Rate of language evolution is affected by population size

Lindell Bromham[a,1,2], **Xia Hua**[a,1], **Thomas G. Fitzpatrick**[b,1], **and Simon J. Greenhill**[b,c,1]

[a]Centre for Macroevolution and Macroecology, Division of Ecology, Evolution, and Genetics, Research School of Biology, Australian National University, Canberra, ACT 0200, Australia; [b]School of Culture, History and Language, ANU College of Asia and the Pacific, Australian National University, Canberra, ACT 0200, Australia; and [c]ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, ACT 0200, Australia

Edited by Tomoko Ohta, National Institute of Genetics, Mishima, Japan, and approved December 29, 2014 (received for review October 14, 2014)

The effect of population size on patterns and rates of language evolution is controversial. Do languages with larger speaker populations change faster due to a greater capacity for innovation, or do smaller populations change faster due to more efficient diffusion of innovations? Do smaller populations suffer greater loss of language elements through founder effects or drift, or do languages with more speakers lose features due to a process of simplification? Revealing the influence of population size on the tempo and mode of language evolution not only will clarify underlying mechanisms of language change but also has practical implications for the way that language data are used to reconstruct the history of human cultures. Here, we provide, to our knowledge, the first empirical, statistically robust test of the influence of population size on rates of language evolution, controlling for the evolutionary history of the populations and formally comparing the fit of different models of language evolution. We compare rates of gain and loss of cognate words for basic vocabulary in Polynesian languages, an ideal test case with a well-defined history. We demonstrate that larger populations have higher rates of gain of new words whereas smaller populations have higher rates of word loss. These results show that demographic factors can influence rates of language evolution and that rates of gain and loss are affected differently. These findings are strikingly consistent with general predictions of evolutionary models.

language evolution | sister-pair comparison | Austronesian | lexical change | Poisson regression

Population size can play a crucial role in the evolution of languages and cultures (1). However, opinions differ on both the possible mechanisms and the expected patterns (2–7). It has been suggested that larger populations will generate more innovations and are less prone to random loss of cultural elements (8–10), but may have less efficient diffusion of innovations than smaller populations (4). Alternatively, languages spoken by small isolated populations of speakers may have lower rates of loss if they maintain tighter cultural norms that improve transmission and resist change (11). Rates of change might be accelerated by founder effects when a new population is started from a small number of colonists, which could result in loss of elements from the ancestral language (11–13). Population size might also influence language complexity if small populations can develop greater linguistic complexity (11), whereas large, widespread languages that are often learned by adults may become simplified (14). Conversely, it has been suggested that the average rate of word turnover is essentially the same in all languages (15–17), or that it is determined primarily by other factors such as language contact (6, 18).

Uncovering systematic patterns of rates of language change may reveal underlying mechanisms of language evolution (13, 19). In particular, investigating rates of language change can demonstrate whether language evolution follows the predictions of population genetic models (20). Systematic variation in the rates of language change may affect attempts to reconstruct the history of human cultures from comparative language data, particularly the estimation of time (21). Theoretical modeling has been used to explore the consequences of population size on rates of language evolution; however, such models unavoidably require prior assumptions about the way language change diffuses through populations (4, 6). Tests of population size effects have so far been equivocal and have been limited by the availability of appropriate data and methodology (3, 7, 22). In particular, similarities and differences across languages cannot be compared as if they were statistically independent data points, because closely related languages are expected to be similar in many aspects. This hierarchical pattern of similarities can confound attempts to find causal correlations between aspects of human language and culture by creating spurious correlations, a methodological challenge sometimes referred to as Galton's problem (23, 24). The comparison of rates of evolutionary change in different lineages presents additional challenges because we need to be able to infer the number of evolutionary changes that have occurred along each lineage (25, 26). Because rates are based on count data, the accuracy of rate estimates will depend not only on the amount of data compared between the languages but also on the time since their divergence (26). It is important to test these hypotheses against observations from real languages within a statistically robust framework that controls for the effect of phylogeny on covariation and explicitly deals with the effect of time of divergence on rate estimation.

