This is a map of the world published in 1707, at the dawn of European Pacific Discovery. (Note the ignorance of the Pacific and the supposition that California was an island!) This map was aware of about half of what we now call Austronesia, the Island Empire settled by people with a common linguistic ancestor. In this talk we will learn a little bit about the history of Austronesian settlement, and see what – if anything – lexical analysis can contribute to our understanding.
Indo-European languages are #1 in the world today, with over 3 billion native speakers, or about 45% of the world population.
But turn back the clock to pre-colonial days and you see these languages covered just Europe and India, as you would expect. The total extent was about 8300 km, from Iceland to Assam.
But in pre-colonial times, the geographic extent for what would come to be called the Austronesian Family of languages was much greater, about 22,000 km, all the way from Madagascar to Rapa Nui (Easter Island). Today Austronesian is #2 in number of languages (1291 per Wikipedia), #5 in number of speakers (over 400 million, or 5.9% of the world’s population) The most widespread language sub-group is called Malayo-Polynesian, and it further divided into Western, Oceanic and a couple of Central subgroups.
That 22,000km extent is enormous. You can’t show it in a single hemisphere…it’s spread over 206 degrees of longitude, and has been for a thousand years or so.
It appears that these languages originated on the island of Taiwan (or Formosa), where 9 out of the 10 sub-families exist, or did exist until recently. All the remaining languages, the ones we’ll be discussing, are in the Malayo-Polynesian subfamily. And because all of these languages are spoken on Islands, and because many of these islands were quite isolated until modern times, the Austronesian Family is particularly interesting to study.
The Technology that Settled Austronesia

Here’s the technology that made it possible for people to spread over such enormous distances: the outrigger canoe and double-canoe, or catamaran. These are such valuable inventions they are in use even today.
And these pacific explorers didn’t venture forth by accident or because they got lost. They were exploring for new islands on which to settle. It was a determined effort, brought about by population pressures and human conflicts on tiny specks of land in a vast sea. The Pacific Rat [rattus exulans] didn’t ride along accidentally, it was brought as a source of food. These rats are now distributed throughout the Pacific. Haplogroup 1 never moved East, only Haplogroup 2 made its way along the New Guinea coast, and only Haplogroup 3 travelled farther east, into Polynesia. This seems to imply that there were two separate eastward migrations, one along the shores of New Guinea, and the other across the wild blue ocean, both originating in an area that is between Indonesia and New Guinea.
Combining archaeological, DNA and linguistic evidence, we can get this picture of how people spread through Austronesia. Polynesia was finally inhabited by about 1000 years ago, as far north as Hawaii, as far south as New Zealand, and as far east as Rapa Nui (Easter Island). Just a few hundred years before the pinnacle of eastward settlement, the western boundary of Austronesia was established with the arrival of Indonesian voyagers in Madagascar.
But the Austronesians seem - to the eye at least – to be completely unrelated to one another. The Polynesians in Rapa Nui, Hawaii and New Zealand all have a similar appearance. The Melanesians, like the Fijian warrior in this picture, look entirely different from the Polynesians. But indigenous Taiwanese look even more different, and the lady from the island of Borneo looks – not surprisingly – southeast Asian. And finally, in Madagascar, most people look like Africans. What ties them together most clearly is just the similarity of their languages.
When you examine the maternal (mitochondrial) and paternal (Y-chromosome) DNA characteristics of Pacific Islanders, you see clear evidence of Founder Effects, where a certain subset of people on one island sailed east to settle a new island, and became the majority of its inhabitants. But notice this: the genetic signal for Polynesians does not show up today in The Philippines or Taiwan. It can be traced back as far as Wallacea only. If it’s true that Austronesian languages originated in Taiwan, it’s a clear example of how genes and languages don’t necessarily travel together through time.
This illustration of the percentages of certain Y-Chromosome Haplotypes found across much of Austronesia makes it clear why the Polynesians look different from the Aboriginal Formosans. Only haplotype O3 (purple) found in Taiwan has survived the trek to Samoa and the Cook Islands. But these Polynesians are dominated by Haplogroup C (orange) which is not present today in Taiwan or the Philippines, but does appear in Borneo. On the other side of the Austronesian world, we have the Malagasy, who have retained Haplogroup O1b (blue-green) from Taiwan and Borneo along with Haplogroup O2a (blue) from Borneo. (More about the Malagasy later).

So the genetics of these people at the antipodes of Austronesia seem quite unrelated, though their languages are not. How do we know their languages actually ARE related?
We can tell this by examining the characteristics of some Austronesian languages for which we have data. Linguists have catalogued and studied these languages in great technical detail for many, many years. But even as amateurs, we can examine the number of similar words with the same meaning in various languages, so-called cognates. I’m going to refer to this process as “Lexical Analysis”.

13
The word for “Woman” seems to be somewhat similar across the realm.
The word for “Fire” seems fairly universal, but with some sound changes.
“Eye” is pretty much the same everywhere you go.
And the number “Three” is pretty uniform across the oceans.
Here is a table of words selected for their uniformity across the western Malayo-Polynesian languages. The words are not identical, they are cognates, words similar enough to one another to imply a causal relationship in their evolution.
As the languages moved from Taiwan to Madagascar, half of the words on this short list were essentially preserved. Almost all were preserved from the Philippines onward.
Cognate Words in Polynesian Languages

<table>
<thead>
<tr>
<th>Direction</th>
<th>Polynesia</th>
<th>Polynesia</th>
<th>Polynesia</th>
<th>Polynesia</th>
<th>Polynesia</th>
<th>Polynesia</th>
<th>Polynesia</th>
<th>Polynesia</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Tongan</td>
<td>Sāmoan</td>
<td>Rarotongan</td>
<td>Tahitian</td>
<td>Māori</td>
<td>Hawai'ian</td>
<td>Rapanui</td>
<td></td>
</tr>
<tr>
<td>sky</td>
<td>/ lanĩi/</td>
<td>/ lanĩi/</td>
<td>/ raŋi/</td>
<td>/ raʔi/</td>
<td>/ raŋi/</td>
<td>/ lani/</td>
<td>/ raŋi/</td>
<td></td>
</tr>
<tr>
<td>woman</td>
<td>/ fefine/</td>
<td>/ fafine/</td>
<td>/ vaʔine/</td>
<td>/ vahine/</td>
<td>/ wahine/</td>
<td>/ wahine/</td>
<td>/ wahine/</td>
<td></td>
</tr>
<tr>
<td>fire</td>
<td>/ aﬁ/</td>
<td>/ aﬁ/</td>
<td>/ aﬁ/</td>
<td>/ auahi/</td>
<td>/ ahi/</td>
<td>/ ahi/</td>
<td>/ ahi/</td>
<td></td>
</tr>
<tr>
<td>tongue</td>
<td>/ ’elelo/</td>
<td>/ alelo/</td>
<td>/ arero/</td>
<td>/ arero/</td>
<td>/ alelo/</td>
<td>/ arero/</td>
<td>/ arero/</td>
<td></td>
</tr>
<tr>
<td>eye</td>
<td>/ mata/</td>
<td>/ mata/</td>
<td>/ mata/</td>
<td>/ mata/</td>
<td>/ mata/</td>
<td>/ mata/</td>
<td>/ mata/</td>
<td></td>
</tr>
<tr>
<td>stone</td>
<td>/ maka/</td>
<td>/ ma’a/</td>
<td>/ pooاةtu/</td>
<td>‘oofa’i/</td>
<td>/ pōhatu/</td>
<td>/ pohaku/</td>
<td>/ ma’ea/</td>
<td></td>
</tr>
<tr>
<td>One</td>
<td>/ taha/</td>
<td>/ tasi/</td>
<td>/ tai/</td>
<td>/ tahi/</td>
<td>/ tahi/</td>
<td>/ kahi/</td>
<td>/ tahi/</td>
<td></td>
</tr>
<tr>
<td>Two</td>
<td>/ ua/</td>
<td>/ iua/</td>
<td>/ rua/</td>
<td>/ piti/</td>
<td>/ rua/</td>
<td>/ lua/</td>
<td>/ rua/</td>
<td></td>
</tr>
<tr>
<td>Three</td>
<td>/ tola/</td>
<td>/ tolu/</td>
<td>/ toru/</td>
<td>/ toru/</td>
<td>/ toru/</td>
<td>/ kolu/</td>
<td>/ toru/</td>
<td></td>
</tr>
<tr>
<td>Four</td>
<td>/ fa/</td>
<td>/ fa/</td>
<td>/ a/</td>
<td>/ maha/</td>
<td>/ wha/</td>
<td>/ ha/</td>
<td>/ ha/</td>
<td></td>
</tr>
</tbody>
</table>

