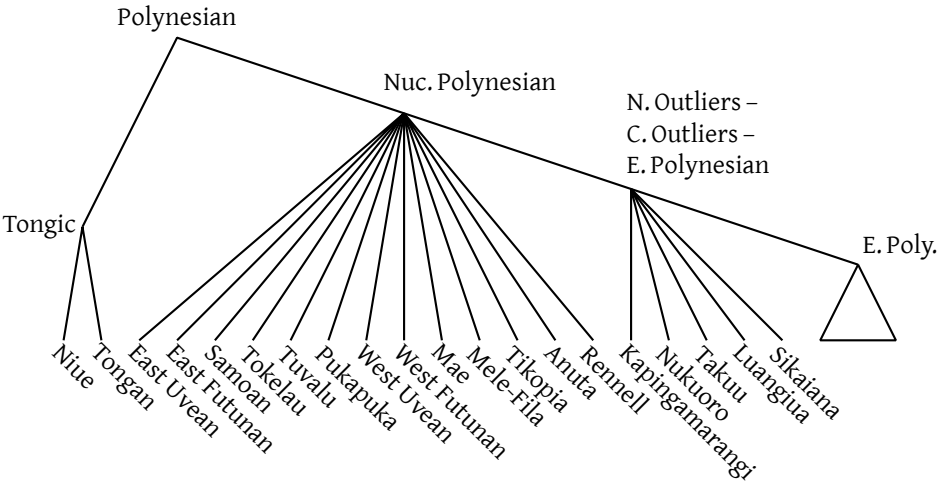
Linguistic mirages and lexical borrowing between Tongan and Samoan

Will Chang

9th International Conference On Oceanic Linguistics
University of Newcastle
4-8 February 2013

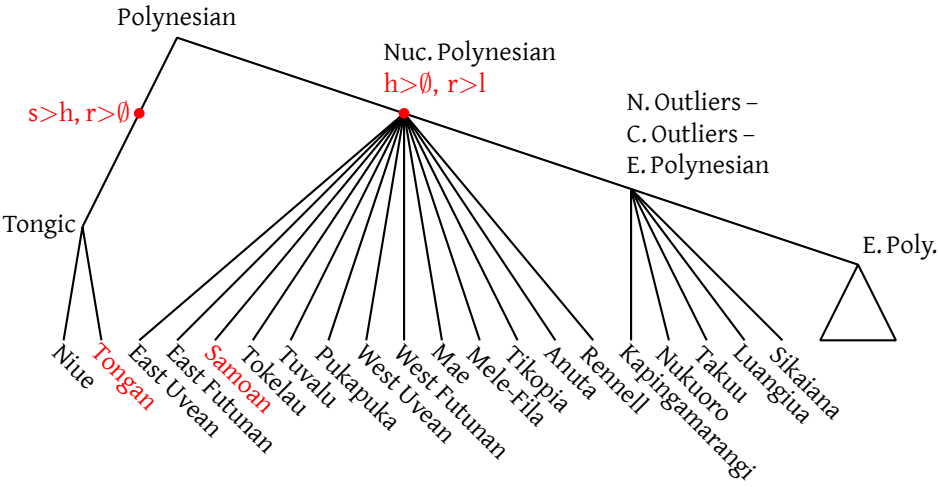
I think it's commonly believed that there had been long-standing cultural exchange between Tongan and Samoan societies before European contact. And if this is true, one would expect to find lots of Tongan-Samoan loanwords, and one may even expect that someone has published a list of such loans with dozens, or even hundreds, of items on it. I've looked for such a list, but have only ever found short ones, with four or five items. I thought about this for a bit before realizing that I would never find a long list. The reason is that any time there are related forms in Tongan and Samoan, it's possible to make a plausible-looking Proto Polynesian reconstruction, so you can never be sure that you're not dealing with true cognates. If you can't rule out that possibility, then you can't, of course, make a definitive list of Tongan-Samoan loans. But it's a somewhat different question to ask, *if* such a list existed, how many items would it contain? And this rather narrow concern is the topic of my talk.

Tongan and Samoan



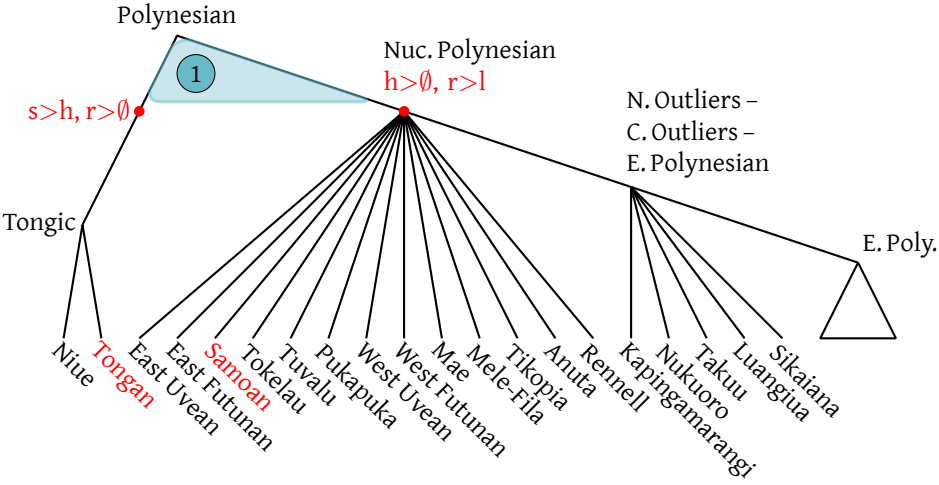
For this talk I assume these subgroups. (The North Outliers – Central Outliers – Eastern Polynesian node is due to William Wilson, 1985.)

Tongan and Samoan



Tongan and Samoan are on opposite sides of the first split, which is defined in part by the regular sound changes shown as red dots. For the sake of simplicity, I'm going to assume that all the sound changes happened at the same time.

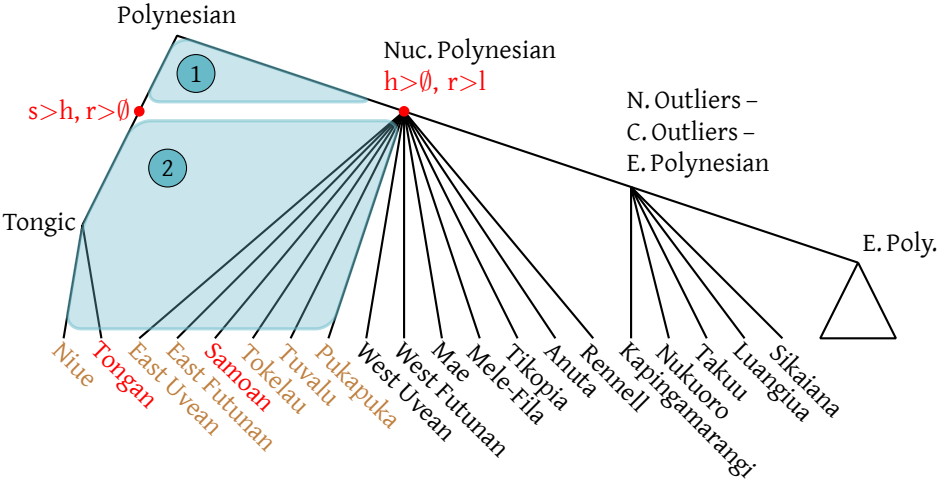
Tongan and Samoan



There may well have been lexical borrowing between Proto Tongic and Proto Nuclear Polynesian before the regular sound changes, but my analysis doesn't permit me to comment on this.