The Polynesian languages provide an excellent test case for examining the effects of population size on rates of language evolution (Fig. 1). They form a distinct lineage within the Austronesian language family, one of the largest language groups in the world, and are well-documented in comparative language databases (27, 28). They arise from a relatively recent history of colonization (29, 30), and the relationships between languages have been extensively investigated (31). Polynesian languages have well-defined areas of occupation, and establishment dates for most language groups

Significance

Evolutionary methods are increasingly being applied to investigating linguistic change. But does language change conform to the predictions of evolutionary theory? Here, we use data from closely related pairs of languages to show that a key prediction of evolutionary theory is met: rates of gain of new words are higher in larger populations whereas rates of word loss are greater in small populations. Our analysis provides, to our knowledge, the first statistically robust evidence of an influence of population size on rate of language change. These results demonstrate the potential for demographic factors to influence language evolution.

Author contributions: L.B., X.H., T.G.F., and S.J.G. designed research; T.G.F. and S.J.G. performed research; X.H. analyzed data; and L.B., X.H., and S.J.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]L.B., X.H., T.G.F., and S.J.G. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: lindell.bromham@anu.edu.au.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1419704112/-/DCSupplemental.

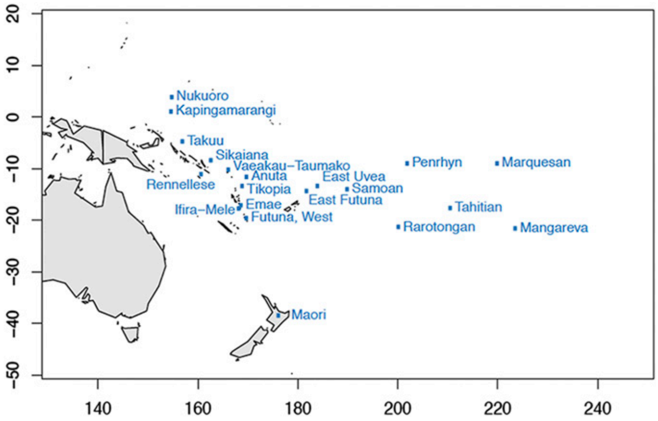


Fig. 1. Map of Polynesian languages included in this study.

can be derived from archaeological data (32). By comparing rates in recently established sister languages occupying similar habitats, we avoid potential confounding social and environmental variables. Due to this combination of factors, there is no better "natural experiment" to test the effect of population size on rates of language evolution. Although the Polynesian languages form only a small subset of the Austronesian language family, there are no other language groups that would allow us to control for age, area, and relationships to the extent we can with the recent Polynesian radiation.

To control for statistical nonindependence due to shared history, we based our analyses on phylogenetically independent sister pairs—that is, we considered differences between pairs of languages that are likely to be each other's closest relatives (31, 33, 34). This approach allowed us to compare changes that have happened in each language since they split from their common ancestral language. Barring borrowing, the changes in cognate sets we counted in each language of the pair occurred independently of changes occurring in any other language in the dataset and so represent statistically independent instances of word gain and loss. However, we formally tested the phylogenetic signal in the data and also analyzed the data without the assumption of phylogenetic structure.

The sister-pairs approach is more appropriate to these data than attempting to estimate rates of change on all edges of a phylogeny. A "whole phylogeny" approach would require branch lengths that are proportional to time. Although we have good archaeological dates for the establishment of most of the populations in our study, we could not extrapolate these dates to the internal divergences in the phylogeny without making prior assumptions about the rates of language change over time (29). Similarly we have carefully selected sister pairs with known relationships, whereas a whole-tree approach would require us to make assumptions about the more distant relationships between the ancestral lineages that gave rise to these contemporary populations. We have selected only pairs of languages where the relationships are well established and there is no evidence of replacement or blending of populations, avoiding populations with complicated settlement histories. Thus, the sister-pairs approach represents a conservative approach to the data that minimizes the assumptions that we must make about the history of the languages and the mechanisms of population and language change.

