Supplementary information for: **Rate of language evolution is affected by population size.**

Authors: Lindell Bromham, Xia Hua, Thomas G. Fitzpatrick, Simon J. Greenhill

Supplementary methods:

Details of the models of language and population change

Table S1: Population data for each language included in this study.

Table S2: Comparisons of models of language evolution and likelihood ratio tests on the effect of population size on language evolution rates

Table S3 : Island area and rates of word gain and loss

Figure S1: Histograms of observed and expected numbers of changes (gains+losses)

Figure S2: Population size and number of identified loan words per language

Figure S3: Illustration of the two modes of language origin modelled

---

SUPPLEMENTARY METHODS:

Details of the models of language and population change

To implement Poisson regression, we make two assumptions. First, the gain or loss of words follows a Poisson process. Second, rates of gain or loss are linear functions of population size on log-log scales. As a result, the probability of observing S_1 words gained or lost in a language and S_2 words gained or lost in its sister language, since they split at time T back in history, is:

$$p(S_1, S_2) = e^{-\int_0^T [\lambda_1(t)+\lambda_2(t)]dt} \frac{[\int_0^T \lambda_1(t)\,dt]^{S_1}[\int_0^T \lambda_2(t)\,dt]^{S_2}}{(S_1)!(S_2)!} \quad \text{Eqn.1}$$

$\lambda_1(t)$ is the gain or loss rate of language 1 and equals $e^{b\log(X_1(t)/X_0)+\lambda_0}$, where X_1 is the population size of language 1 at time t, X_0 is the population size of the common ancestor of the language pair and λ_0 is its gain or loss rate, b measures the effect of population size on gain or loss rate. Since X_0 and λ_0 are unknown, they can be grouped into a single parameter, such that $\lambda_1(t)=e^{b\log(X_1(t))+a}$. Similarly, $\lambda_2(t)=e^{b\log(X_2(t))+a}$. We then estimate parameter b under models of language and population change that differ in four aspects as described below.

Phylogenetic structure

If language evolution is not phylogenetically structured, the relationship between language change and population size may be independent of the state of the common ancestor. In this case, we can treat changes in different languages as independent experiments on the same relationship between language change and population size, defined by the two parameters a and b. Otherwise, if, for example, an ancestral language evolves faster than others of the same population size and its descendent languages inherited the high rate of language changes, then the relationship between language change and population size in those descendent languages should have a larger intercept (parameter a) than languages descending from other ancestors. We account for such a process of descent by fitting different intercepts of language evolving rates for different pairs of languages

Constant population size vs. growing population

If each of the populations grows slowly following colonization of a new area, then we expect a long period in each language's history in which the historical population was much smaller than the current population. To account for this period of population growth, we model population growth in each language as a continuous density-dependent process with carrying capacity equal to its current population size, such that $X_1(t) = \frac{X_1(T)}{1+(\frac{X_1(T)}{N}-1)e^{-rt}}$, for which a common population growth rate (r) and initial population size (N) are fitted to all language pairs. Otherwise, if population grows rapidly to the carrying capacity of the inhabited area then stabilised, the current population size is a good approximation of population size at any time point.

Fission vs. colonization

We account for different modes of the origination of a new language by using the archeological dates that most closely approximate the age of the split between two sister languages (T in equation 1). If the sister languages originated by splitting an ancestral population (Fission model in Figure S2), the older date of the establishment dates of the two languages should more accurately represent the age of the split (t_A in Figure S2). If a language is originated through colonization, where a founder population is established in a new area while the original population continues to occupy the original area (Colonization model in Figure S2), then the younger of the establishment dates of the two languages should more accurately estimate their age of the split (t_B in Figure S2).