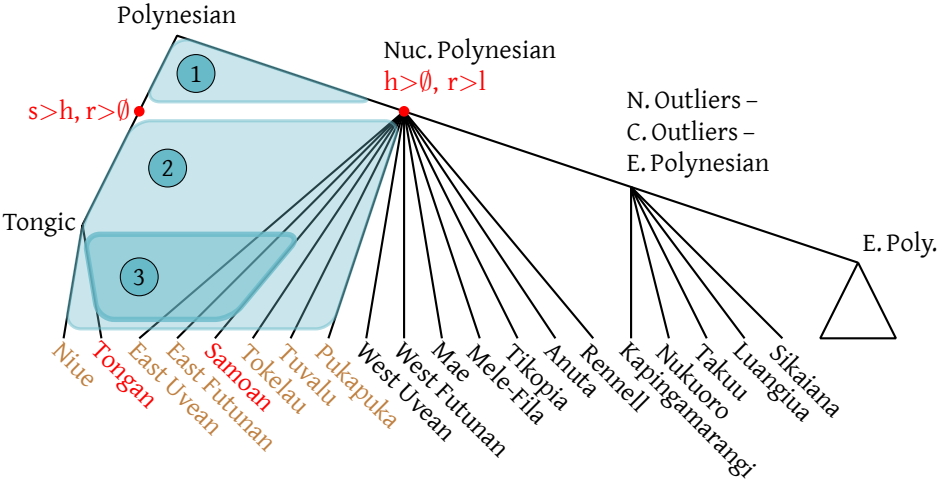
In response to a comment on early diversity: I agree that prior to the regular sound changes, there was already diversity in Nuclear Polynesian that would have been continued by the fan-like split underneath the node for Proto Nuclear Polynesian. Borrowing between Tongan and Samoan at this time may well account for some of the diversity in NP languages. But I mostly weasel out of having to deal with this issue by restricting the scope of my analysis to Tongan-Samoan borrowing after the regular sound changes have occurred.

Tongan and Samoan



The lexical borrowing that my analysis picks out comes after these regular sound changes. I find evidence of lots of borrowing between these eight languages of geographical Western Polynesia.

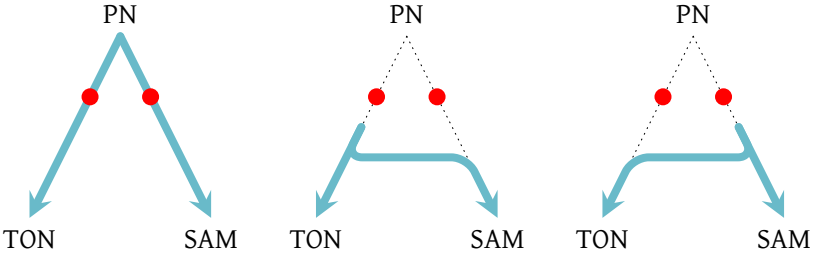
Tongan and Samoan



I find yet more borrowing between Tongan, Samoan, East Uvean, and East Futunan. If we take Western Polynesia to be a linguistic area, these four languages comprise the core.

Descent versus borrowing

- Three ways to get related forms in Tongan and Samoan:



- When the true history of an etymon contains borrowing, it is hard to prove it, because a PPn reconstruction is always possible.

To bring the problem into focus, I've made three simplified figures of possible histories of related forms in Tongan and Samoan. If the related forms are true cognates, the history is as shown on the left. But if borrowing was involved, the history may be as shown in the other two figures.

As I mentioned before, in all three cases, it's possible to come up with a plausible-looking reconstruction to Proto Polynesian. If the true history involved borrowing, the reconstruction to Proto Polynesian would be a *false reconstruction*, or a *mirage*, in the sense that it is merely an artifact of the comparative method, and we should have no reason to believe that the form really existed in Proto Polynesian.

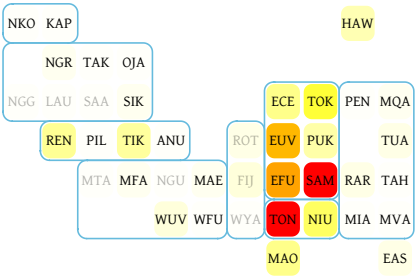
So, for individual reconstructions, it is impossible to rule out the possibility that it is a true reconstruction, unless there is another witness of the form that contradicts the reconstruction, which often there isn't. However, I submit to you that in the aggregate, it is easy to distinguish between true and false reconstructions. That is, if somehow you were able to round up a bunch of true reconstructions, and you were also able to round up a bunch of false reconstructions, that there would be obvious differences between the two sets.

Assemble two sets of PN etyma

From among etyma labeled .PN in POLLEX*, not attested outside Central Pacific, assemble:

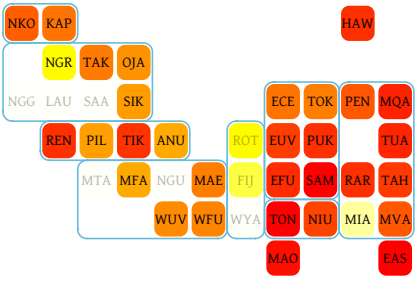
X: 113 etyma attested in

- Tongan &
- Samoan &
- up to 3 other Polynesian languages.



Y: 115 etyma attested in

- Tongan &
- Samoan &
- Rapanui &
- one of Hawaiian or NZ Maori &
- 4 Outliers &
- at least 4 more Polynesian languages anywhere.



*POLLEX (B. Biggs & R. Clark) is a giant comparative word list for the Polynesian languages.

Let me show this with a little experiment. I'm going to construct two sets of etyma using POLLEX. POLLEX is a very large comparative Polynesian word list, without which none of this would be possible. One nice thing about POLLEX is that it tells you the highest level to which a form can be reconstructed. I considered only etyma that reconstruct to Proto Polynesian and no further. Out of these etyma I assembled two sets.

For X, I chose etyma with reflexes in Tongan and Samoan, and no more than three other Polynesian languages. For Y, I chose etyma with reflexes in Tongan, Samoan, Rapanui, either Hawaiian or NZ Maori, in four Polynesian Outliers, and in four more Polynesian languages anywhere.

My goal in constructing X was to find etyma whose geographical distributions were as small as possible, but were still attested in Tongan and Samoan. I've plotted its distribution in the top map. The languages of POLLEX are laid out roughly geographically. The red color for Tongan and Samoan denote that all of the etyma in X have reflexes in these languages, as stipulated. EFU (East Futunan) and EUV (East Uvean) are orange, which means that most of the etyma also have reflexes in these languages as well. The light coloring elsewhere indicates that these etyma are seldom found elsewhere.

My goal in constructing Y was to find etyma that are widespread in Polynesian languages, and the bottom map indicates that I've basically succeeded.

Now, I think it would be natural to suppose that the etyma in X were at one point peculiar to either Tongan or Samoan, and then they spread to nearby languages. This would be the most parsimonious way to explain their present distributions.

On the other hand, it would be most natural to suppose that the etyma in Y descended from a form in Proto Polynesian, and that's how they achieved their present distribution.