And the batting average is even better across Polynesia, all the way from Tonga to Rapanui. But these are “Cherry-picked” words. I like to call them “Golden words”. But what if we take a larger sample? Will some degree of relationship hold up? And if so, can we make a measurement of how closely these languages are actually related?
A Swadesh List of 200 Selected Words

This is what’s called a Swadesh List. It’s named after Lexical Analysis pioneer Morris Swadesh, who first generated a similar list in 1952. This list attempts to sample a language using words that are thought to be universally understood in all languages. This 200-word version contains a wide range of words that have been found and understood in many languages. It is a popular tool for measuring the similarity between related languages.

There are a lot of problems with these lists that I don’t have time to discuss in this lecture. But basically, they work pretty well. So for now, let’s use these lists – carefully – to see what we can learn about the Austronesian languages.

|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| 1 | all | 21 | cloud | 41 | far | 61 | good | 81 | in | 101 | narrow | 121 | root | 141 | smell | 161 | that | 181 | water |
| 2 | and | 22 | cold | 42 | fat | 62 | grass | 82 | kill | 102 | near | 122 | rope | 142 | smoke | 162 | there | 182 | we |
| 3 | animal | 23 | come | 43 | father | 63 | green | 83 | know | 103 | neck | 123 | rotten | 143 | smooth | 163 | they | 183 | wet |
| 4 | ashes | 24 | count | 44 | fear | 64 | guts | 84 | lake | 104 | new | 124 | rub | 144 | snake | 164 | thick | 184 | what |
| 5 | at | 25 | cut | 45 | feather | 65 | hair | 85 | laugh | 105 | night | 125 | salt | 145 | snow | 165 | thin | 185 | when |
| 6 | back | 26 | day | 46 | few | 66 | hand | 86 | leaf | 106 | nose | 126 | sand | 146 | some | 166 | think | 186 | where |
| 7 | bad | 27 | die | 47 | fight | 67 | he | 87 | left | 107 | not | 127 | say | 147 | spit | 167 | this | 187 | white |
| 8 | bark | 28 | dig | 48 | fire | 68 | head | 88 | leg | 108 | old | 128 | scratch | 148 | split | 168 | you (s) | 188 | who |
| 9 | because | 29 | dirty | 49 | fish | 69 | hear | 89 | lie | 109 | one | 129 | sea | 149 | squeeze | 169 | three | 189 | wide |
| 10 | belly | 30 | dog | 50 | five | 70 | heart | 90 | live | 110 | other | 130 | see | 150 | stab | 170 | throw | 190 | wife |
| 11 | big | 31 | drink | 51 | float | 71 | heavy | 91 | liver | 111 | person | 131 | seed | 151 | stand | 171 | tie | 191 | wind |
| 12 | bird | 32 | dry | 52 | flow | 72 | here | 92 | long | 112 | play | 132 | sew | 152 | star | 172 | tongue | 192 | wing |
| 13 | bite | 33 | dull | 53 | flower | 73 | hit | 93 | louse | 113 | pull | 133 | sharp | 153 | stick | 173 | tooth | 193 | wipe |
| 14 | black | 34 | dust | 54 | fly | 74 | hold | 94 | man | 114 | push | 134 | short | 154 | stone | 174 | tree | 194 | with |
| 15 | blood | 35 | ear | 55 | fog | 75 | how | 95 | many | 115 | rain | 135 | sing | 155 | straight | 175 | turn | 195 | woman |
| 16 | blow | 36 | earth | 56 | foot | 76 | hunt | 96 | meat | 116 | red | 136 | sit | 156 | suck | 176 | two | 196 | woods |
| 17 | bone | 37 | eat | 57 | four | 77 | husband | 97 | mother | 117 | right | 137 | skin | 157 | sun | 177 | vomit | 197 | worm |
| 18 | breathe | 38 | egg | 58 | freeze | 78 | I | 98 | mountain | 118 | right (hand) | 138 | sky | 158 | swell | 178 | walk | 198 | you (pl) |
| 19 | burn | 39 | eye | 59 | fruit | 79 | ice | 99 | mouth | 119 | river | 139 | sleep | 159 | swim | 179 | warm | 199 | year |
| 20 | child | 40 | fall | 60 | give | 80 | if | 100 | name | 120 | road | 140 | small | 160 | tail | 180 | wash | 200 | yellow |
Lexical Distance: A way of Specifying Language Relationships

Lexical Distance =

$$1 - \frac{\text{Number of Cognate Differences Between 2 Languages or Dialects}}{\text{Number of Cognate Sets in All Languages Being Compared}}$$

e.g. 1 - (50 cognates / 200 sets) => Lexical Distance = 1 – 0.25 = 0.75

Lexical Distance = 1 means 0% Cognacy of Words from Swadesh List
Lexical Distance = 0 means 100% Cognacy of Words from Swadesh List

But first, let’s take a moment to define a term: Lexical Distance. By counting the number of cognates between all the words on Swadesh Lists for two languages, we can derive a useful number called “Lexical Distance” between them. For example, if two dialects of languages share 50 words in a list of 200, the Lexical Distance is 1 minus 50 over 200 equals 0.75. Complete matching is a lexical distance of zero; no matching is a lexical distance of one. Defining language differences as distances is useful because it is both intuitive and amenable to technical analysis, as we will see later on.
Lexical Distance = 1 means 0% Cognacy of Words from Swadesh List

To get a feel for what lexical distance means to someone listening to a language, let’s use some familiar languages. About half the English words on the list are cognate with German or Dutch, leading to lexical distance of about 0.5.