Our analysis measured one specific aspect of language change, the gain and loss of cognate (homologous) words of basic vocabulary (27). We compared rates of word gain and loss in pairs of closely related languages, with estimates of current population sizes as well as historical population sizes estimated at the time of European contact. The age of each language was derived from archaeological data, and area inhabited by summing the land area of all islands on which the language is traditionally spoken. We estimated the relative number of word gains and losses in each language of each pair by comparing the presence or absence of cognate terms (words related by descent from a common ancestral form) in 210 semantic units in basic vocabulary, such as kinship terms, body parts, and numbers (27). By using cognate terms for basic vocabulary items, we could compare terms that were homologous across languages and had retained a common meaning. Restricting our analysis to basic vocabulary items helped to limit the problem of semantic shift, where cognate terms take on new meanings in some descendant languages. Basic vocabulary is considered to be relatively resistant to borrowing (35). It also helped ensure that we had comparable data for each of our language pairs: each of the languages had nearly all of the basic vocabulary items recorded but had varying amounts of general lexicon available in other databases (28). Furthermore, because rate of word change has been linked to frequency of use (19), comparing the same set of semantic units across all languages in the study helped prevent bias in rates due to word selection.

For each language, we counted as a gain any word that had no identified cognates in any other language in the database, representing adoption of a new word for one of the basic vocabulary items. Any cognate set present in one language of the pair but not the other, and which was recognized in at least one other language in the family, was considered to have been present in the common ancestor of the pair then lost in one language (Fig. 2). Gain and loss rates were estimated separately because the addition of a new word does not necessarily involve the loss of an existing word because languages can have multiple lexemes (word items) per semantic unit (basic vocabulary category). We did not include borrowed words or language pairs known to have a high rate of borrowing.

We used Poisson regression to test the effect of population size on the rate of gain, loss, and total change (see *SI Appendix* for details). We modeled rates as linear functions of population size on log–log scales to confine rates to positive values (where all logs are natural logs, or $\log_e$). Numbers of cognate gains and losses were standardized by the total number of semantic units compared in each pair of languages. We applied the Akaike information criterion with correction for small sample size

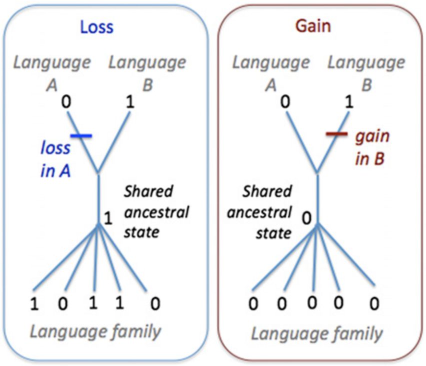


Fig. 2. Coding scheme for counting word gains and losses in a pair of sister languages, A and B. The presence (signified by a 1) of a cognate in one member of the pair and any members of the language family indicates the presence of that cognate in the shared ancestor of the pair. Absence (signified by a 0) of that cognate in the other member of the pair is evidence that it has been lost from that language since it split from the other language. Any word (lexeme) present in one language but with no cognates in any other language in the family is considered to have been gained in that semantic category in that language since it split from the other member of the pair and so is counted as a word gain in that language.

Table 1. Significance tests for the effect of population size on rate of word gain and loss

Rate	Mean	SE	Upper 95% CI	Lower 95% CI	R^2	Power	Likelihood* ratio
Gain	**0.29**	0.064	0.435	0.145	0.144	0.68	**22.8**
Loss	**−0.12**	0.032	−0.048	−0.194	0.186	0.69	**14.5**
Total	−0.03	0.028	−0.092	0.033	0.010	0.08	1.2

*Best-fitting model: phylogenetically structured analysis based on the colonization model using a pair-wise approach with 10 language pairs. Total is the sum of estimated gains and losses for each language. The mean, SE, and 95% confidence intervals (CIs) of regression coefficients, as well as R^2, power of the test, and likelihood ratio against a null model (where population size has no effect), are reported. The values in bold indicate a significant effect of population size on the rate of language change.

(AICc) to test the fit of alternative models of language evolution to the observed data. The models varied in (*i*) whether the data were phylogenetically structured or not, (*ii*) whether new populations rapidly achieved a stable population size upon colonizing a new area, or underwent a more protracted period of growth, (*iii*) whether new languages formed by an allopatric process, where a parent population underwent fission to produce two daughter populations, or by a peripatric process, where the parent population gave rise to a daughter population by a process of colonization, and (*iv*) whether or not new languages experienced a founder effect, losing a number of words when a new population became established (*Methods*). We estimated the power of Poisson regression to detect the effect of population size on the rate of gain, loss, and total change, at the significance level of 0.05 (36).