Founder effect vs. gradual loss of words

If the founding population of a language does not use all the lexemes from the ancestral language, there may be an initial loss of lexemes when the population is founded (i.e., founder effect). We model this sudden loss by introducing a new parameter S_f to describe the absolute number of words lost due to founder effect, such that if both languages were subject to founder effect since they split (Fission model in Figure S2), equation 1 becomes:

$$p(S_1,S_2)=e^{-\int_0^T[\lambda_1(t)+\lambda_2(t)]dt}\frac{[\int_0^T\lambda_1(t)\,dt]^{S_1-S_f}[\int_0^T\lambda_2(t)\,dt]^{S_2-S_f}}{(S_1-S_f)!(S_2-S_f)!}.$$

If a language, say language 1, is derived from its sister language (Colonization model in Figure S2), only language 1 was subject to founder effect since the split of the two languages, then equation 1 becomes:

$$p(S_1,S_2)=e^{-\int_0^T[\lambda_1(t)+\lambda_2(t)]dt}\frac{[\int_0^T\lambda_1(t)\,dt]^{S_1-S_f}[\int_0^T\lambda_2(t)\,dt]^{S_2}}{(S_1-S_f)!(S_2)!}.$$

We investigate all possible models that vary in the above four aspects. When accounting for phylogenetic structure in language evolution, we cannot estimate founder effect separately for each language pair due to constraints on degree of freedom. Thus, we assume equal number of words lost due to founder effect in all the language pairs. When accounting for phylogenetic structure and assuming constant population size over time, each different origination mode of a new language gives same fit to the data because the split age of a language pair becomes a part of the intercept to optimize. This fact allows us to use all the ten language pairs, including those whose establishment dates are not available.

Table S1: Population data for each language included in this study.

Language		Population size			Area	Age
Name	ISO†	Current (total)	Current (in area)	Pre-contact	(km^2)	(yr BP)
Anuta	aud	270	270	150	0.4	500
East Futuna	fud	3600	3600	2000	65	-
East Uvea	wls	10400	9620	4000	59	-
Emae	mmw	400	400	-	32	-
Ifira-Mele	mxe	3500	3500	-	1.5	400
Kapingamarangi	kpg	3000	1000	-	1.1	300
Mangareva	mrv	600	-	4000	15	970
Marquesas	mrq	6000	5390[i]	35000	1057	855
NZ Maori	mri	60660	60000	115000	501776	891
Nukuoro	nkr	1000	730	150	1.7	500
Penrhyn	pnh	200	200	-	9.84	730
Rarotongan	rar	39090	13100[ii]	15000[iii]	240	982
Rennellese	mnv	4390	-	-	60	600
Samoan	smo	364257	199000	80000	3134	3062
Sikaiana	sky	730	-	-	2	500
Tahitian	tah	68260	63000[ii]	45000	1536	982
Takuu	nho	1750	-	-	0.9	-
Tikopia	tkp	3320	-	1250	4.6	800
Vaeakau-Taumako	piv	1660	-	-	15	500
West Futuna	fut	1500	-	-	11	1000

† The ISO-639-3 Language Identification Code (ISO) is a unique identifier assigned to each language under the International Organisation for Standardisation. Current population size estimates are from Ethnologue.com: where given, we report both the population within the area and the total estimated number of speakers, including immigrant communities. Dates of establishment from archaeological estimates are given in years before present (yr BP)

[i] includes speakers of the language residing within French Polynesia
[ii] includes speakers of the language residing within the Cook Islands
[iii] includes Penrhyn and Pukapuka

Table S2. Comparisons of models of language evolution and likelihood ratio tests on the effect of population size on language evolution rates. Values for each model are the negative log maximum likelihood (*-lnL*), number of parameters (*k*), adjusted *AIC* for small sample size (*AICc*), and the *-lnL* of the corresponding null model that assumes no effect of population size on language evolution rates. Bold *-lnL* values indicate a significant effect of population size after Bonferroni correction. *AICc* values in bold indicate the best-fitting model for each language evolution rate.