About three quarters of Dutch words are cognate with German, leading to a lexical distance of about 0.25. We’ll use this diagram as a yardstick for visualizing the Lexical Distances between other languages. Realize that none of these languages are mutually intelligible. There are a lot more words and some important differences in syntax. But this Lexical Distance number can still be a rough indicator of relatedness. Notice that you can’t simply add two lexical distances to get the resultant distance to a third language.
Let’s apply Lexical Distance Analysis to some Austronesian words for which I have data.
Today, a remarkable amount of data is publicly-available on Austronesian languages. The results of decades of field work by many researchers has been compiled by Professor Richard Blust and others at the University of Auckland and University of Hawaii. In this example, we see data on the Samoan language as spoken today.
I combined their data for major Polynesian words into a table for word-by-word comparison. Notice that the entries are multi-valued in many cases, because the researchers obtained synonyms for many of the words.
Here is a portion of those cognate sets. Horizontal rows highlighted in yellow are sets of cognate words, sometimes more than one word per meaning. For example, the meaning “dirty” (no. 10) has two cognate sets: “repo” and “hava”. The flow of words can be visualized. “Lima” (no. 1) survives with only slight modifications throughout Eastern Polynesia. “Hava” (10) and “koeloe/kokoma” (16) are unique to Marquesan and Rapanui. “Tapono” is found only in 18th century and modern Tahitian. “uu” and similar forms signify “breast” everywhere except in modern Tahitian, where it has been replaced by the word “titti”, probably borrowed from English sailors. Continued

The Modern Tahitian word “mana’o” is cognate to several other languages, but was missing from the 18th century data, even though it, or something very similar to it, had to have been present in the 18th century. So I added it to the table (21, in red) to further reconstruct the 18th century Tahitian language. By analyzing the cognate connections between these columns, we can calculate a measure of Lexical Distance between languages. And by comparing the forms of each cognate set’s members across languages, we can follow the language evolution.
Lexical Distance Matrix for Selected Polynesian and Melanesian Languages

<table>
<thead>
<tr>
<th></th>
<th>KIRIBATI</th>
<th>KAPINGAMARANGI</th>
<th>FIJIAN</th>
<th>FUTUNA</th>
<th>TONGAN</th>
<th>SAMOAN</th>
<th>MARQUESAN</th>
<th>TAHITIAN</th>
<th>18c TAHITIAN</th>
<th>HAWAIIAN</th>
<th>RAPANUI</th>
<th>RARATONGAN</th>
<th>MAORI</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIRIBATI</td>
<td>0.93</td>
<td>0.93</td>
<td>0.94</td>
<td>0.91</td>
<td>0.92</td>
<td>0.90</td>
<td>0.91</td>
<td>0.95</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>KAPINGAMARANGI</td>
<td>0.93</td>
<td>0.88</td>
<td>0.88</td>
<td>0.75</td>
<td>0.76</td>
<td>0.71</td>
<td>0.73</td>
<td>0.78</td>
<td>0.74</td>
<td>0.75</td>
<td>0.74</td>
<td>0.70</td>
<td>0.74</td>
</tr>
<tr>
<td>FIJIAN</td>
<td>0.94</td>
<td>0.88</td>
<td>0.86</td>
<td>0.86</td>
<td>0.85</td>
<td>0.89</td>
<td>0.91</td>
<td>0.88</td>
<td>0.88</td>
<td>0.89</td>
<td>0.89</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>FUTUNA-ANIWA</td>
<td>0.91</td>
<td>0.75</td>
<td>0.86</td>
<td>0.75</td>
<td>0.68</td>
<td>0.71</td>
<td>0.79</td>
<td>0.73</td>
<td>0.75</td>
<td>0.72</td>
<td>0.68</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>TONGAN</td>
<td>0.92</td>
<td>0.76</td>
<td>0.86</td>
<td>0.75</td>
<td>0.66</td>
<td>0.75</td>
<td>0.78</td>
<td>0.73</td>
<td>0.76</td>
<td>0.73</td>
<td>0.71</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>SAMOAN</td>
<td>0.90</td>
<td>0.71</td>
<td>0.85</td>
<td>0.68</td>
<td>0.65</td>
<td>0.72</td>
<td>0.64</td>
<td>0.67</td>
<td>0.65</td>
<td>0.65</td>
<td>0.60</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>MARQUESAN</td>
<td>0.91</td>
<td>0.73</td>
<td>0.89</td>
<td>0.71</td>
<td>0.75</td>
<td>0.65</td>
<td>0.68</td>
<td>0.63</td>
<td>0.61</td>
<td>0.62</td>
<td>0.55</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>TAHITIAN</td>
<td>0.95</td>
<td>0.78</td>
<td>0.91</td>
<td>0.79</td>
<td>0.78</td>
<td>0.72</td>
<td>0.68</td>
<td>0.52</td>
<td>0.67</td>
<td>0.69</td>
<td>0.62</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>18c TAHITIAN</td>
<td>0.92</td>
<td>0.74</td>
<td>0.88</td>
<td>0.73</td>
<td>0.73</td>
<td>0.64</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.64</td>
<td>0.57</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>HAWAIIAN</td>
<td>0.92</td>
<td>0.75</td>
<td>0.88</td>
<td>0.75</td>
<td>0.76</td>
<td>0.67</td>
<td>0.61</td>
<td>0.67</td>
<td>0.63</td>
<td>0.63</td>
<td>0.61</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>RAPANUI</td>
<td>0.92</td>
<td>0.74</td>
<td>0.89</td>
<td>0.72</td>
<td>0.73</td>
<td>0.65</td>
<td>0.62</td>
<td>0.69</td>
<td>0.64</td>
<td>0.63</td>
<td>0.58</td>
<td>0.63</td>
<td>0.58</td>
</tr>
<tr>
<td>RARATONGAN</td>
<td>0.91</td>
<td>0.70</td>
<td>0.86</td>
<td>0.68</td>
<td>0.71</td>
<td>0.60</td>
<td>0.55</td>
<td>0.62</td>
<td>0.57</td>
<td>0.61</td>
<td>0.58</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>MAORI</td>
<td>0.92</td>
<td>0.74</td>
<td>0.88</td>
<td>0.72</td>
<td>0.74</td>
<td>0.65</td>
<td>0.63</td>
<td>0.68</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.56</td>
<td>0.56</td>
</tr>
</tbody>
</table>