Another nice thing about POLLEX is that it gives you a reconstruction for every etymon. Now I will look at the reconstructions for the etyma in X and Y, and compare them in the aggregate.

Protosound frequencies in X and Y

	Number of etyma with reconstructed sound																		
	a	u	i	o	k	t	e	l	f	m	p	n	ʔ	s	ŋ	r	w	h	v
X	91	47	39	41	41	42	22	29	33	23	21	13	17	24	17	2	10	2	1
Y	80	40	27	35	24	33	31	33	23	28	11	14	29	7	10	7	6	9	0

- X and Y have different distributions:
  $p = 0.008$ by Pearson's chi-squared test.
- X and Y cannot both be purely PPn in origin.

I will compare the sets by measuring how often a reconstruction contains a particular sound, as tabulated here.

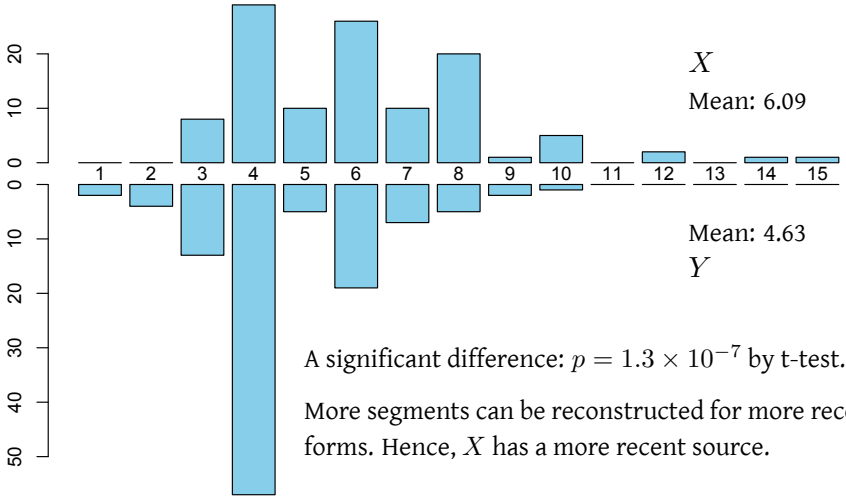
Protosound frequencies in X and Y

	Number of etyma with reconstructed sound																	
	a	u	i	o	k	t	e	l	f	m	p	n	ʔ	s	ŋ	r	w	h	v
X	91	47	39	41	41	42	22	29	33	23	21	13	17	24	17	2	10	2	1
Y	80	40	27	35	24	33	31	33	23	28	11	14	29	7	10	7	6	9	0

- X and Y have different distributions:
  $p = 0.008$ by Pearson's chi-squared test.
- X and Y cannot both be purely PPn in origin.

Note that the counts are especially skewed for *s*, *r*, and *h*, but also for some other sounds. It is a common question in statistics to ask whether two sets for which we have a bunch of measurements are significantly different. It turns out that this difference is very significant. This means that the reconstructions in X and the reconstructions in Y could not have come from the same pool. That is, they cannot all have come from the same time and place. Thus, they cannot all be true Proto Polynesian reconstructions.

Number of segments in reconstructed forms

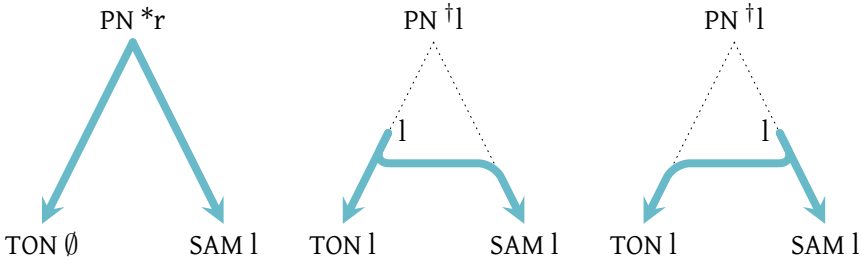


A significant difference: $p = 1.3 \times 10^{-7}$ by t-test.

More segments can be reconstructed for more recent forms. Hence, X has a more recent source.

Another thing that can be measured is the length of each reconstruction. I find that reconstructions in X are on average about 1.5 segments longer than reconstructions in Y. It stands to reason that more segments can be constructed for more recent forms, so this suggests that the reconstructions in X correspond to more recent forms than the reconstructions in Y.

When do we reconstruct PPn *r?



We can gain more insight by trying to understand the skewed counts for *r, *h, and *s. It turns out that *r appears only in true reconstructions, and never in false reconstructions. The reason for this is simple. PPn *r is lost on its way into Tongan, and is reflected in Samoan as *l*. Thus we only ever reconstruct it when one languages has a liquid and the other language does not. However, if there had been borrowing, then either both languages would have the liquid, in which case we reconstruct *l; or both would lack it, in which case we reconstruct zero. Thus a false reconstruction can never contain *r.

Number of reconstructed forms with *r*

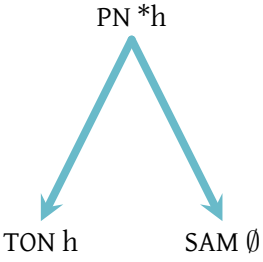
	Number of etyma with reconstructed sound																		
	a	u	i	o	k	t	e	l	f	m	p	n	ʔ	s	ŋ	r	w	h	v
X	91	47	39	41	41	42	22	29	33	23	21	13	17	24	17	2	10	2	1
Y	80	40	27	35	24	33	31	33	23	28	11	14	29	7	10	7	6	9	0

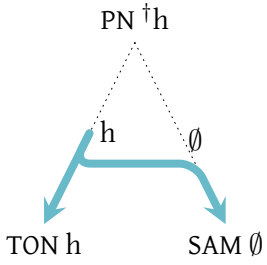
Estimate Tongan-Samoan loans in X.

- Suppose Y contains only true cognates.
- $2/7$ of X must be true cognates, so
- $5/7$ of X must be loans.

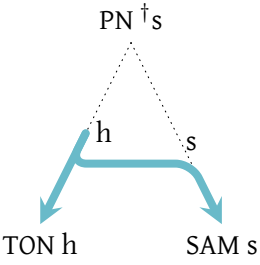
Now it makes sense that X has lower counts for *r. We can do a simple calculation to estimate the fraction X that consist of loans, if we assume that Y contains no loans. Observe that X and Y are about equal in size. So if we get 7 r's for Y and 2 r's for X, then $2/7$ of X must consist of non-loans, and $5/7$ of X must consist of loans.

When do we reconstruct PPn *h?





Doesn't actually happen!



Cf. TON or TAH hamala 'hammer' > SAM samala

With a bit more work, we can do a similar sort of analysis for reconstructed *h. PPn *h is retained in Tongan but lost on its way into Samoan. Now the question is, what happens when a Tongan word with *h* is borrowed into Samoan? Is it deleted, or does it get borrowed in as *s*? Some evidence for the latter is that European words that get borrowed into a Polynesian language with *h*, and subsequently into Samoan, show h>s. *Hammer* became Tongan or Tahitian *hamala* which then became Samoan *samala*. Likely the Samoans were aware of the equation between Tongan *h* and Samoan *s*, and made conscious use of it in borrowing from Tongan.

So if we assume that h>s is what always happens, then we can conclude that *h appears only in true reconstructions. This is because h:s correspondence is the regular outcome of PPn *s.