Results

Our analyses revealed two significant patterns in the data: languages that had larger speaker populations tended to have higher rates of gain of new words in basic vocabulary than their smaller sister languages, and smaller populations tended to have higher rates of word loss from basic vocabulary (Table 1 and Fig. 3). Variation in population size explained about 10–20% of the variation in word gains and losses (Table 1). Because gain and loss showed opposing patterns, overall rates of change (gain plus loss) were not significantly related to population size (Table 1 and *SI Appendix*, Fig. S1).

We found that both current and precontact population sizes were strongly predicted by area (Fig. 4). Accounting for phylogenetic structure provided a better fit to the data than a nonhierarchical model where each language group would be considered to be an independent data point (*SI Appendix*, Table S2). Allowing for a period of population growth does not provide a better fit to the data, nor does accounting for founder effects. Therefore, the results presented here are from the best-fitting model: the phylogenetically structured analysis assuming stable population size over time and no founder effect.

Our data suggest that the size of the speaker population of each language is determined primarily by the area available (*SI Appendix*, Table S3). Because we do not have all measures of population size for each language group, each analysis has different numbers of data points (for example, 11 populations have precontact population size and island size, thus 9 degrees of freedom, and 20 populations have present population size and island size, thus 18 degrees of freedom). Precontact population size is strongly correlated with language area ($t_9 = 6.13$, $P = 1.7 \times 10^{-4}$), suggesting that precontact populations were typically at a carrying capacity. Current population size is also strongly correlated with both precontact population size ($t_9 = 4.84$, $P = 9.2 \times 10^{-4}$) and island size ($t_{18} = 5.36$, $P = 4.3 \times 10^{-5}$). The finding that population size is linked to area, together with the result that accounting for the period of population growth does not provide a better fit to the data, is consistent with evidence suggesting high rates of population growth in Polynesia, which in some cases led to the introduction of cultural mechanisms for curtailing population growth (37).

The pattern of higher rates of word gain in larger populations was confirmed when the Poisson regression analysis was repeated using only the pairs that had estimates of precontact population size (mean = 0.30, likelihood ratio = 53.3) and those that had data on current population size "in area," not including speakers living in migrant populations on other islands (mean = 0.67, likelihood ratio = 30.2). However, these reduced datasets did not have a significantly higher rate of loss in smaller populations (precontact, mean = −0.00, likelihood ratio = 0.0; current, mean = −0.05, likelihood ratio = 3.0). Given the strong correlation between island size and population size, we also repeated the Poisson regression analysis using island size instead of current population size: these results confirmed higher rates of word gain in larger populations, but not higher rates of loss in smaller populations (*SI Appendix*, Table S3).

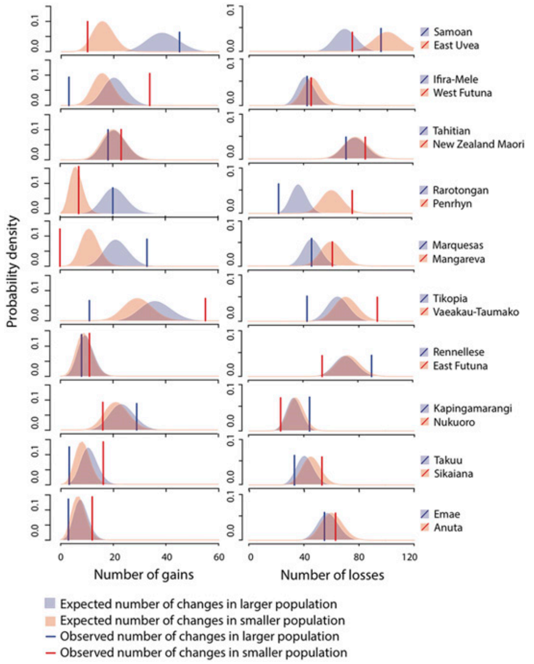


Fig. 3. Histograms of observed and expected numbers of gains and losses of cognates from basic vocabulary in 10 language pairs under the best-fitting model (phylogenetically structured, constant population size, no founder effects). Plotted distributions show the expected probability of having a certain number of gains or losses for each language, given the model fitted to all data points. Vertical lines show the observed numbers of gains or losses in each language. The language with the larger speaker population size is colored blue whereas the language with smaller population size is colored red, and the panels are ordered from the pair with the oldest divergence at the top to the youngest at the bottom. More often than expected by chance, the smaller population has a higher rate of word loss and the larger population has a higher rate of gain, even though specific circumstances in particular languages may override this general trend.