355 Cognate Sets

The lexical distances between languages computed from the preceding table of 13 languages and 206 meanings are organized in this 13 by 13 matrix.
These languages are found mostly within the so-called “Polynesian Triangle”, but there are a couple of what are called “Polynesian Outliers” included (Futuna and Kapingamarangi). The languages of Fiji and Kiribati, which belong to the Melanesian family, are included for comparison.
Our brains are not capable of visualizing all the data in the 13 full dimensions of the matrix. So we turn to a technique called Multidimensional Scaling (MDS), which allows us to view an approximate representation of this 13-dimensional distance matrix in three dimensions. The two Melanesian languages from Fiji and Kiribati stand far from the Polynesian languages. Notice that the Polynesian outliers Futuna and Kapingamarangi along with Tongan stand apart from the main group of eastern Polynesian languages. And Samoan is sort of halfway in between the outliers and the Eastern Polynesian languages.
Let’s now focus on this group of major Polynesian Languages. They are all fairly closely related. Average Lexical Distance between any language and any other in this region is just a little bit more than English to German. Can this tell us anything about the settlement of Polynesia?
Polynesian Language Relationships per Multi-dimensional Scaling of Lexical Distance Matrix

Notice that four branches appear to emerge from this 3D MDS plot: Samoan, Tahitian, Rarotongan/Maori, and Marquesan/Hawaiian/Rapanui. This grouping gives us a first-pass estimate as to how these languages may have descended from a common ancestral tongue.
When we plot the lexical distances of three of these groups relative to the modern Samoan language, we see there was a very large change in Tahitian between the 18th century and the present. It increased from 0.64 to 0.70, the largest LD to Samoan. Raratongan and Maori show similar changes in lexical distance to Samoan, 0.59 for Raratongan to 0.64 for Maori. And it’s the same for Hawaiian and Rapa Nui compared to Marquesan.

Also notice that 18th century Tahitian is closer to Samoan than is Modern Tahitian, as you would expect since Modern Tahitian is descended from 18th century Tahitian. So Maori is farther away than Raratongan; Hawaiian and Rapanui are farther away than Marquesan. This must imply an order of descent within these groups.
One of the clear implications of the previous plots is that Hawaii and Rapa Nui were settled from The Marquesas. But there is an alternate theory for the settlement of Hawaii and Rapa Nui, one that assumes Tahiti was the source of settlement. The lexical distance numbers show that Marquesan is slightly closer to these two final settlements than it Tahitian. But this is not very compelling evidence by itself. The MDS evidence is fairly convincing, but not ironclad. Can we actually learn anything more certain from our data about the source of these settlements?
Yes we can! But to do that we need to dig into the details of the cognate words. It turns out that there are 6 cognate sets whose only members are Marquesan and Hawaiian, and 5 sets whose only members are Marquesan and Rapa Nui. When we look for these rare two-language cognate sets between Tahitian and Hawaiian and Tahitian and Rapa Nui, we come up empty. This is pretty strong evidence that words coined in the Marquesas, or which survived from the original Prototype language only in the Marquesas, were propagated to Hawaii and Rapa Nui. There is no evidence that such a thing happened from Tahiti. So our cognate-counting detective work tends to support the standard theory of migration.
While we are talking about Tahiti, and before we do any more analysis, let’s take a look at the observations of a young amateur linguist upon first contact with the people of Tahiti. This young Scotsman was named Sidney Parkinson. He journeyed to the South Pacific with Captain Cook on his first voyage, working as an illustrator for the soon-to-be-famous naturalist Joseph Banks. Parkinson collected many Tahitian words, of which 124 appear in the Austronesian Basic Vocabulary Database. Because of Parkinson’s linguistic observations, and from other sources, linguists have been able to piece together an estimate of what Tahitian was like when spoken in the 18th century, before it was much influenced by Europeans. The language has changed substantially since that time.
Parkinson’s Observations on the Tahitian Language

"REMARKS on the Otaheitean Language.
The language is very soft, having a great number of vowels, diphthongs, and triphthongs. Every word, almost, begins with a vowel, which they most commonly drop.
The natives could not repeat, after us, the sounds of the letters, Q, X, and Z, without great difficulty; G, K, and S, they could not pronounce at all. Many of the names of the people of our ship having the G, K, or S, in them, they could not approach nearer the sound of them than as follows:

Toote, for Cook …
Tolano — Solander …
Treene — Green …
Patine — Parkinson

"They have various sounds peculiar to themselves, which none of us could Imitate; some of them they pronounced like B and L mingled together; others between B and P, and T and D. Some like Bh, Lh, and Dh."
Evolution of the Basic Tahitian Vocabulary over ~240 Years
18th Century Tahitian –vs- Modern Tahitian

...About a Quarter of the Words Have Changed

<table>
<thead>
<tr>
<th>% of Comparisons:</th>
<th>Identical</th>
<th>Evolved</th>
<th>Innovated</th>
<th>Synonym</th>
<th>Abandoned</th>
<th>Borrowed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>58%</td>
<td>16%</td>
<td>19%</td>
<td>1%</td>
<td>5%</td>
<td>2%</td>
</tr>
</tbody>
</table>

Evolved: teiaha >> teimaha
Innovated: mata’u >> ri’ari’a
Synonym: “Salt” miti >> “Sea” miti
Abandoned: tuumuu >>______
Borrowed: “Father” metua taane >> paapaa

Analyzed from: The Austronesian Basic Vocabulary Database, 2014

Because we have this observation of 18th century Tahitian (then an unwritten language), we can see exactly how the words on our list have changed over the years. This clearly illustrates an important point: languages evolve in situ, not just in a settlement relocation. Over time, the lexical distance in the place of origin will approach ONE, whether spoken by settlers on a new island or those who remained behind.
It is possible to arrange languages in something called a Phylogenic Tree. This is similar to trees generated in genetics, or as an attempt to indicate the relationships between various species of plants or animals. But there are some real pitfalls to doing this, and it’s easy to be misled by what seems to be a scientifically-derived diagram of relationships. This tree is the effort by Greenhill, Blust and Gray (the same people who compiled the ABVB list). Let’s call this the GBG tree and take a close look at the Polynesian part of it.
Here I have simplified the GBG tree to show the portion of their tree applicable to our list of major Polynesian languages. This tree is synthesized by a computer based on lexical distances, and seems to imply that Raratongan was the progenitor of Tahitian, Maori and Hawaiian. (Numbers are Lexical Distances from Samoan).
But as you can see, this sequence makes almost no sense. According to it, settlers would have to have backtracked from the Marquesas to Raratonga before settlement could have proceeded. And we’ve already seen that settlement of Hawaii from Tahiti is extremely unlikely. This is surely not the correct settlement pattern. In Eastern Polynesia, the GBG tree fails the reality check.
Here is a tree I put together based on the “rough draft” from the Multidimensional Scaling Analysis 3D plot. This tree is based on human judgment, not automated calculation. It has three major branches: Raratongan, Marquesan, and Tahitian. The percentages of cognates surviving from each previous node are shown.
Here is the same tree, but with those percentages of surviving cognates converted to Lexical Distances. Each distance in this case is NOT from Samoan, but from reconstructed prototype languages. These prototypes were constructed by summing all the cognates of the presumed descendant languages and eliminating overlaps. Blue is the proto language for Raratongan and Maori; green is for Marquesan, Hawaiian and Rapanui; brown is the grand prototype for all the languages.
When we generated a new matrix based on the tree with a grand prototype language (red) and sub-prototype languages, we get a very clean, intuitive 3D MDS plot of the lexical distances. The radial distances from the prototype language in the center (red) are proportional to the lexical distances from each language to the prototype. What is NOT shown in this plot are actual lexical distances between modern languages, only the distances each of them is from the prototype.
The most likely physical locations for the speakers of these prototype languages were: Samoa (brown); Raratonga (blue); Marquesas (green).
Summary