Comment from audience: *hamala* > *samala* is the only possible case of TON *h* to SAM *s* that he knows about.

Number of reconstructed forms with *h*

	Number of etyma with reconstructed sound																	
	a	u	i	o	k	t	e	l	f	m	p	n	ʔ	s	ŋ	r	w	h	v
X	91	47	39	41	41	42	22	29	33	23	21	13	17	24	17	2	10	2	1
Y	80	40	27	35	24	33	31	33	23	28	11	14	29	7	10	7	6	9	0

Estimate Tongan-Samoan loans in X.

- Suppose Y contains only true cognates.
- $2/9$ of X must be true cognates, so
- $7/9$ of X must loans.

This, then, explains the fewness of *h in reconstructions in X. We can again estimate the fraction of X that consists of loans. The result is similar (but not identical, since this estimate is based on different data).

An aside: reconstructed forms with *s*

	Number of etyma with reconstructed sound																	
	a	u	i	o	k	t	e	l	f	m	p	n	ʔ	s	ŋ	r	w	h	v
X	91	47	39	41	41	42	22	29	33	23	21	13	17	24	17	2	10	2	1
Y	80	40	27	35	24	33	31	33	23	28	11	14	29	7	10	7	6	9	0

Estimate the frequency of *s* in false reconstructions.

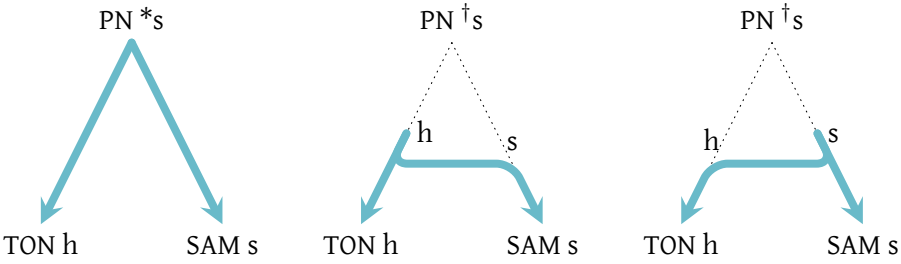
- Since Y has 7 reconstructions with *s*,
  X has 2 reconstructions with *s* from true cognates.
- The other 22 reconstructions with *s* are due to loans.
- False reconstructions contain *s* with $22/5$ times the frequency of *s* in true reconstructions!

Where do all these false reconstructions with *s* come from?

A curious thing in the counts is that there are a lot more reconstructed *s*'s in X than in Y. It turns out, if you do the math, that false reconstructions contain *s with more than 4 times the frequency as true reconstructions. The upshot of this is that if you are attempting to make a Proto Polynesian reconstruction and it contains *s, it could very well be a bogus reconstruction.

So where do all these *s*'s come from?

When do we reconstruct PPn *s?



- TON-to-SAM loans result in PPn †s when Tongan has *h*.
- Since PPn *h and PPn *s merged in Tongan as *h*, it should have a higher frequency of *h*.
- Since PPn *h and PPn *s have roughly the same frequency, this doubles the frequency of *h* in false reconstructions.
- This does not suffice to explain the much higher observed frequency, which is $22/5$ times the frequency of *h* in true reconstructions.

As I mentioned already, PPn *s produces a correspondence of h:s, and this happens when Tongan *h* is loaned into Samoan. This also happens when Samoan *s* is loaned into Tongan. But the TON-to-SAM case is the most interesting. Since *h* and *s* merged as *h* in the history of Tongan, Tongan contains around double the frequency of *h* as Proto Polynesian. (PPn *h and *s seem to have about the same frequency.) Thus, with TON-to-SAM loans, we would expect to reconstruct *s with double the frequency as we would have in true reconstructions.

But this doesn't explain the factor of four that we actually observe. Perhaps it is not possible to say more than that Proto Polynesian and Tongan are separated by a lot of time, and that this is enough time for the frequencies of sounds to randomly drift quite a bit. For unknown reasons, *h* simply became a more frequent sound in Tongan, or *s* became a more frequent sound in Samoan.

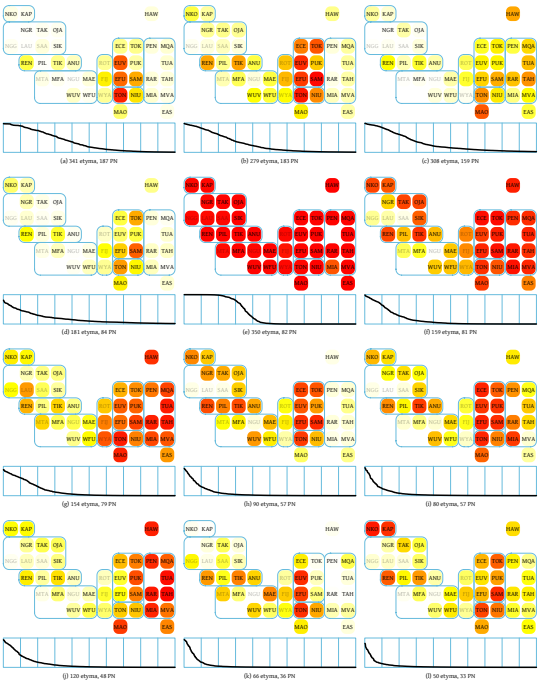
This concludes the part of my talk where I walk through my reasoning process. Next I will briefly summarize what happens when you automate the analysis and apply it to all of the etyma in POLLEX.

Clustering results

Cluster the 4,000+ etyma in POLLEX by their distributions.

Each etymon n is in cluster k with probability q_{nk}.

For each cluster k, the black curve shows q_{nk} for the 1000 etyma with highest q_{nk}.



Previously I constructed sets of etyma by hand. Now I used a statistical model to cluster together etyma in POLLEX on the basis of their distribution among the languages of POLLEX.

The results are shown here. These are the 12 clusters that contain the most PN etyma. Each map shows the geographical distribution of the etyma in the cluster. Since I used a probabilistic model, it's the case that each etyma is assigned to each cluster with some probability, so cluster memberships are graded. Underneath each map, I plot the degree of membership of the 1000 etyma with the highest degrees of membership. You can tell by how sharply the curve drops whether the category has sharp or fuzzy boundaries.

Each cluster is labeled with the mean number of etyma in it, and also how many are putatively reconstructed to Proto Polynesian.

The nice thing about a clustering model is its objectivity. When I was constructing X and Y, I could be accused of finagling the sets to obtain favorable results, but here, what comes out is what comes out.

