S^{(i)}} \mathbf{A}_{xb}}.$$

The language of a newly settled island is copied from the island that settles it.

3.4 Relation to the radiation and network breaking models

We have described a relatively simple stochastic model with two parameters, τ and μ. During each phase, the amount of exchange (borrowing) between different islands is

Parameters	
τ	Rate of exchange (transfer of cognates)
μ	Rate of mutation/innovation
Constants	
$\mathbf{A}$	Accessibility matrix
N	Number of semantic slots
r	Number of possible cognates per semantic slot
$S^{(i)}$	Islands inhabited during phase i.
$\sigma(a, b)$	Probability that island b is settled from island a.

Table 3: Parameters and constants in the basic model.

determined by τ. The probabilities of exchange between two islands correlate with the ease of migrating from one island to another. There would be a lot of exchange between some islands, while remote islands will be relatively isolated. The rate of innovation is determined by μ, and innovations are more likely to persist on remote islands where there is less chance they will be 'corrected'.

If we set $\tau = 0$ then we obtain the radiation model, with no exchange between islands after settlement. Cognate patterns generated according to this model will be completely tree-like. If we want to distinguish between the radiation model and the network breaking model we can study the effect of increasing the parameter τ.

4 Derivation of the expected differences between languages

Having formalised our model, we now study the patterns of linguistic differences that it generates. In this section we derive expectations for the pair-wise differences between sites.

For the remainder of the section we consider a single, arbitrary, semantic slot. The expected number of differences between languages follows directly, irrespective of whether or not the different slots are independent.

Let $C_{ab}(t)$ denote the event that island a and island b have cognate words (that is, the same state) for the semantic slot at time t. Let $c_{ab}(t)$ denote the probability $\mathbb{P}[C_{ab}(t)]$ that $C_{ab}(t)$ holds. If t occurs during the middle of phase i then, by definition $c_{ab}(t) = 0$ whenever either a or b are not in $S^{(i)}$, that is, when a and b have not both been settled by phase i.

We will determine the probabilities $c_{ab}(t)$ by solving a system of ordinary differential equations (ODEs). Let t be a point in time in the middle of one of the settlement phases. Let h be a value that is small enough such that $t + h$ is in the same phase. We will consider the limit of

$$\frac{c_{ab}(t+h) - c_{ab}(t)}{h} \tag{1}$$

as h goes to zero. This equals the derivative $\frac{d}{dt}c_{ab}(t)$.

4.1 Background: 'little oh' $o(h)$ notation

Before proceeding we briefly review the 'little oh' notation $o(h)$. Writing

$$\mathbb{P}[\text{Event } X] = \lambda h + o(h)$$

for some real number λ is identical in meaning to writing

$$\lim_{h\to 0} \frac{\mathbb{P}[\text{Event } X] - \lambda h}{h} = 0.$$

Hence if can find some function $F(t)$ such that

$$c_{ab}(t+h) = c_{ab}(t) + hF(t) + o(h)$$

then we have

$$\begin{aligned} 0 &= \lim_{h\to 0} \frac{c_{ab}(t+h) - c_{ab}t - hF(t)}{h} \\ &= \lim_{h\to 0} \frac{c_{ab}(t+h) - c_{ab}t}{h} - F(t) \\ &= \frac{d}{dt}c_{ab}(t) - F(t). \end{aligned}$$

Also note that if $\mathbb{P}[\text{Event } X_1] = \lambda_1 h + o(h)$ and $\mathbb{P}[\text{Event } X_2] = \lambda_2 h + o(h)$ then

$$\mathbb{P}[\text{Event } X_1] + \mathbb{P}[\text{Event } X_2] = \lambda_1 h + \lambda_2 h + o(h).$$

For any Poisson process with rate λ the number of events occurring in the interval $[t, t+h]$ has a Poisson distribution

$$\mathbb{P}[k \text{ events}] = e^{-\lambda h}\frac{(\lambda h)^k}{k!}.$$

If you expand out $e^{-\lambda h}$ into series form one gets

$$e^{-\lambda h} = 1 - \lambda h + \frac{\lambda^2}{2!}h^2 - \frac{\lambda^3}{3!}h^3 + \cdots$$

The sum of the terms involving $h^2, h^3, \ldots$ is $o(h)$. We therefore get

$$\mathbb{P}[1 \text{ event}] = \lambda h + o(h) \tag{2}$$

$$\mathbb{P}[2 \text{ or more events}] = 0 + o(h). \tag{3}$$

4.2 Setting up the ODE

We now derive a formula for $c_{ab}(t+h)$ in terms of the probabilities at time t. Suppose that t and $t+h$ are both in settlement phase i. As we have just seen, the probability of more than one event occurring between time t and time $t+h$ is $o(h)$. This leaves four possible outcomes.