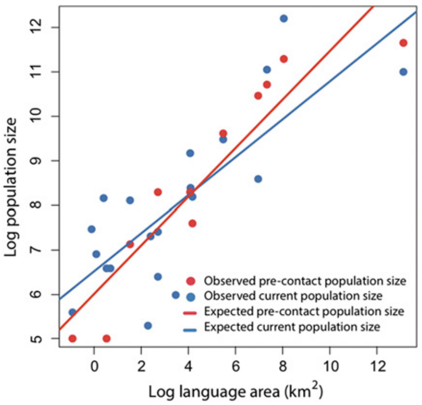


Fig. 4. Relationship between log population size and log language area for the languages included in this study, where logs are the natural logs (ln, log$_e$), for both current population size and historical population time estimated at the time of European contact (pre-contact).

Discussion

We found that Polynesian languages with larger speaker-population sizes had higher rates of gain of new words than their smaller sister languages, consistently across all analyses. We also found that languages with a smaller number of speakers had higher rates of loss of lexemes from basic vocabulary items although this result was not always significant in the reduced datasets. Total rate of change (gain plus loss) was not significantly related to population size.

Our results have a number of important implications for the study of language evolution. Most importantly, we demonstrate that population size can influence the patterns and rates of language change. Our results do not support the hypothesis that smaller populations gain new words rapidly due to greater rate of uptake of innovation. Instead, our analyses suggest that rates of gain and loss of basic vocabulary items may, in some ways, conform to some general predictions of evolutionary models. Population genetic theory predicts that larger populations should have higher rates of adaptation because there are more individuals to generate novelty and fewer disruptive influences of random sampling on the process of fixation of new variants. The observation that rate of gain of new words is greater in larger populations is compatible with the establishment of new words by a process analogous to positive selection. An alternative explanation for this pattern is that smaller populations might have denser social networks that tend to be conservatizing, enforcing linguistic norms and resisting change (5, 11, 38).

Our finding that small populations have greater rates of word loss is compatible with the expectation from evolutionary theory that small populations are prone to loss of diversity due to the chance of incomplete sampling of variants each generation. Although lost words could be replaced by novel innovations, the lower rate of gain in smaller populations suggests that the loss of cognates in these languages is not due to higher rates of word turnover. Instead, the pattern of loss is compatible with stochastic fluctuation in frequency of lexeme use (akin to the process of genetic drift): in a small population, such fluctuations are more likely to go to zero, resulting in the loss of one lexeme for that unit of basic vocabulary (39).

These results suggest that languages with consistently small speaker populations undergo a greater rate of word loss from basic vocabulary. This process is distinct from the founder effect, which predicts loss of diversity associated with colonization events (11, 12). Although a small population of founders cannot carry all of the genetic alleles of a large parent population, it may be possible for a small number of speakers to use all of the basic vocabulary from the parent language (40). Despite the likely establishment of new Polynesian languages from relatively small colonizing populations, we do not find that a founder model describes our data better than one of gradual loss of words.

If the finding that rates of word gain are greatest in large populations is a general feature of language change, then it has implications for understanding the connection between rate of language change and the diversification of language families. Higher rate of language change has been reported in diverse language groups, based on word turnover rates (13). One possible explanation for this association is that change is highest when populations are divided. However, our results suggest that small subdivided populations may be more likely to undergo word loss than accelerated gain of new words. More work is needed to establish both the generality and causes of these patterns.