Statistics on clusters

	# Pn etyma	Mean length	# total seg	Frequency of reconstructed sound, per thousand segments																		
				a	u	i	o	k	t	e	l	f	m	p	n	ʔ	s	ŋ	r	w	h	v
(a)	187	5.78	1080	235	95	76	77	77	68	56	45	58	42	30	17	37	35	28	4	14	6	1
(b)	183	5.32	972	223	90	87	83	59	75	60	50	49	40	31	20	58	28	23	5	11	8	0
(c)	158	5.18	821	238	79	95	78	66	66	61	50	41	55	33	28	27	22	33	6	15	6	0
(d)	85	5.39	460	242	97	78	86	53	67	44	56	41	56	37	25	37	33	26	7	8	8	0
(e)	82	4.73	387	268	51	67	87	56	63	57	71	38	73	22	13	48	11	36	11	11	17	0
(f)	81	4.51	365	210	101	78	94	45	60	63	68	42	44	37	38	36	17	31	10	17	9	0
(g)	79	5.17	411	243	95	88	65	58	77	62	62	39	55	25	22	32	17	24	10	13	12	0
(h)	57	5.05	289	232	103	92	72	63	64	42	54	34	49	29	10	71	34	21	8	15	6	0
(i)	56	5.33	301	265	75	76	77	76	59	52	59	37	49	31	33	37	21	28	4	8	9	3
(j)	48	5.00	240	213	66	95	98	75	56	76	75	30	55	32	17	24	29	19	15	18	8	0
(k)	36	5.20	189	220	105	64	97	65	39	46	62	64	46	21	16	42	35	40	19	6	13	0
(l)	33	5.33	178	280	100	58	75	60	75	28	56	47	43	19	19	58	9	44	12	6	11	0

Just as with X and Y, I take measurements of the reconstructions in these clusters. To identify a cluster as containing lots of loans, we look for low frequencies for *h* and *r*, and a high frequency for *s*. Note that for *h* and *r*, a frequency of about 10 or 12 per thousand can be considered high.

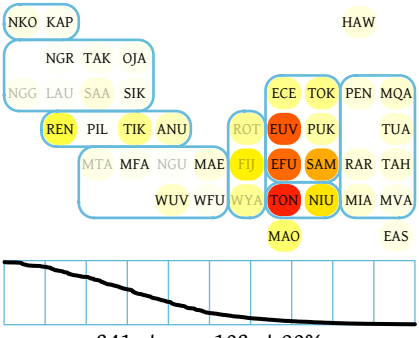
Question from audience: Have I try to explain the counts for glottal stop? And have I looked into exploiting the k:ʔ correspondence?

Answer: A TON-to-SAM loan with ʔ would result in †ʔ, while a SAM-to-TON loan can never result in †ʔ. I will look into whether I can use this fact to estimate the direction of borrowing in loans. As for the k:ʔ correspondence, my impression is that it's fairly late, and in any case, there is probably nothing to exploit, since the Tongans and Samoans would have understood this equation when borrowing from one another.

Reply: There is not a lot of evidence for this, but it's possible that the k:ʔ correspondence continues dialectal alternations from Proto Central Pacific.

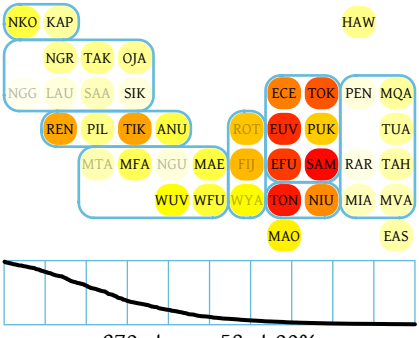
Western Polynesia

	# Pn etyma	Mean length	# total seg	Frequency of reconstructed sound, per thousand segments																		
				a	u	i	o	k	t	e	l	f	m	p	n	ʔ	s	ŋ	r	w	h	v
(a)	187	5.78	1080	235	95	76	77	77	68	56	45	58	42	30	17	37	35	28	4	14	6	1
(b)	183	5.32	972	223	90	87	83	59	75	60	50	49	40	31	20	58	28	23	5	11	8	0



341 etyma, 103 at 90%

(a) 187 PN



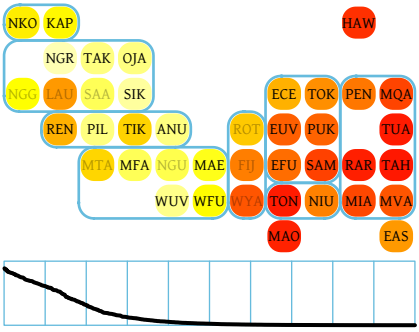
279 etyma, 58 at 90%

(b) 183 PN

These two clusters seem to depict related phenomenon. The one on the left shows the core of a linguistic area, and the one on the right shows an area that surrounds the core. The counts give clear evidence of loans.

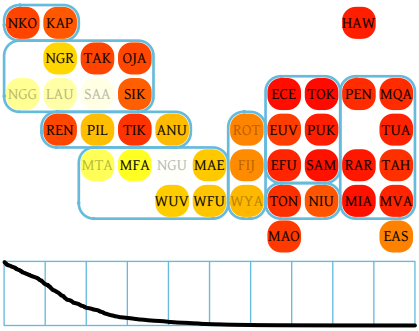
Greater Polynesia

	# Pn etyma	Mean length	# total seg	Frequency of reconstructed sound, per thousand segments																	
				a	u	i	o	k	t	e	l	f	m	p	n	ʔ	s	ŋ	r	w	h	v
(f)	81	4.51	365	210	101	78	94	45	60	63	68	42	44	37	38	36	17	31	10	17	9	0
(g)	79	5.17	411	243	95	88	65	58	77	62	62	39	55	25	22	32	17	24	10	13	12	0



154 etyma, 0 at 90%

(g) 79 PN



159 etyma, 27 at 90%

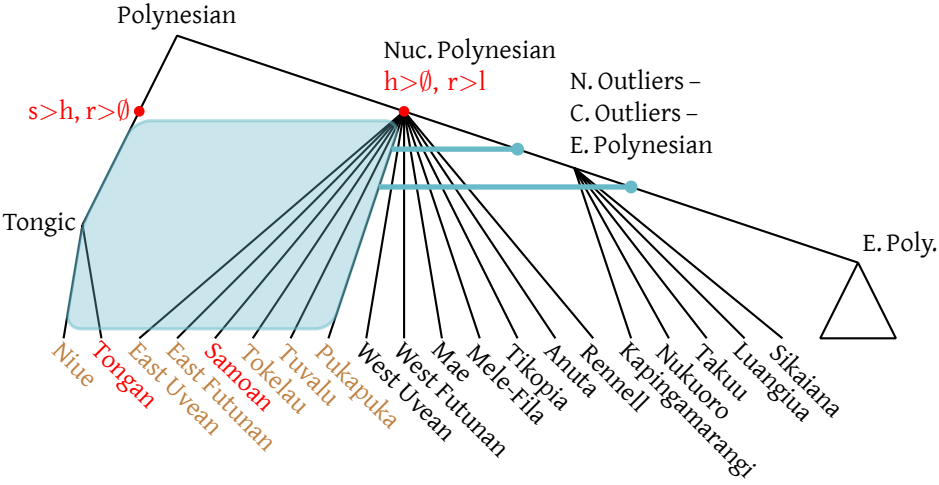
(f) 81 PN

These two clusters consist of etyma spread throughout Western Polynesia, that are also in Eastern Polynesia, and in the case of the cluster on the right, in the Nothern Outliers as well. I pondered for a long time whether these etyma descended straight from Proto Polynesian, or whether lexical borrowing was involved.

The counts — the highish *h* and *r* frequencies, and the lowish *s* frequencies, suggest the former, but that seems wrong. If they were descended from Proto Polynesian, we would have to explain how they were lost the Southern Outliers or in the Outliers as a whole. Since the Southern Outliers do not form a clade, the loss would have had to be at multiple points in the family tree, which seems improbable.