• Austronesia is an “Empire” of Language Similarity Only
• Original Human Settlement of Eastern Polynesia Occurred between 3200 and 1000 years ago
• Careful Lexical Analysis shows Polynesian Language Groupings are Consistent with the Standard Theory of Population Expansion
• Specifically: Pioneers emerged from Samoa to inhabit Raratonga, Tahiti, and The Marquesas. Later Voyages brought People to Hawaii, Rapanui, and New Zealand.
References:

Kirch, P V. and R. C. Green, “Hawaiki, Ancestral Polynesia” (2001)
Diamond, J., “Guns, Germs and Steel” (1997)

www.roryvantuyl.com/Linguistics.html
Polynesia is only part of Austronesia. Let’s shift gears now and look at the Western Malayo-Polynesian half of the picture. Specifically, let’s look at the westernmost island in Austronesia: Madagascar, and its people, the Malagasy. When people brought the first permanent settlement to Madagascar about 1200-1400 years ago, they brought with them a culture, their genes, and of course an Austronesian language.
These are Malagasy people from 16 of the 20 or so ethnic groups of Madagascar. If you were transported to Madagascar without knowledge of where you were going, you’d surely think you had landed in Africa once you saw the people. But if you were familiar with African language sounds, you would be totally mystified by what you heard: these people may look African, but they sure don’t sound African! And you wouldn’t be the first person to be confused by this situation.
In 1603, Dutch Merchant Frederick Houtman noticed the Malagasy natives spoke a language “...very similar to Malay.”
And in 1613, Portuguese Jesuit Fr. Luis Mariano wrote that he noticed the similarity of Malagasy speech to the languages of Southeast Asia.

He traveled up and down the coasts of Madagascar noting:

“their language...is the same throughout the island...the natives of the South and North understand each other with ease.”
So Europeans were puzzled from the earliest days as to why people who looked like This (left)… would talk like people who looked like This (right)?
The eminent 19th century Malagasy scholar Alfred Grandidier called this “Le plus belle énigme du monde” – “The most beautiful mystery in the world.”
The answer came clear to Europeans in 1777, when French merchant-explorer Nicholas Mayeur ventured into the central highlands and for the first time discovered the Merina People.

He wrote...

"In the interior of this great island entirely surrounded by savage peoples there is more enlightenment, more industry and a more active administration than on the coasts where the inhabitants are in constant relations with foreigners."
Here is the Imerina Kingdom in the highlands of Madagascar. It was unified in the 18th century by the great Merina king Andrianampoinimerina. Many Merina - especially the upper classes - don’t look like they came from Africa at all. Could they have come from the Malay Peninsula as Houtman hypothesized? Or from SE Asia as Mariano supposed? And did they bring the Malagasy language with them?
Here’s what the most recent DNA study tells us: Both the Merina and the coastal dwellers from the southern and eastern sections share both African and Oceanic heritage. In fact, the female line (mtDNA) is quite similar for both the Merina and the coastal dwellers: about 60% Oceanic, 40% African. In the male line, the Merina are about 50% African, while the coast dwellers are about three-quarters African.
Madagascar was Peopled from Borneo and Southern Africa
Permanent Settlement circa 600–800 AD

One theory holds that the Indonesians sailed from Borneo to mainland Africa, where they hooked up with some coastal Africans, and they then migrated together to Madagascar. Given the sailing expertise of the Pacific Austronesians, it’s no mystery how they could have sailed to Madagascar from Borneo. The big question is why wasn’t this place settled much earlier from Africa? Well, Southern Africa wasn’t settled by the Bantu until about 1500 years ago, about the same time the Indonesians sailed on the scene. And the Bantu did not at that time have the sailing technology of the Indonesians. So hitching a ride from expert Indonesian sailors would have been an effective way for them to get to Madagascar.
And as you’ve probably figured out, the Malagasy language comes from Indonesia: probably the island of Borneo. The most closely related language is that of the Ma’anyan Dayak people. But the Ma’anyan are land-bound and no longer go to sea. So the actual sailors who set off for the west may have been ancestors of the Ma’anyan who lived near the coast.
None of the modern Indonesian languages that contain words similar to Malagasy are particularly closely related to it. We see that Ma’anyan and Malagasy are not nearly as close as English to German or Latin to Italian.
All the data I used in preparing the Malagasy portion of this presentation was generated from field research directed by Prof. Maurizio Serva, an Italian physicist and expert on Madagascar and its languages. He compiled his data into 200-word Swadesh lists of 23 Malagasy dialects, arranged by location. And he pioneered the use of technical analysis of this data.
Serva and his collaborators spoke to Malagasy people from 23 different towns and 20 different tribal groups and assembled these lists. What is really good about these lists is their identification of the LOCATION each speaker came from. This turns out to be more important than tribal affiliation when it comes to language. Notice that I have marked in pink the words on this list I judge to be COGNATE to Merina Malagasy. Those highlighted in blue are cognate to Ma’anyan as well as Merina.
Here is a diagram of the lexical distances between Merina Malagasy and some Indonesian languages for which Swadesh Lists are available. Although Ma’anyan is the closest to Merina in terms of lexical distance, it is quite a bit farther away than German is from English, for example. Nevertheless, it is significantly closer to Merina than is Malay or any of the other Indonesian languages we’ve studied. Does this mean that the people who sailed from Indonesia to Madagascar were only Ma’anyan speakers? Or that Ma’anyan and Malay speakers all got in a boat together and sailed to Madagascar? Probably Not!
More likely, this is what happened: The Indonesian languages and Merina Malagasy share a common ancient ancestral language. Sub-prototype languages evolved from this ancestral language by the time people sailed for Madagascar. Malagasy settlers and the forebears of the Malay and Ma’anyan speakers all spoke different languages, but these languages were much closer than they are today. These ancient languages diverged over time. Today’s Merina Malagasy is a direct descendant of one of those ancient languages, Ma’anyan is a descendant of another. Merina is not directly descended from Ma’anyan.
But it is nevertheless clear, when you look at the list of all 59 cognates to Merina (highlighted in yellow), that Ma’anyan is the closest relative to Merina Malagasy. There are 29 “Golden Words” on the list. They have cognates to 3 or more Indonesian languages. And there are 8 words on the list that are cognate to Merina from only languages other than Ma’anyan. This is further evidence that the true source language was NOT Ma’anyan, but rather a common ancestral tongue.
So Language analysis and DNA data confirm this theory: Many Ancestors of the Malagasy People came from Indonesia.

Though Derived from Indonesian Language, Merina Malagsy is farther removed from Indonesian than English is from German. Malagasy is unintelligible to Indonesians today.