Phylogenetic structure	Population growth	Population divergence	Founder effect	*-lnL*	*k*	*AICc*	Null *-lnL*
Gain							
Tip-wise	Constant	Fission	--	110.0	2	225.3	110.1
		Colonization	--	125.4	2	256.1	126.2
	Growth	Fission	--	107.3	4	228.3	110.1
		Colonization	--	123.6	4	260.9	126.2
Pair-wise	**Constant**	**--**	--	81.6	7	**205.2**	85.6
	Growth	Fission	--	81.6	9	271.2	85.6
		Colonization	--	81.7	9	271.4	85.6
Loss							
Tip-wise	Constant	Fission	--	85.3	2	175.9	85.5
		Colonization	--	86.7	2	178.7	91.2
	Growth	Fission	--	79.1	4	171.9	85.5
		Colonization	--	85.3	4	184.3	91.2
	Constant	Fission	Multiple	**54.5**	8	173.0	72.1
		Colonization	Multiple	50.9	8	165.8	51.1
Pair-wise	**Constant**	**--**	**--**	**46.4**	7	**134.8**	65.0
	Growth	Fission	--	**46.5**	9	201.0	65.0
		Colonization	--	**44.4**	9	196.8	65.0
	Constant	Fission	Single	**46.4**	8	156.8	65.0
		Colonization	Single	**37.4**	8	138.8	60.0
Total (gain + loss)							
Tip-wise	Constant	Fission	--	113.9	2	233.1	114.0
		Colonization	--	131.1	2	267.5	136.2
	Growth	Fission	--	112.1	4	237.9	114.0
		Colonization	--	134.6	4	282.9	136.2
	Constant	Fission	Multiple	**90.3**	8	244.6	103.8
		Colonization	Multiple	63.4	8	190.8	64.2
Pair-wise	**Constant**	**--**	**--**	**70.6**	7	**183.2**	78.5
	Growth	Fission	--	70.6	9	249.2	78.5
		Colonization	--	**69.6**	9	247.2	78.5
	Constant	Fission	Single	**70.6**	8	205.2	78.5
		Colonization	Single	**62.7**	8	189.4	69.6

Table S3: The relationship between island area and rates of word gain and loss from Polynesian language pairs.

Rate	Mean	s.e.	95 % CIs		R^2	Likelihood ratio
			Upper	Lower		
Gain	**0.26**	0.039	0.351	0.174	0.333	**53.1**
Loss	-0.01	0.017	0.033	-0.044	0.001	0.1
Total	**0.05**	0.015	0.081	0.012	0.079	**9.2**

Figure S1: Histograms of observed and expected numbers of total change (gains plus losses) of cognates from basic vocabulary in 10 language pairs under the best-fitting model (phylogenetically structured, constant population size, no founder effects). Plotted distributions show the expected probability of having a certain number of changes (gains or losses) in each language. Vertical lines show the observed numbers of gains or losses in each language. The language with the larger speaker population size is colored blue while the language with smaller population size is colored red. There is no significant association between population size and total change (see Table 1).