(i) *Exchange affecting this slot between islands a and b.* An exchange from island a to island b or from b to a affecting this slot occurs with probability $\tau(\mathbf{A}_{ab} + \mathbf{A}_{ba})h + o(h)$. If such a transfer occurs, the probability of $C_{ab}(t+h)$ equals one.

(ii) *Exchange affecting this slot from another settled island x to a or b.* For each settled island $x \in S^{(i)}$ there is an exchange from x to a affecting this slot with probability $\tau\mathbf{A}_{xa}h + o(h)$. If this transfer occurs the probability of $C_{ab}(t+h)$ is $C_{bx}(t)$. The formula for a transfer to b is the same with a and b reversed. Note that we need to sum these probabilities over all $x \in S^{(i)}$ such that $x \neq a, b$.

(iii) *Innovation for this slot on island a or b.* A word innovation in this slot occurs on a or b with probability $2\mu h + o(h)$. If there is a word innovation then the probability $c_{ab}(t+h)$ of a and b having cognate words at time $t+h$ is 0.

(iv) *None of the above* The probability of no transfers affecting a or b at this slot, and no innovations in a or b at this slot is simply 1 minus the probability of all of the above events, or

$$\begin{aligned} & 1 - 2\mu h - \tau(\mathbf{A}_{ab} + \mathbf{A}_{ba})h - \tau \sum_{x \in S^{(i)} - \{a,b\}} (\mathbf{A}_{xa} + \mathbf{A}_{xb})h + o(h) \\ = \;& 1 - 2\mu h - \tau \sum_{x \in S^{(i)}} (\mathbf{A}_{xa} + \mathbf{A}_{xb})h + o(h). \end{aligned}$$

If none of these transfers or innovations do take place, then the probability, $c_{ab}(t+h)$, of identity at time $t+h$ is the same as the probability $c_{ab}(t)$ at time t.

Bringing everything together, we obtain

$$\begin{aligned} c_{ab}(t+h) &= \tau(\mathbf{A}_{ab} + \mathbf{A}_{ba})h + \tau \sum_{x \in S^{(i)} - \{a,b\}} (\mathbf{A}_{xa}c_{bx}(t) + \mathbf{A}_{xb}c_{ax}(t))h \\ &\quad + \left(1 - 2\mu h - \tau \sum_{x \in S^{(i)}} (\mathbf{A}_{xa} + \mathbf{A}_{xb})h\right) c_{ab}(t) + o(h). \end{aligned}$$

Subtracting $c_{ab}(t)$ from both sides, dividing by h and taking the limit as $h \to 0$ gives the system of differential equations

$$\begin{aligned} \frac{d}{dt}c_{ab}(t) &= \tau(\mathbf{A}_{ab} + \mathbf{A}_{ba}) + \tau \sum_{x \in S^{(i)} - \{a,b\}} (\mathbf{A}_{xa}c_{bx}(t) + \mathbf{A}_{xb}c_{ax}(t)) \\ &\quad -2\mu c_{ab}(t) - \tau \sum_{x \in S^{(i)}} (\mathbf{A}_{xa} + \mathbf{A}_{xb})c_{ab}(t). \end{aligned} \tag{4}$$

Here, a, b ranges over all pairs in $S^{(i)}$.