But what about the various dialects of Malagasy spoken today? Can people throughout the island communicate well with speakers of other dialects?
How Well Do the Malagasy Communicate?

Here’s what foreign auditors have said in historical times:

(1613) “THEIR LANGUAGE...IS THE SAME THROUGHOUT THE ISLAND...THE NATIVES OF THE SOUTH AND NORTH UNDERSTAND EACH OTHER WITH EASE.” --Fr. Luis Mariano

(1777) “...I WAS UNDERSTOOD EVERYWHERE. HOWEVER, I RECOGNIZED A DIFFERENCE IN THEIR WAY OF [PRONOUNCING] CERTAIN WORDS FROM ONE PROVINCE TO ANOTHER.” --Nicolas Mayeur
“Based on my own experience of staying in a non-Merina region, I feel comfortable to claim that if two speakers from different regions distant from each other speak to each other, they typically have problems communicating if they only use their own speech varieties. However, in an actual situation, such speakers negotiate with words and expressions they know of other varieties, eventually establishing a form of communication.”
--Ritsuko Kikusawa, National Museum of Ethnology, Japan

“Dialects from close regions are usually perceived as being similar by Malagasy people while distant dialects usually have a low degree of mutual intelligibility. Most of the people are able to understand the Merina dialect, which is the official language, but outside of the Imerina region only cultivated people are able to speak it.”
--Maurizio Serva, Universitá dell’Aquila, Italy
Here is the difference between three Malagasy dialects as compared to our English/German/Dutch reference. As we see, these inter-dialect distances are comparable to the German/Dutch distance.
The average lexical distance between Merina and the other Malagasy dialects is 0.26±0.05, slightly more than the distance from Dutch to German. This gives us some idea of how different these dialects are today. And this doesn’t apply only to the Merina dialect. The average Lexical Distance between any two dialects is 0.28.
As we have seen, the distance between Merina and all other dialects is 0.26, similar to the distance from German to Dutch. But the Merina and Betsileo dialects are extremely close because the Betsileo, another plateau tribe, were conquered and subjugated by the Merina in the 18th century. Conquerors can impose speech on the conquered.
Lexical Distance Matrix for 23 Malagasy Dialects

1000x Lexical Distance using 372 cognate sets denominator

|     | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  | 13  | 14  | 15  | 16  | 17  | 18  | 19  | 20  | 21  | 22  | 23  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1   | 546 | 535 | 543 | 567 | 581 | 575 | 570 | 591 | 597 | 599 | 621 | 583 | 589 | 602 | 608 | 597 | 610 | 629 | 648 | 640 | 629 |
| 2   | 546 | 556 | 556 | 562 | 575 | 567 | 565 | 581 | 597 | 586 | 583 | 605 | 570 | 565 | 583 | 597 | 607 | 618 | 640 | 624 | 643 | 634 |
| 3   | 535 | 556 | 508 | 556 | 567 | 575 | 565 | 581 | 597 | 602 | 605 | 613 | 634 | 594 | 599 | 599 | 621 | 618 | 640 | 654 | 634 |
| 4   | 543 | 556 | 508 | 554 | 573 | 567 | 556 | 591 | 594 | 591 | 605 | 629 | 597 | 602 | 624 | 616 | 599 | 618 | 640 | 634 | 677 |
| 5   | 567 | 556 | 554 | 527 | 554 | 551 | 597 | 599 | 599 | 599 | 624 | 608 | 597 | 613 | 648 | 634 | 637 | 602 | 586 | 583 | 660 |
| 6   | 581 | 575 | 567 | 573 | 527 | 559 | 559 | 602 | 605 | 602 | 581 | 613 | 559 | 562 | 591 | 605 | 583 | 608 | 621 | 640 | 637 |
| 7   | 575 | 567 | 578 | 567 | 554 | 559 | 524 | 570 | 583 | 570 | 565 | 597 | 535 | 551 | 567 | 589 | 573 | 597 | 608 | 642 | 632 |
| 8   | 570 | 565 | 556 | 556 | 551 | 559 | 524 | 556 | 565 | 562 | 556 | 591 | 546 | 554 | 565 | 583 | 562 | 597 | 609 | 642 | 632 |
| 9   | 591 | 581 | 597 | 591 | 597 | 602 | 570 | 556 | 524 | 548 | 575 | 605 | 562 | 567 | 581 | 586 | 594 | 619 | 640 | 651 | 626 |
| 10  | 591 | 597 | 602 | 594 | 599 | 605 | 583 | 565 | 524 | 551 | 551 | 583 | 570 | 573 | 570 | 597 | 594 | 610 | 602 | 640 | 651 |
| 11  | 597 | 586 | 605 | 591 | 599 | 602 | 570 | 562 | 548 | 551 | 556 | 575 | 556 | 565 | 578 | 586 | 602 | 594 | 642 | 645 | 648 |
| 12  | 599 | 583 | 613 | 605 | 599 | 581 | 565 | 565 | 575 | 551 | 556 | 538 | 562 | 554 | 543 | 575 | 559 | 599 | 624 | 637 | 624 |
| 13  | 621 | 605 | 634 | 629 | 624 | 613 | 597 | 591 | 605 | 583 | 575 | 538 | 575 | 562 | 551 | 562 | 586 | 608 | 613 | 637 | 642 |
| 14  | 583 | 570 | 594 | 597 | 583 | 559 | 535 | 546 | 562 | 570 | 556 | 562 | 575 | 484 | 527 | 567 | 562 | 594 | 609 | 637 | 640 |
| 15  | 589 | 565 | 599 | 602 | 591 | 562 | 551 | 554 | 567 | 573 | 551 | 554 | 562 | 484 | 516 | 559 | 559 | 564 | 598 | 634 | 626 |
| 16  | 602 | 583 | 621 | 624 | 613 | 591 | 567 | 565 | 581 | 570 | 565 | 543 | 551 | 527 | 516 | 567 | 573 | 599 | 608 | 637 | 640 |
| 17  | 608 | 597 | 618 | 616 | 618 | 605 | 589 | 583 | 586 | 597 | 578 | 575 | 562 | 567 | 596 | 619 | 530 | 589 | 578 | 583 | 583 |
| 18  | 597 | 578 | 597 | 599 | 597 | 583 | 573 | 562 | 594 | 594 | 586 | 599 | 586 | 586 | 586 | 602 | 573 | 578 | 578 | 581 | 586 |
| 19  | 610 | 613 | 621 | 618 | 618 | 605 | 589 | 583 | 586 | 597 | 578 | 602 | 594 | 589 | 599 | 535 | 573 | 530 | 589 | 578 | 583 |
| 20  | 629 | 618 | 618 | 624 | 621 | 621 | 608 | 591 | 597 | 602 | 594 | 599 | 613 | 599 | 599 | 608 | 595 | 578 | 530 | 581 | 583 |
| 21  | 648 | 640 | 640 | 634 | 645 | 640 | 642 | 624 | 640 | 640 | 642 | 624 | 637 | 637 | 634 | 637 | 602 | 578 | 589 | 581 | 527 |
| 22  | 640 | 624 | 634 | 637 | 637 | 632 | 624 | 651 | 651 | 645 | 637 | 648 | 624 | 626 | 640 | 640 | 602 | 581 | 578 | 583 | 527 |
| 23  | 629 | 634 | 634 | 640 | 634 | 637 | 629 | 618 | 626 | 637 | 648 | 642 | 624 | 640 | 632 | 620 | 586 | 583 | 594 | 527 | 530 |

372 Cognate Sets List

We can further understand our Malagasy language data by carefully analyzing lexical distance data.