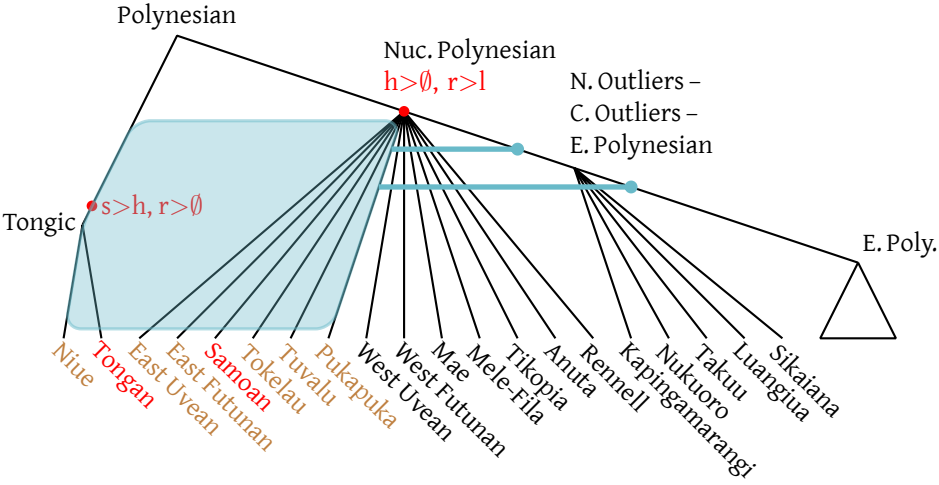
The alternative is to see the ancestor of Nothern and Central Outliers and Eastern Polynesia as belonging to the periphery of a linguistic area in which etyma diffused fairly freely, as shown...

The periphery



...here. But then how to explain the counts? I think at this point, there are many possibilities, so my analysis loses much of its cogency. One possibility is that some of the sound changes that define Tongic or Nuclear Polynesian didn't take place until later,...

The periphery



...as shown here. Now, if borrowing happened between the time of the red dots, we would still reconstruct *h, and at least if the borrowing was from Tongan into Samoan, we would still reconstruct *r. But as I've said, there seem to be many possibilities, so I won't dwell further on this.

How many etyma were borrowed?

Posit borrowing to account for drops in the frequency of *r and *h, and for surges in the frequency of *s.

	10%ile	med	90%ile
The number of PPn etyma reflected in both TON and SAM:	—	864	—
Of these, the number that were borrowed between TON and SAM:	313	379	448
The percentage of true reconstructions with *h:	6.5	7.4	8.7
The percentage of true reconstructions with *s:	1.9	3.4	5.4
The percentage of true reconstructions with *r:	5.8	6.6	7.7
The percentage of true reconstructions with *l:	24.9	27.8	30.5
The percentage of false reconstructions with *h is set to zero.	—	—	—
The percentage of false reconstructions with *s:	19.3	22.5	26.8
The percentage of false reconstructions with *r is set to zero.	—	—	—
The percentage of false reconstructions with *l:	20.4	24.1	27.8

Just as with X, we can estimate the fraction of each cluster that consists of loans. This estimate is based on the frequency of h, s, r, and l phonemes in reconstructions for loans and non-loans. As before, I stipulate that false reconstructions cannot contain h or r. The model figures out the frequencies of the other sounds. Jointly the model estimates the number of loans in each cluster, and the sum is reported in the second row.

The model inferred that, of the 864 PN etyma in POLLEX that are attested in Tongan and Samoan, around 45% were loans. This is surprisingly high, but in fact there were many details that I swept under the rug, which if I took into account, would only make this figure higher.

Likely loans (1/2)

PPn Recon.	Loan	Distribution (+ TON SAM)	Gloss
masuʔa	0.9570	EUV EFU	Overflow.
kawasasa	0.9564	EUV EFU	A creeper used to posion fish.
sapotu	0.9556	EUV EFU REN	To pant.
fesikiʔaki	0.9552	NIU EUV EFU	(Ex)change places.
matamoso	0.9550	EUV TIK	A plant with red seeds.
sou	0.9550	ROT EFU	To be agitated (of the sea).
saumi	0.9512	EUV MAO	Banana variety.
saakato	0.9478	EFU	Fern sp.
sawili	0.9454	FIJ NIU EUV EFU	To blow of the wind; breeze.
silo	0.9434	EUV TUA	Entrails.
akatasa	0.9412		Herb sp. (Rorippa sarmentosa).
sela	0.9404	EUV EFU	Asthma; gasp for breath.
suŋalu	0.9368	EFU	Driftwood and shells worn by wave action; jetsam.
sasake	0.9334	EVU ECE PUK	East.
seku	0.9288	MQA	Fantail (Rhipidura sp.).
ʔaʔasi	0.9260	NIU EFU TIK REN	Visit.
matapisu	0.9214	NIU TIK ANU	Sheelfish sp., Limpet.

Finally, we can ask the model to tell us which etyma are most likely loans. I list these on this slide and the next. Interestingly, all of the etyma contain *s*. No semantic generalizations seem possible, aside from the fact that none are basic vocabulary items. Of individual interest is ʔaʔasi 'visit'. I wonder if it's the case that when people go places, they say that they're 'visiting', and that's how the form spread.

Likely loans (2/2)

PPn Recon.	Loan	Distribution (+ TON SAM)	Gloss
masisi	0.9206	EUV EFU MFA REN	Cut or broken lengthwise.
ŋaosi	0.9174	EUV TOK ECE	Make, do deal with.
maasoli	0.9172	NIU EUV EFU MFA	A type of banana.
masunu	0.9108	NIU EFU WUV REN MAO	Singed, scorched, burnt.
sulali	0.9094	EFU HAW	Bêche-de-mer sp.
salii	0.9048	EFU WUV ECE	A small fish.
sana	0.8988	NIU TIK PUK	Job's Tears (Coix sp.) [...] used for necklaces.
masele	0.8958	FIJ EUV EFU OJA MAO	A sedge.
ʔasi	0.8852	ROT NIU EUV EFU WUV MAE TIK REN PIL TOK ECE TAK SIK MQA PUK	Visit.
sikuleʔo	0.8830	REN ECE	Faint voice; echo (Clk).
sali	0.8812	EFU TOK PUK	Scoop out, up.
sopo	0.8796	NIU EUV EFU WUV WFU MAE MFA TIK REN TOK ECE KAP NKO TAK OKA SIK NGR MAO	Jump (up or down), cross a boundary.
masiki	0.8788	EUV EFU TIK ECE TUA	Be lifted, raised.

On this slide is an unreduplicated form of the same word. It has an unusual distribution.

Comment from audience: *sana is a late borrowing from outside Polynesia. *Note to self*: I should look into the what the form is in the donor language.

FIN

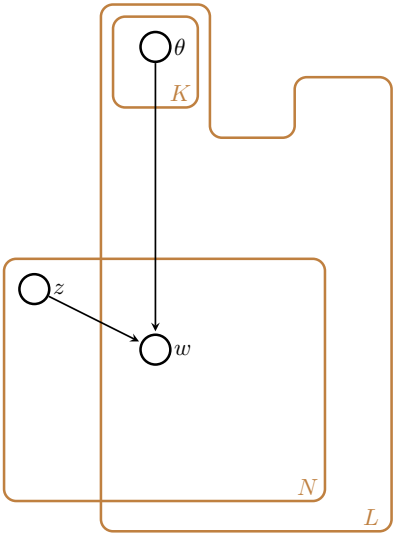
Distributions, observed and actual

- The *distribution* of an etyma is the set of languages that contain it.
- POLLEX gives the *observed distribution* of each etymon.
- The *actual distribution*, a superset, must be inferred.