Methods

Comparisons. Pairs of sister languages were chosen based on established relationships (31, 33, 41), ascertained using the linguistic comparative method, which uses patterns of language change, such as systematic sound correspondences, to identify language relationships (42). We did not include any comparisons between languages that are considered to have a high degree of borrowing and contact: for example, the languages of Tonga, Niue, and Pukapuka. Any undetected loan words in basic vocabulary will reduce the apparent rate of loss and gain because loan words will be counted as retentions, reducing the power of the tests to detect differences in rates. Borrowing could bias our results only if it is systematically affected by population size. For example, if smaller languages are more prone to borrowing than their larger sister languages, then we might detect fewer losses in smaller populations, if losses are replaced by reborrowing from the sister language. Borrowing should not affect the gain rate, which counts only words with no cognates recorded in any other language in the family. A plot of the number of identified loan words against population size for these languages provides no evidence that loan rate is influenced by population size (*SI Appendix*, Fig. S2).

Population Data. We included several measures of population size. Contemporary population size for each language was taken from the Ethnologue (43), representing the number of speakers for each language group, estimated between 1993 and 2011 (*SI Appendix*, Table S1). We recorded both the total number of speakers reported in the Ethnologue and also the "in area" population of speakers found within the language area, excluding speakers in migrant communities established outside the area. We also included, where available, estimates of population size at the time of first contact with Europeans (32). These "precontact" estimates were included to represent the likely population size of each language group before changes in population size brought about by contact with Europeans (37). The languages included in our analysis ranged in population size from 200 to over 360,000 speakers.

We estimated the area of each language included in our analysis (Fig. 1) from the sum of the area of the islands on which the language was traditionally spoken (hereafter referred to as "language area"). Language area was derived primarily from published sources (32, 44), with additional information from the United Nations Environment Program (UNEP) "Island Directory" (islands.unep.ch). Distance between each pair was calculated using the "Distance Measurement Tool" function of Google Maps to estimate the shortest coast-to-coast distance. However, this distance was not found to have any explanatory power in preliminary data analyses and so was not included in final analyses.

To calculate absolute rates of change, we sought establishment dates for each language, based on archaeological evidence of settlement in each language area (30, 32, 45). Because we were concerned with the establishment of separate language groups, we took the conservative approach of taking the first evidence of permanent, continuous settlement, rather than the earliest evidence of human presence. For example, evidence of human presence on Kapingamarangi dates to 1,000 y B.P., but archaeologists suggest that the island was only sporadically occupied between 700 y B.P. and 300 y B.P., after which stabilization of the islet allowed permanent occupation (46). We therefore took 300 y B.P. as the establishment of the population of speakers of the Kapingamarangi language. We have included only language groups considered to have a continuous history in the current area, without replacement by other language groups. For this reason, we excluded languages with complicated settlement histories, such as Fiji and Rotuma.

Counting Word Gains and Losses. For this study, we considered only word gain and loss, not other forms of language evolution such as phonological change, morpho-syntactic change, and semantic change (12, 14, 47). We used the Austronesian Basic Vocabulary Database (27), which records up to 210 basic semantic units for more than 1,100 languages spoken in the Pacific region. By using basic vocabulary, we ensured that cognate terms not only had a common history but a common meaning across language comparisons.

We considered each of the 210 identified basic vocabulary sets as semantic units. So, for example, one semantic unit is "one hundred," which may be represented by different words in different languages (e.g., *rau* in Anuta, *gau* in Rennellese, *vahiki* in Vaeakau-Taumako). We used the term "cognate set" to represent a set of lexical units that are clearly related by descent and have been identified by linguists as being derived from a common ancestral word. A classic example is the word for "five," for which many Polynesian languages share related terms, such as *lima* in Fijian, *nima* in Tongan, *gima* in Rennellese, *'ima* in Marquesan, and *rima* in New Zealand Maori. In this case, the cognate set was those words that are recognizably descended from the ancestral word, reconstructed as the protoform **lima* in the ancestral Proto-Malayo-Polynesian language (27). When we refer to a word in one language having a cognate in another language, we mean that both languages contain words from the same cognate set in the same semantic unit. We will refer to the particular version of the word found in each language as "lexemes." A language might have more than one lexeme for a given semantic unit. For example, the Marquesan language has three alternative lexemes in the semantic unit "good" (*'eka, meitai, kanahau*) whereas the Mangaraven language has two recorded lexemes for "good" (*reka, meitetaki*).