Observe that the right hand side is just a linear combination of the probabilities $c_{xy}(t)$. Hence we can re-express the entire system in one matrix equation. There are $11 \times 10/2 = 55$ different pairs of island a, b. Let $\mathbf{c}(t)$ be the vector with entries indexed by pairs of islands and

$$\mathbf{c}(t)_{ab} = c_{ab}(t).$$

Define the 55×55 matrix $\mathbf{M}^{(i)}$ with rows and columns indexed by pairs of islands and entries given by

$$\mathbf{M}^{(i)}_{ab,cd} = \begin{cases} 0 & \text{if } a,b,c,d \text{ are not all in } S^{(i)}\text{; else:} \\ -2\mu - \tau \sum_{z \in S^{(i)}} (\mathbf{A}_{za} + \mathbf{A}_{zb}) & \text{if } \{a,b\} = \{c,d\}, \\ \tau \mathbf{A}_{bd} & \text{if } a = c; \\ \tau \mathbf{A}_{bc} & \text{if } a = d; \\ \tau \mathbf{A}_{ad} & \text{if } b = c; \\ \tau \mathbf{A}_{ac} & \text{if } b = d; \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{x}^{(i)}$ denote the vector, also indexed by pairs of islands, with

$$\mathbf{x}^{(i)}_{ab} = \begin{cases} \tau(\mathbf{A}_{ab} + \mathbf{A}_{ba}) & \text{if } a,b \in S^{(i)}; \\ 0 & \text{otherwise.} \end{cases}$$

Let t_i be the start of the current settlement phase. The system of equations (4) can be rewritten as

$$\frac{d}{dt}\mathbf{c}(t) = \mathbf{M}^{(i)}\mathbf{c}(t) + \mathbf{x}^{(i)}$$

which has solution

$$\mathbf{c}(t) = e^{\mathbf{M}^{(i)}(t-t_i)}((\mathbf{M}^{(i)})^{-1}\mathbf{x}^{(i)} + \mathbf{c}(t_i)) - (\mathbf{M}^{(i)})^{-1}\mathbf{x}^{(i)}. \tag{5}$$

Equation (5) gives the probabilities of languages having cognate words for a slot given the probabilities at the beginning of a phase. These probabilities can change *between* phases as well, due to the settlement process. Consider the settlement stage between phase i and phase $i+1$. If a is a settled island and b is an island settled in this turn, then the probability that b is settled from a is

$$\frac{\mathbf{A}_{ab}}{\sum_{x \in S^{(i)}} \mathbf{A}_{xb}}$$

which we denote by $\sigma(a,b)$. This leads immediately to formulae for the probability $\mathbf{c}_{ab}$ that languages on islands a and b have cognate words in that slot after settlement:

If a, b both already settled then $\mathbf{c}_{ab}$ is unchanged.

If a is already settled and b is newly settled then for b to have the same cognate as a it must either have been settled from a or from an island with a word that is cognate to the word used in a. Hence

$$\mathbf{c}_{ab} = \sigma(a,b) + \sum_{z \in S^{(i)} - \{a\}} \sigma(z,b)\mathbf{c}_{az}.$$

If a and b are both newly settled then the probability that they have cognate words in this slot equals the probability that they were settled from the same island, or from islands with words that are cognate. Hence in this case

$$\mathbf{c}_{ab} = \sum_{z \in S^{(i)}} \sigma(z,a)\sigma(z,b) + \sum_{y \in S^{(i)}} \sum_{z \in S^{(i)} - \{y\}} \sigma(y,a)\sigma(z,b)\mathbf{c}_{yz}$$

Each of the above cases involves a linear combination of the elements in $\mathbf{c}$ and some constant terms. We can therefore construct a matrix $\mathbf{B}^{(i)}$ and vector $\mathbf{b}^{(i)}$ such that if $\mathbf{c}$ is the vector of probabilities before settlement then

$$\mathbf{B}^{(i)}\mathbf{c} + \mathbf{b}^{(i)}$$

is the vector of probabilities after settlement.

4.3 Solution of the system: expected differences

We now bring everything together and compute the expected proportion of semantic slots at which the current eleven languages have cognate words (given the values for the parameters α, μ). For each phase i we let t_i denote the time just at the beginning of phase i (after settlement). We also define $\mathbf{y}^{(i)} = -(\mathbf{M}^{(i)})^{-1}\mathbf{x}^{(i)}$. Then $\mathbf{y}^{(i)}$ contains the *equilibrium probabilities* for phase i, in the sense that if the system stayed in this phase for $t \to \infty$ then the limiting cognate probabilities would be $\mathbf{y}^{(i)}$. Generally, these are either all one or all zero, depending on the relative rate of innovations and exchanges.