Cluster the 4,000+ etyma in POLLEX by their inferred actual distributions.

In order for the model to work properly, I had to make an inferential leap: I needed to infer the real distribution of an etymon based on what POLLEX gives as its distribution.

ETYMDIST



$\theta_{kl} \in (0, 1)$
Intensity of class k in language l.

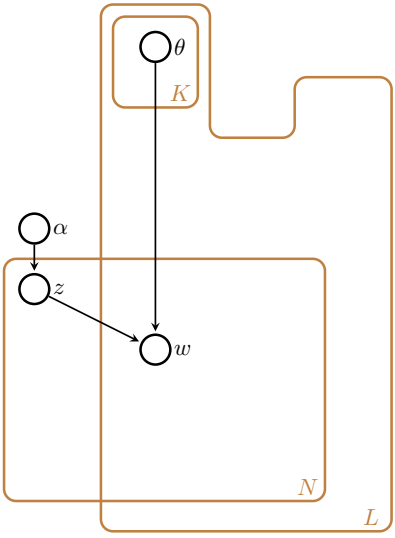
$z_n \in \{1, 2, \ldots\}$ (α)
Class of eytmon n.

$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

This is the heart of the model. The variable z_n denotes the cluster to which etymon n belongs. Each cluster k is defined by a bank of *intensities* $\theta_{k1}, \ldots, \theta_{kL}$, one for each language l. The intensity θ_{kl} denotes the probability that an etymon of class k will exist in language l.

The variable w is an $N \times L$ binary matrix, with each entry w_{nl} denoting whether etymon n exists in language l. To generate w_{nl}, first look up the cluster of the etymon z_n, then look up the intensity of that cluster in that language ($\theta_{z_n l}$, i.e. θ indexed by z_n and l), and then perform a weighted coin toss.

ETYMDIST

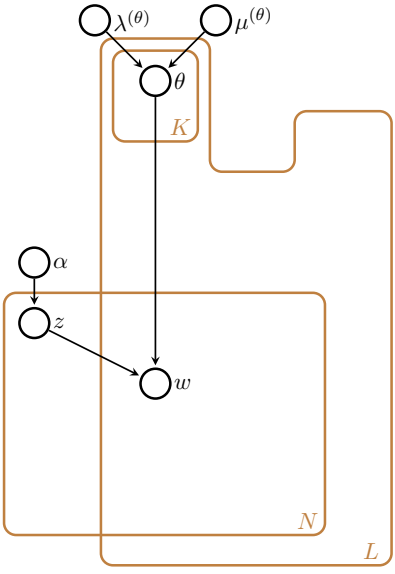


$\theta_{kl} \in (0, 1)$
Intensity of class k in language l.

$\alpha \in (0, \infty)$
Diversity parameter
$z_n \in \{1, 2, \ldots\}, fromCRP(\alpha)$
Class of eytmon n.
$w_{nl} \in \{0, 1\},\ w_{nl} \sim$ Bernoulli$(\theta_{z_n l})$
Is eytmon n actually in language l?

The cluster variables z_n correspond to table assignments of a draw from a Chinese restaurant process parameterized by α. In less technical terms, this is simply a prior over all possible cluster configurations, ranging from putting each etymon in its own cluster, to putting all etyma into the same cluster. The hyperparameter α governs how likely it is for etyma to clump together. Smaller α means fewer, larger clusters.

ETYMDIST

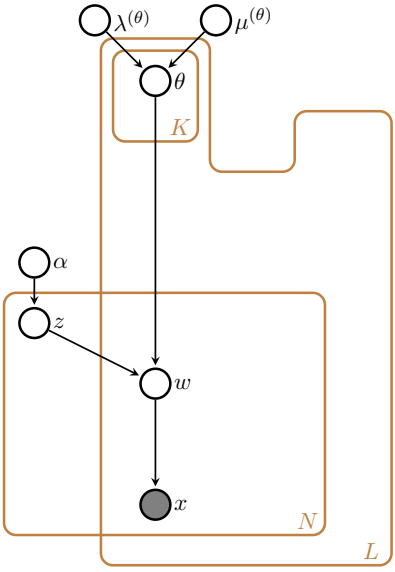


$\lambda^{(\theta)} \in (0, \infty),\ \mu^{(\theta)} \in (0, 1)$
Hyperparameters for intensity.
$\theta_{kl} \in (0, 1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1 - \mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\alpha \in (0, \infty)$
Diversity parameter
$z_n \in \{1, 2, \ldots\}, fromCRP(\alpha)$
Class of eytmon n.
$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

The intensities are all generated from the same beta distribution, parameterized by $\lambda^{(\theta)}$ and $\mu^{(\theta)}$. I strongly suspect that there are easy ways to improve this prior, but I have yet to think of any.

ETYMDIST



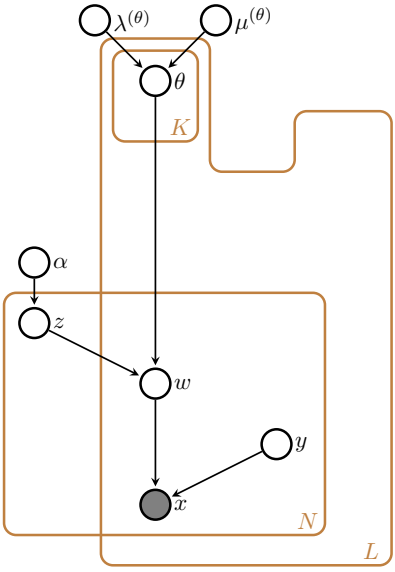
$\lambda^{(\theta)} \in (0, \infty),\ \mu^{(\theta)} \in (0, 1)$
Hyperparameters for intensity.
$\theta_{kl} \in (0, 1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1 - \mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\alpha \in (0, \infty)$
Diversity parameter
$z_n \in \{1, 2, \ldots\}, fromCRP(\alpha)$
Class of eytmon n.
$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

$x_{nl} \in \{0, 1\}$
Is etymon n observed in language l?

The difference between actual and observed distributions is the difference between w and x. The variable x_{nl} indicates whether etymon n is attested in language l. It is a function of w_{nl}, but also of ...

ETYMDIST

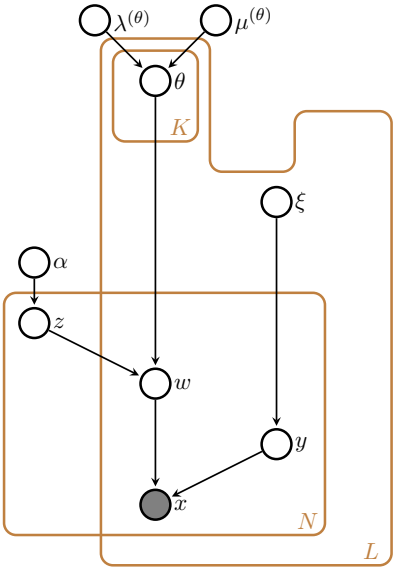


$\lambda^{(\theta)} \in (0, \infty),\ \mu^{(\theta)} \in (0, 1)$
Hyperparameters for intensity.
$\theta_{kl} \in (0, 1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1 - \mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\alpha \in (0, \infty)$
Diversity parameter
$z_n \in \{1, 2, \ldots\}, fromCRP(\alpha)$
Class of eytmon n.
$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?
$y_{nl} \in \{0, 1\}$
Observability of etymon n in language l.
$x_{nl} \in \{0, 1\},\ x_{nl} = w_{nl} \cdot y_{nl}$
Is etymon n observed in language l?