Our approach to estimating rates of language change differed from other studies of language evolution. The standard approach for investigating rates of change in historical linguistics is to calculate "retention rates" within defined lists of basic vocabulary (15). The retention rate—or the number of shared cognates—provides a useful indicator of the amount of change between a language and its ancestor (48). However, the retention rate does not distinguish between gains and losses and thus cannot allow comparison of these two processes. Another approach has been to quantify lexical distance between words using a Levenshtein distance metric to quantify how different two languages are (6). However, the Levenshtein distance does not distinguish between true homology—cognate words—and chance similarity (49). A third approach is to use computational phylogenetic methods to estimate the rate of change on the branches of a phylogeny (50, 51). Although the sister-pairs approach lets us constrain the date of origin of sister lineages, the "whole tree" approach would require us to solve dates, rates, and branch lengths simultaneously. Our approach overcame these limitations by identifying changes in homologous word sets, estimating separate rates of word gain and loss, and using independently established relationships between languages to correct for phylogenetic nonindependence. Our aim was not to establish the relationships between languages or their general levels of similarity. Instead, we considered the presence or absence of cognates on a pair-wise basis to localize the gain or loss of particular words to the history of each language.

Consider a pair of languages, A and B, that are each other's closest relatives, such that they share a more recent common ancestor with each other than either does with any other member of the language family (Fig. 2). We want to be able to count the relative number of word gains and losses that have occurred in each language since they shared a common ancestral language. We consider that, if a lexeme present in either of these languages has a cognate in at least one other member of the language family, then it must have been retained from the ancestral language (assuming that two independent gains of the same cognate are unlikely and that borrowing has been eliminated from the database). If a cognate is present in language A but not in B, then we assume it has been retained in A but lost in B. Loan words have been removed from the database so we discount the possibility that the cognate was lost from the common ancestor and then regained in language A. Although the database may contain unidentified loan words, our analysis is based on items of basic vocabulary that are more resistant to being borrowed (52). Moreover, the most likely source of loan words is a language's sister (2, 53), the effect of which would be to weaken the signal in our data rather than create false patterns. Simulation studies of borrowing suggest that any unidentified loan words would make the sister languages seem more similar than they actually are by masking innovations or losses (53).

If language A contains a lexeme in a semantic unit in the word list that has no identified cognates in language B or in any other language in the family, then it is considered to be a novel invention in language A after it split from language B. Although we cannot discount the possibility that it was gained in the common ancestor of the pair and then lost in B, this explanation is less parsimonious. Note that gain and loss in a given sematic unit may be either linked or unlinked: a new lexeme may be added without the loss of an existing lexeme, or it may replace an existing cognate with a new lexeme (in which case it will be counted as both a loss of one cognate and a gain of a novel word).

This approach provides two rates-informative patterns in the data: cognates present in one member of the pair and not the other (evidence of loss) and noncognate terms present in one language but not in any other language in the family (evidence of gain) (Fig. 2). This approach is somewhat similar to the Tajima test (54, 55) in that it considers patterns of shared characters in the data to compare the rate of change in each member of a pair of lineages with respect to an outgroup. We took a conservative approach to counting word changes, excluding any lexemes marked as possible borrowings or those that seemed to show cognacy with other lexemes in the same semantic unit even if not labeled as such in the database. Semantic units where one member of the pair had no recorded lexemes were recorded as "missing" and did not contribute to analysis of rates.

The number of lexemes compared varied between pairs because not all languages have the full set of 210 semantic units recorded in the Austronesian Basic Vocabulary Database (27) and languages differ in the number lexemes per semantic units. In all analyses, numbers of gains, losses, and total changes (gains plus losses) between languages in a pair were standardized by their total number of semantic units compared.

Statistical Analysis. Because languages evolve through a process of descent with modification, we expect them to show phylogenetic signal, such that closely related languages will show similarities due to features inherited from their shared common ancestor (56). The process of descent leaves a pattern of hierarchical similarity in the data that creates statistical nonindependence in language features. Here, we accounted for phylogenetic signal by applying a pairwise approach by fitting different intercepts of language evolving rates for different pairs of languages. We refer to this approach as "pair-wise."