Under our model, the initial conditions are that Samoa and Tonga were the first settled and, in the beginning, they shared the same language. Thus $\mathbf{c}(t_1)_{12} = 1$ and $\mathbf{c}(t_1)_{ab} = 0$ for all other islands a, b.

From equation (5) the cognate probabilities are the end of phase 1 are

$$e^{(t_2-t_1)\mathbf{M}^{(1)}}(\mathbf{c}(t_1) - \mathbf{y}^{(1)}) + \mathbf{y}^{(1)}$$

so the probabilities after settlement, at the beginning of phase 2, are

$$\mathbf{c}(t_2) = \mathbf{B}^{(1)}\left[e^{(t_2-t_1)\mathbf{M}^{(1)}}(\mathbf{c}(t_1) - \mathbf{y}^{(1)}) + \mathbf{y}^{(1)}\right] + \mathbf{b}^{(1)}.$$

In the same way

$$\begin{aligned}
\mathbf{c}(t_3) &= \mathbf{B}^{(2)}\left[e^{(t_3-t_2)\mathbf{M}^{(2)}}(\mathbf{c}(t_2) - \mathbf{y}^{(2)}) + \mathbf{y}^{(2)}\right] + \mathbf{b}^{(2)}, \\
\mathbf{c}(t_4) &= \mathbf{B}^{(3)}\left[e^{(t_4-t_3)\mathbf{M}^{(3)}}(\mathbf{c}(t_3) - \mathbf{y}^{(3)}) + \mathbf{y}^{(3)}\right] + \mathbf{b}^{(3)}, \\
\mathbf{c}(t_5) &= \mathbf{B}^{(4)}\left[e^{(t_5-t_4)\mathbf{M}^{(4)}}(\mathbf{c}(t_4) - \mathbf{y}^{(4)}) + \mathbf{y}^{(4)}\right] + \mathbf{b}^{(4)},
\end{aligned}$$

and the present day probabilities are given by

$$\mathbf{c} = \mathbf{c}(3004) = e^{(3004-t_5)\mathbf{M}^{(5)}}(\mathbf{c}(t_5) - \mathbf{y}^{(5)}) + \mathbf{y}^{(5)}.$$

4.4 Estimating the parameters

For each pair of islands a, b and each semantic slot we have that $\mathbf{c}_{ab}$ is the probability that both islands have words that are cognate. As the expectation of a sum is the sum of expectations, $1 - \mathbf{c}_{ab}$ equals the expected proportion of slots for which islands a and b have words that are not cognate. By comparing these expected differences with the observed differences, we can obtain rough estimates for the rate of innovations μ and

the rate of exchanges τ, as well as test whether our estimates of τ are large enough to reject the radiation model.

The values $\mathbf{c}_{ab}$ depend on μ and τ so we write them as a function $\mathbf{c}_{ab}(\mu, \tau)$. Let δ_{ab} denote the observed proportion of differences between the islands a and b. We minimise the least squares residue

$$\sum_{a<b} (\mathbf{c}_{ab}(\mu, \tau) - \delta_{ab})^2. \tag{6}$$

We used the `fminsearch` routine in MATLAB to find values of μ and τ giving the smallest residue.

5 Results

Our model has two parameters, one determining the rate of innovation and the other determining the rate at which words are exchanged between islands. By fitting the model to the linguistic data, we can obtain estimates for these parameters and identify the extent to which linguistic differences between the Polynesian islands support a network breaking or radiation model.

The residue (eqn. 6) is minimised when the rate of innovation μ is 9.09×10^{-5} and the rate τ of transfers is $\tau = 0.0041$ words exchanged per year. All 100 random starting points converging to the same minimum. The value for μ corresponds to a rate of change of 9% of slots per 1000 years, roughly in concordance with the rate obtained by Swadesh [13]. The value of τ corresponds to about 4 transfers between all islands per slot per thousand years.

Note that if we hold the transfer rate at $\tau = 0$ (no exchanges) as in the radiation model, the optimal value for μ is $\mu = 8.07 \times 10^{-5}$.