...y_{nl}, which indicates whether an etymon n in language l would be observable, if it were already to exist in language l.

ETYMDIST



$\lambda^{(\theta)} \in (0, \infty),\ \mu^{(\theta)} \in (0, 1)$
Hyperparameters for intensity.

$\theta_{kl} \in (0, 1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1 - \mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\xi_l \in (0, 1)$
Coverage for language l.

$\alpha \in (0, \infty)$
Diversity parameter

$z_n \in \{1, 2, \ldots\}, fromCRP(\alpha)$
Class of eytmon n.

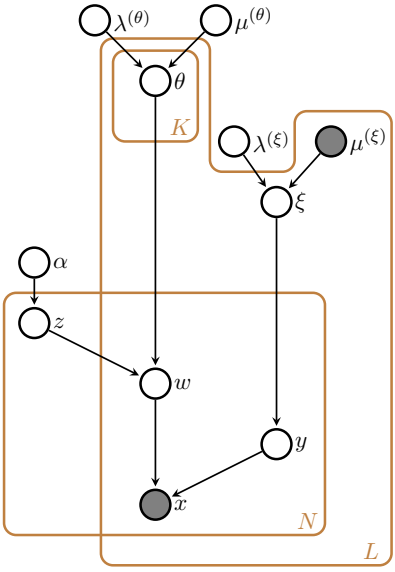
$w_{nl} \in \{0, 1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

$y_{nl} \in \{0, 1\},\ y_{nl} \sim \text{Bernoulli}(\xi_l)$
Observability of etymon n in language l.

$x_{nl} \in \{0, 1\},\ x_{nl} = w_{nl} \cdot y_{nl}$
Is etymon n observed in language l?

Each y_{nl} is derived via a Bernoulli distribution parameterized by the lexicographic coverage for language l, ξ_l.

ETYMDIST



$\lambda^{(\theta)} \in (0,\infty),\ \mu^{(\theta)} \in (0,1)$
Hyperparameters for intensity.

$\theta_{kl} \in (0,1),\ \theta_{kl} \sim \text{Beta}(\mu^{(\theta)}\lambda^{(\theta)}, (1-\mu^{(\theta)})\lambda^{(\theta)})$
Intensity of class k in language l.

$\lambda^{(\xi)} \in (0,\infty),\ \mu_l^{(\xi)} = 0.9N_l / \max\{N_1, N_2, \ldots, N_L\}$
N_l = #entries for language l.

$\xi_l \in (0,1),\ \xi_l \sim \text{Beta}(\mu_l^{(\xi)}\lambda^{(\xi)}, (1-\mu_l^{(\xi)})\lambda^{(\xi)})$
Coverage for language l.

$\alpha \in (0,\infty)$
Diversity parameter

$z_n \in \{1,2,\ldots\}, fromCRP(\alpha)$
Class of eytmon n.

$w_{nl} \in \{0,1\},\ w_{nl} \sim \text{Bernoulli}(\theta_{z_n l})$
Is eytmon n actually in language l?

$y_{nl} \in \{0,1\},\ y_{nl} \sim \text{Bernoulli}(\xi_l)$
Observability of etymon n in language l.

$x_{nl} \in \{0,1\},\ x_{nl} = w_{nl} \cdot y_{nl}$
Is etymon n observed in language l?

It stands to reason that ξ_l correlates positively with N_l, the number of forms in POLLEX for language l. This correlation is encoded via a beta distribution. The degree of correlation is encoded in the hyperparameter $\lambda^{(\xi)}$.

Note 1: One reason that the correlation is imperfect, is that for non-Polynesian languages in POLLEX, the lexicographic coverage is much higher than N_l would suggest. The reason is that POLLEX selectively contains etyma that appear in Polynesian languages, which artificially limits the N_l for non-Polynesian languages.

Note 2: Supplying the model with some knowledge of N_l seemed critical for getting the model to learn reasonable values for ξ_l; or perhaps I did not try hard enough to get it to work without N_l.

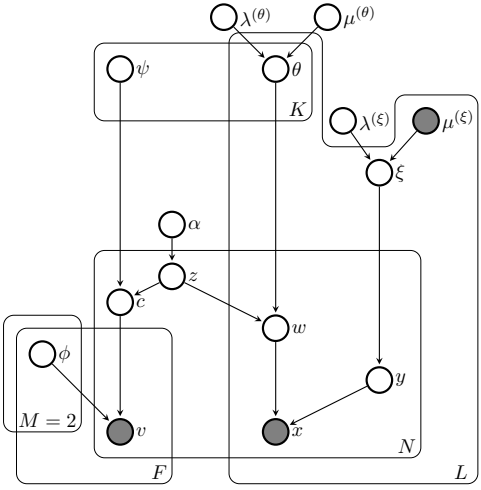
ETYMDIST with sound frequencies

$\psi_k \in [0, 1]$
Amount of TON-SAM borrowing in cluster k.

$c_n \in \{0, 1\}, c_n \sim \text{Bernoulli}(\psi_k)$
Whether etymon n is a TON-SAM loan.

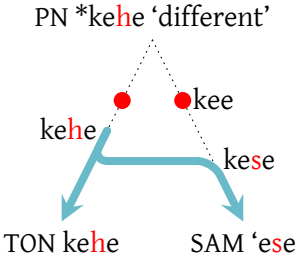
$\phi_{mf} \in [0, 1]$
Frequency of feature f in a loan ($m = 1$) or non-loan ($m = 0$). Zero for *h or *r in loans.

$v_{nf} \in \{0, 1\}, v_{nf} \sim \text{Bernoulli}(\phi_{c_n f})$
Presence or absence of feature f in etymon n.



A plate diagram of an augmentation of the above model, used for jointly inferring etyma clusters and instances of Tongan-Samoan borrowing.

Dialect borrowing



*r, *h in *X*

.PN	fakalotolotorua	Uncertain, of two minds.
TON	fakalotolotoua	causing or tendency to cause uncertainty (Cwd).
EUV	fakalotolotolua	faire hesiter, hesitant (Rch).
EFU	fakalotolotolua	faire hesiter, en hesitation (Rch).
SAM	faʔalotolotolua	
.PN	tuqurua	Cut in two; half way point
TON	tuʔuua	Cut in two.
REN	tuʔugua	Break in two, sever, divide.
SAM	tuulua	Interval.
TOK	tuulua	Half full (Sma).
.PN	gahele	Soft.
TON	ŋahele	Move slowly and noiselessly; creep, crawl.
SAM	gaele	Shake, oscillate, as a bog.
MAO	ŋaere	Soft, quake, oscillate, as a bog.
HAW	naele	Soft.
.PN	holi	Desire, long for.
TON	holi	Want, desire, crave, wish for (Cwd).
EUV	holi	(Rch).
EFU	oli	Désirer, envie (Mfr).
SAM	olioli	Be eager for (Mnr).
ECE	holi	Favour one side or contender to win (Rby).