I generated this matrix using a table of 372 cognate sets derived from 198 word meanings for 23 dialects of the Malagasy language. Each entry is 1000x the lexical distance between the language in its row and the language in its column. The locations and number assigned to each dialect are shown on the left.
Multi Dimensional Scaling Plot for Malagasy Dialects

As with the Polynesian languages, I converted the matrix into lexical distances and performed a 3D multidimensional scaling analysis of it. Plotting the first two principal components gives us an important insight: There appear to be 4 identifiable zones of similarity for the Malagasy dialects. And remarkably, this 3D map of lexical distances seems to somewhat replicate the numerical order in which the languages appear on the map of Madagascar.
To really understand what’s going on, we need to look at the words themselves, in gory detail. Here, the 23 dialects are arrayed in columns and the 200 words in rows. I’ve sorted them into categories of cognacy and color-coded them as indicated.
An edited version of the table separates cognates into categories for easier interpretation. At the top of this edited table we see a field of yellow words. These are words all cognate to the Merina dialect.
Further down the table we see the yellow words have been replaced in many locations by *Innovations* – words not cognate to Merina that occur in ONE DIALECT ONLY. I’ve shown these in light green. These innovations are not shared with surrounding dialects so they appear to have been spontaneously generated. There are all told 199 of these unique innovations. If we assume the Malagasy language has been on the island for 1200 years, this translates to one unique innovation per dialect every 139 years, on average. This is analogous to a genetic *mutation rate*. 
Here we see the main reason for the regional clustering in the MDS visualization of the Malagasy Matrix: words not cognate to Merina and with high affinity to a geographic region.

Here we see the blue words and green words. The blue words are non-cognate to Merina and are very popular in the south and west of the island, and much less so elsewhere. The green words concentrate mostly in the southeast coastal dialects, but some of them are also popular to the north and south.
The red words are heavily concentrated – but not exclusively - in the north of the island. But the purple words are popular throughout the coastal region of Madagascar and have even crept into the non-Merina Plateau dialects.
When we re-generate the Malagasy Matrix with the Regional Words Omitted, we can immediately see what was responsible for the order in the original MDS plot. It’s those regional words! Without them, we have no information about the regional variation of the lexical distances.
Multi Dimensional Scaling Plot for Malagasy Dialects

For comparison: here is the MDS plot with regional words included. Clearly, they are the determining factor in the shape of this MDS map.
We have two basic ways to slice the data: Vertically and Horizontally. The vertical method, the one we’ve discussed so far, counts the number of cognate differences between dialects to obtain a single number for each dialect pair: the Lexical distance.
Method 2: Compare Across All Dialects on a Word-by-Word Basis

The second method, the one we’ll be discussing now, takes a horizontal slice through the data and studies the way each word changes from dialect to dialect across all 23 dialects.
Case 1: Minor Changes Between Dialects

<table>
<thead>
<tr>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
<th>mamono</th>
</tr>
</thead>
<tbody>
<tr>
<td>velo</td>
<td>veloge</td>
<td>velo</td>
<td>velo</td>
<td>velo</td>
<td>velogny</td>
<td>velogna</td>
<td>velogna</td>
<td>velogna</td>
<td>velogno</td>
<td>velona</td>
<td>velona</td>
</tr>
<tr>
<td>atene</td>
<td>ate</td>
<td>aty</td>
<td>aty</td>
<td>aty</td>
<td>aty</td>
<td>aty</td>
<td>aty</td>
<td>aty</td>
<td>aty</td>
<td>aty</td>
<td></td>
</tr>
<tr>
<td>fitava</td>
<td>vava</td>
<td>vava</td>
<td>vava</td>
<td>vava</td>
<td>vava</td>
<td>vava</td>
<td>vava</td>
<td>vava</td>
<td>vava</td>
<td>vava</td>
<td>vava</td>
</tr>
<tr>
<td>agnare</td>
<td>agnaragne</td>
<td>agnara</td>
<td>agnara</td>
<td>agnara</td>
<td>agnara</td>
<td>agnara</td>
<td>agnara</td>
<td>agnara</td>
<td>agnara</td>
<td>anarana</td>
<td>anarana</td>
</tr>
<tr>
<td>lala</td>
<td>lalagne</td>
<td>lala</td>
<td>lala</td>
<td>lala</td>
<td>lalagna</td>
<td>lalagna</td>
<td>lalana</td>
<td>lalana</td>
<td>lalana</td>
<td>lalana</td>
<td>lalana</td>
</tr>
</tbody>
</table>

Most of the words vary not at all, or only slightly, across the island. The differences that have evolved result in slightly different pronunciation, that’s all.
Case 2: Independent Innovations...No Sharing

<table>
<thead>
<tr>
<th>oro</th>
<th>fagnatsogne</th>
<th>oro</th>
<th>oro</th>
<th>oro</th>
<th>vatagnoro</th>
<th>orogna</th>
<th>orogna</th>
<th>orona</th>
<th>orona</th>
<th>orona</th>
</tr>
</thead>
<tbody>
<tr>
<td>taly</td>
<td>taly</td>
<td>taly</td>
<td>taly</td>
<td>taly</td>
<td>tady</td>
<td>tady</td>
<td>tady</td>
<td>tady</td>
<td>tady</td>
<td>tady</td>
</tr>
<tr>
<td>mate</td>
<td>mate</td>
<td>maty</td>
<td>maty</td>
<td>maty</td>
<td>mangatsiaka</td>
<td>maty</td>
<td>maty</td>
<td>diso</td>
<td>maty</td>
<td>maty</td>
</tr>
<tr>
<td>ala</td>
<td>ala</td>
<td>ala</td>
<td>ala</td>
<td>ala</td>
<td>hazo</td>
<td>ala</td>
<td>ala</td>
<td>ala</td>
<td>atiala</td>
<td>ala</td>
</tr>
</tbody>
</table>

But in every dialect you find innovations, words that just pop up seemingly from nowhere, and replace the standard word IN THAT LANGUAGE ONLY.
Case 3: Independent Innovations with Local Sharing

Sometimes, these local innovations are shared with a single neighbor language. But this is pretty rare, occurring less than 1% of the time.
And sometimes, these local innovations start a trend that continues on down the line. First, TSINE gets elaborated into TSIKOLIKY. Only the first syllable is retained. Then, as the word moves south, the first syllable is abandoned altogether, leaving just OLIKY. But the original word TSINAY dominates outside the local area on the east coast where that first innovation occurred.
Case 4: Shared Substitutions

Frequently, it appears that some words travel together as synonyms, and change places in popularity along the way, as in the case of RA and LIO, competing words for “Blood”.