To test whether this is a significant amount of transfer, we generated random data under our model with parameters $\tau = 0$ (no exchanges) and $\mu = 8.07 \times 10^{-5}$. The goal was to test whether the high estimates of τ could be obtained randomly even when the true model had no transfers. The experiment was repeated 100 times. For each replicate we generated a simulated language data set and then determine the values of μ and τ that gave the best fit. We found that, out of the 100 runs, only 5% gave estimates for τ that were greater than 0.0041 word exchanges per year. Hence, if our model is correct but there is no migration then we would expect to estimate the amount of migration we did get with probability at most 0.05. We can interpret this as a fairly significant indication that there was indeed non-trivial amounts of migration between the different Polynesian islands.

We stress that these results come with the obligatory warning that our statistical tests are only as good as our model. We have made a great number of simplifying assumptions when setting up our model, and these could well have misled our analysis. The long and arduous process of model validation, and detection of systematic biases, remains.

6 Discussion

We believe that this is the first explicit model for the effect of network breaking on linguistic patterns. There is, understandably, a lot of work remaining, and a lot of model assumptions that need testing. We have identified the most important directions for future research:

1. *Analysis of larger data sets.* In order to reduce the variance in our estimators, and properly compare different modes of language evolution, we need to increase both the number of semantic slots and the number of islands in the analysis. The POLLEX database [1] is a rich source of linguistic data - but this needs to be complemented with accessibility data for a larger set of islands.

2. *Improved models of language change.* Following Ringe et al. [10] we have shoehorned the linguistic data into a simple multi-state, infinite alleles, model. Unfortunately this model does not allow polymorphism in the data (more than one word for a concept), and any polymorphism will bias our estimates for mutation and transition rates.

3. *More sophisticated models for transfers.* There are several factors which should perhaps be incorporated into our model of exchange. For example, the varying population sizes on difference islands would influence the rate of language innovation and the rate of exchanges with neighbouring islands. As well, it is not clear that our naive use of the accessibility matrix really captures the relative ease of exchange between different islands.

4. *More efficient statistical techniques.* We have determined explicit formulae for the expected differences between islands. However to obtain really accurate estimates we also need to model variances and higher level joint probabilities. A full likelihood model would be ideal, possibly within a Bayesian framework.

With the appropriate statistical tools and enlarged data sets we can start testing and refining our model and, in the process, shed light on the features of Polynesian language evolution.

References

[1] B. Biggs. POLLEX: Proto Polynesian lexicon., 1998.

[2] D. Bryant, F Filimon, and R. Gray. *The Evolution of Cultural Diversity: A Phylogenetic Approach*, chapter Untangling our past: languages, trees, splits and networks, pages 69–86. UCL, 2005.

[3] D. Bryant and V. Moulton. Neighbornet: An agglomerative algorithm for the construction of planar phylogenetic networks. *Molecular Biology And Evolution*, 21:255–265, 2004.

[4] Andreas Buja and Deborah F. Swayne. Visualization methodology for multidimensional scaling. *J. Classification*, 19(1):7–43, 2002.

[5] W. H. Goodenough. Oceania and the problem of controls in the study of cultural and human evolution. *Journal of the Polynesian Society*, 66:146–55, 1957.

[6] G. J. Irwin. *The prehistoric exploration and colonisation of the Pacific*. Cambridge University Press, Cambridge, UK, 1992.

[7] M. King. *The Penguin History of New Zealand*. Penguin, Auckland, New Zealand, 2003.

[8] P. Kirch and R. Green. *Hawaiki, Ancestral Polynesia*. Cambridge University Press, Cambridge, UK, 2001.

[9] A. Pawley and R. Green. The Proto-Oceanic language community. *Journal of Pacific History*, 19:123–146, 1984.

[10] D. Ringe, T. Warnow, and A. Taylor. Indoeuropean and computational cladistics. *Trans. Philol. Soc.*, 100:59–129, 2002.

[11] A. Sharp. Ancient voyagers in the Pacific. *Polynesian Society Memoir*, 32, 1956.

[12] M. Stephens and P. Donnelly. Inference in molecular population genetics. *J. R. Statistical Society B*, 62(4):605–655, 2000.

[13] M. Swadesh. Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.*, 96:453–463, 1952.

[14] J.E. Terrell, T. Hunt, and C. Gosden. The dimensions of social life in the Pacific: Human diversity and myth of the primitive isolate. *Current Anthropology*, 38(2):155–195, 1997.