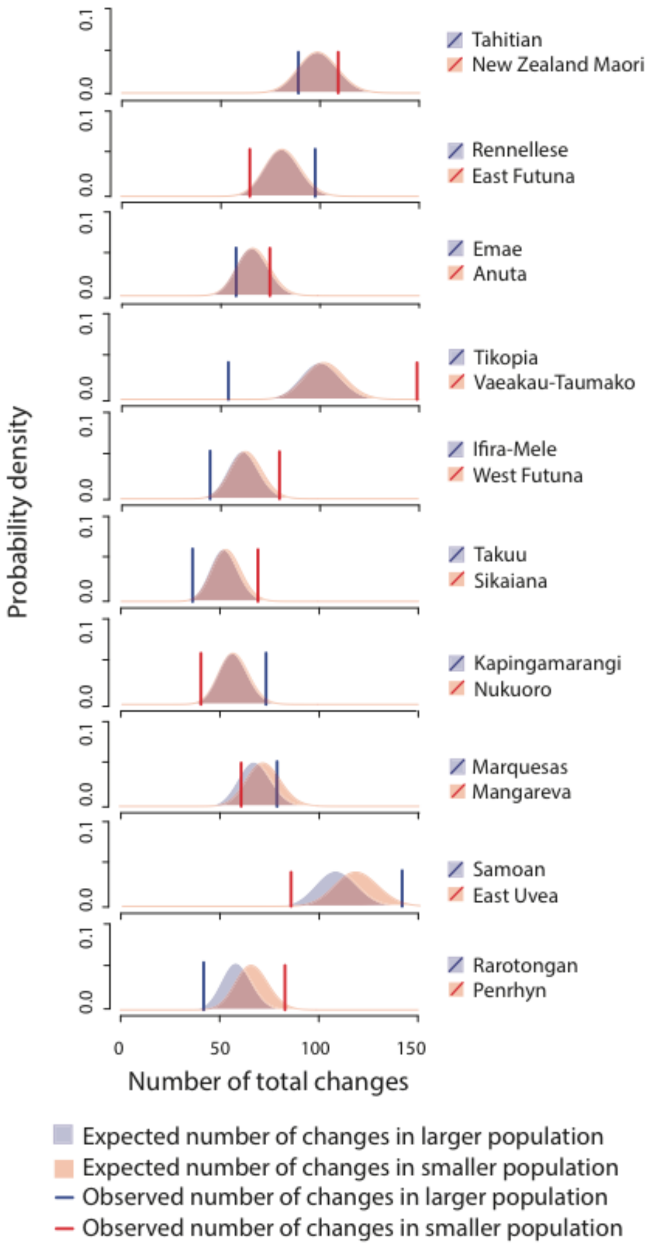

Figure S2: Log (ln) population size and number of loan words identified in the Austronesian Basic Vocabulary Database for the languages included in this study. There is no evidence of an association between population size and identified loan words, with or without the point on the extreme right of the graph (East Uvea, 10,400 speakers, 34 identified loan words in basic vocabulary).

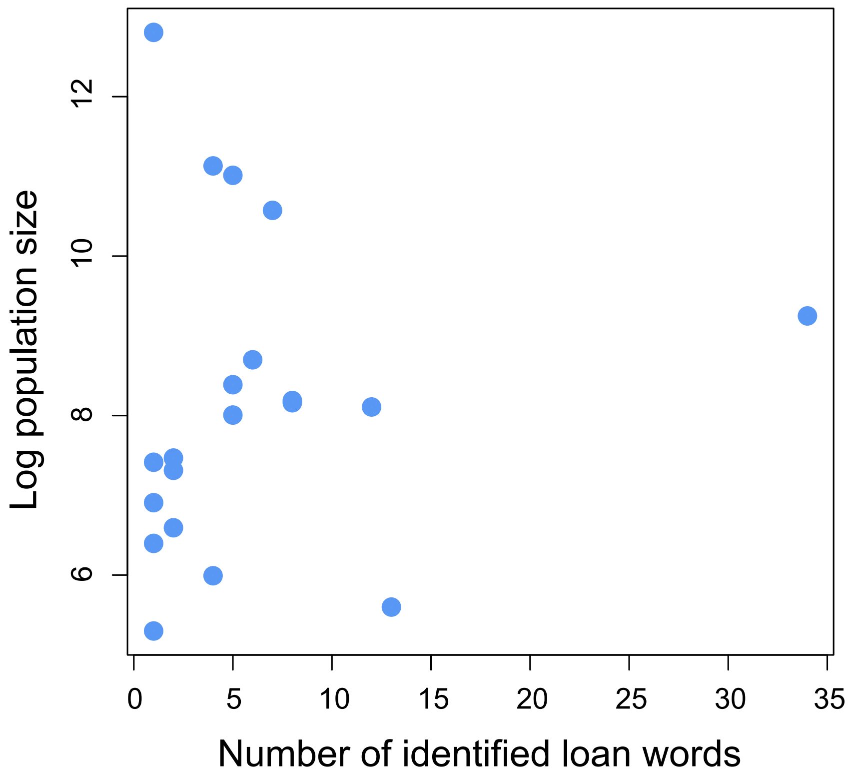

Figure S3: Illustration of the two modes of language origin modelled, and their relationship to the establishment dates of the two languages of the pair (t_A and t_B). For the fission model, the older date (t_A) provides the best estimate of date of divergence of the two languages in the pair. For the colonization model, the younger of the two dates (t_B) is the most appropriate.