Merina
And in a couple of cases, it’s obvious that a foreign word has been taken up and used preferentially, as in the case of the French word for “Ice”
Here is our word table pruned down to contain only the 69 words which are known to be cognate to an Indonesian word. The colors other than yellow indicate these words are not cognate to the Merina dialect but are cognate to one of five Indonesian languages studied.
43% of these words survived with only minor variations in all dialects. An additional 20% survived in most dialects, with only occasional local innovations. The remaining 37% were subjected to widespread substitutions, some from Bantu but for the most part from unknown sources.
Consider the Ma’anyan word for “Tree”. In the north of the island it is rendered as KAKAZO. In the east coast and central highlands it was shortened to HAZO. In the far south it morphed in HATAY. Could this have happened in reverse order, with HAZO morphing into KAKAZO, a word so obviously cognate to the Ma’anyan KAKAO? Not likely! So support for a North-to-South migration hypothesis seems to come from the Ma’anyan word KAKAO, meaning “Tree”.
Another case is that of the word ALEM from Ngaju Dayak, meaning “Night”. It continually evolved as it moved south, eventually becoming ALINA in Merina. It further evolved along the east coast into ARIVA, finally becoming ALIKY on the west coast after a series of stepwise changes. However, once again we have an ambiguous piece of evidence. The word could have entered on the east coast and propagated unchanged northward and -in a changed form- southward.
Here is the strongest evidence I’ve found for immigration from the north. The Ma’anyan word for “Old” is MATUEH, and cognates to it are found only in the three northernmost dialects. All other dialects use ANTITRY, or something similar. So either there was a cognate to ANTITRY we were unaware of in the Proto language, or (more likely) there was an innovation that changed MATEUH to ANTITRY, which then caught on and propagated throughout the rest of the island.
Here is another case of a word with cognates unique to the northern dialects. The Bantu word for “Fire”, MOTO, became MOTRO in the northwest of the island.

AFO, is the Ma’anyan cognate for “Fire” is used in most other dialects. It’s possible that the Bantu influence was a later introduction, due to later immigration or trading. But it’s just as likely it was introduced by the original Bantu settlers, and if so argues for a northern point of entry.
When we plot the frequency of these regionally-clustered words versus location, a pattern begins to emerge. If we set a threshold of 5% as a definition for the regional boundaries, we can justify the lumping of these words into regions.
So the four zones we first noticed in the MDS plot are reaffirmed by analysis of the frequency of occurrence for words not cognate to Merina.
One hypothesis for language dispersal is this: both the northern and southern dialects started out in mid-island, along with the plateau dialects, then spread northward and southward as shown on the MDS plot and map.
Language Dispersal Hypothesis 2: North to South

A second hypothesis has the dialects moving south from the north cape, jumping up to the plateau, and continuing southward, eventually wrapping completely around the island.
Etienne de Flacourt wrote in 1658 that the area surrounding Fort Dauphin was divided into two hierarchies, one classified as ‘white’ (fotsy), the other ‘black’ (mainty). The royal family, at the top of the ‘white’ hierarchy, was descended from a group of immigrants known as Zafiraminina who had reached Madagascar some seventeen generations previously and migrated gradually down the east coast before arriving in the far southeast. The system of double authority reflected an accommodation between the Zafiraminina immigrants and the local peoples.

This might explain the pattern of word transmission we observe in the Malagasy dialects, where the plateau dialects and the coastal dialects could coexist in two separate groups, the Mainty and the Fotsy, as they gradually moved southward from their point of landing.
Here is a map of Madagascar settlement according to Hypothesis 1. Indonesians and Africans landed on the east coast and spread together, as Flacourt recorded, but both northward and southward. The south and west was settled mostly by Africans as Indonesians dominated the other areas and the Merina conquered and developed the high plains. The biggest problem with this hypothesis is the unlikelihood of an east coast landing by a mixed African/Indonesian group. The possibility of Africans and Indonesians did not meet until the Indonesians landed alone is remote. How could a small band of seafarers just happen to meet a similarly small group of African settlers on the east coast?
Here is a map of Madagascar settlement consistent with Hypothesis 2. It is driven by historical and archaeological findings as well as by common sense, and it is supported by the linguistic results. It assumes that the Fotsy and Mainty people had migrated together from the North Cape all the way to the south. At some point before the 18th century, the Merina moved en masse to the highlands and conquered or dominated other tribes in the area. Meanwhile, the mostly-African tribes, perhaps augmented by newcomers from Africa to the south cape, continued their settlement of the island south cape and south-west coast.
This tree diagrams the Northern settlement hypothesis. It is not based on Lexical Distance, but rather on a hypothetical settlement history supported by lexical data, not driven by it. From the northern landing, the Fotsy and Mainty migrated southward while keeping their linguistic preferences. The Mainty became the coastal dwellers shown in Red, Green and Blue variations, while the Fotsy maintained the purest Indonesian dialect, eventually becoming the Merina and exerting their linguistic domination over the Betsileo and Tsimehety highland tribes.
Support for Northern Immigration Hypothesis

• Early Archaeological Remains (685-745 CE) in Caves in the north, the most likely landing spot for a joint expedition from Africa.

• An East Coast Landing is Logically Unlikely

• Flacourt’s story about Mainty/Fotsy N>>S Migration

• North-to-South Word Evolution:

![Diagram showing word evolution from North to South]

• The presence of imported words Motro and Matoe in the North Only
Conclusions

• Lexicography and DNA indicate Borneo and Mozambique are Origins of Malagasy People
• Malagasy Languages Evolved from a Prototype Austronesian Language no Longer in Existence
• Lexical Distance Analysis Alone Does Not Accurately Predict Settlement Patterns
• But Analysis of Shared Cognates and Lexical Distances Do Inform Understanding of Settlement
• North-to-South Mainty/Fotsy Settlement Hypothesis is Supported, but not Proved by Lexical Analysis
We’ve seen how Polynesia could be differentiated from Melanesia based on Lexical Analysis and Multidimensional Scaling. And how we could develop a Phylogenic Tree for the languages of eastern Polynesia. For Madagascar, we saw how regional word patterns dominated the lexical distances between Malagasy dialects, and how word-by-word analysis along with Lexical Distance and MDS support the Northern Settlement Hypothesis for the world’s 4th largest island.
References:
YouTube Videos:

Tsimihety Girls (Music Video):
https://www.youtube.com/watch?v=bvoFt3UvO3w

How to Speak Malagasy (taught by a cool instructor):
https://www.youtube.com/watch?v=Dl3oMPLUNwY

Trials and Tribulations of Road Transport in Northern Madagascar:
https://www.youtube.com/watch?v=UOjxSNbuTqM

Vezo Fishing People at Toliara, West Coast:
https://www.youtube.com/watch?v=xEmPJC6soAA

Antananarivo City:
https://www.youtube.com/watch?v=oyi_dUQVi-I
This Presentation is viewable at:

http://www.roryvantuyl.com/Linguistics.html