However, it may be the case that the specific traits we are interested in—population size and rates of language change—do not always show hierarchical patterns of similarity. If we assume that all of the Polynesian populations included in this study have the same intrinsic population growth rate, then population size may be largely determined by the available area for each population (Fig. 4) rather than by heritable features of each cultural group. Similarly, because Polynesian languages are broadly similar in structure, the determinants of language change may be effectively independent of the state of the common ancestor and largely a function of the particular circumstances of each particular group. In this case, we might consider each language group to be an independent experiment in language evolution, triggered by the establishment of a new population on an uninhabited island. To allow for this possibility, we also analyzed by fitting a common intercept for all of the pairs. We refer to this approach as "tip-wise" because it considers each separate "tip" of the phylogeny as an independent data point.

Polynesia was largely peopled through relatively small numbers of colonists making ocean voyages and establishing new settlements on uninhabited islands (37, 57). Therefore, we expect that each language group was started by a relatively small founding population that then grew in size. However, it is not clear whether the small founder population would have a significant impact on the estimation of the overall average rate of language change (*Discussion*). To account for this uncertainty, we compared different models: the "constant" model assumes that the founding population rapidly grew to the carrying capacity of the area and then stabilized, and the "growth" model assumes a continuous density-dependent population growth, for which a common population growth rate and initial population size are fitted to all language pairs. Note that the constant model under the pair-wise approach does not require absolute date estimates

We considered two alternative models for the origination of a new language: "fission," where an ancestral population splits to form two daughter populations, and "colonization," where a small number of colonists from a parent population establish a single daughter population (*SI Appendix*, Fig. S3). Fission is analogous to allopatric speciation whereas colonization is analogous to peripatric speciation. Clearly, there is no clear line between fission and colonization because the difference between the two models lies in the relative change in population size from the parent population to the founder population. However, we made this distinction because the two models differed in the way we used establishment dates to estimate age of pairs. Under a fission model, we considered the older date of the two establishment dates to most closely approximate the age of the split between the two languages. However, it is likely to often be the case for this dataset that new languages are established through colonization (29, 30), and therefore the age of establishment of the parent population overestimates the date of the split between the parent and daughter populations. So under the colonization model we used the younger of the two dates to approximate the time since the two languages started accumulating differences from each other.

In addition to the gradual loss of words over time as a language evolves, there may be an initial loss of lexemes when a population is founded, if the founding population does not use all of the lexemes used in the parent language. To model this initial sampling effect on word loss, we implemented additional Poisson regressions under the founder scenario where words were lost due to sampling error. For the pair-wise approach, a single estimate of the absolute number of initial losses was fitted to all language pairs under the fission scenario, or to just the daughter population in the colonization scenario (because the parent population is assumed to undergo no founder effect at the time when its daughter population establishes). For the tip-wise approach, multiple estimates of the initial losses were fitted to each language, with each pair of languages sharing an estimate in the fission model. Details of the models of language and population change are given in *SI Appendix*.

We applied likelihood ratio tests with Bonferroni correction to each of the scenarios against their null models that assume no effect of population size on language-evolving rates. We applied the Akaike information criterion with correction for small sample size (AICc) to examine which scenario best fits the observed changes in language evolution. Confidence intervals of regression coefficients between population size (or island size) and rates of language change were derived from the information matrix. Effect size was calculated as the pseudo R^2 measures for Poisson regression (58). Statistical power was estimated using the R package *asypow* (36).

ACKNOWLEDGMENTS. We are grateful to Russell Gray, Andrew Pawley, and Bethwyn Evans for feedback and thank Adrian Bell for providing archaeological dates. We thank the Research School of Asia and the Pacific at the Australian National University and the Australian Research Council for support.

1. Levinson SC, Gray RD (2012) Tools from evolutionary biology shed new light on the diversification of languages. *Trends Cogn Sci* 16(3):167–173.
2. Nichols J (1997) Modeling ancient population structures and movement in linguistics. *Annu Rev Anthropol* 26:359–384.
3. Moran S, McCloy D, Wright R (2012) Revisiting population size vs. phoneme inventory size. *Language* 88(4):877–893.
4. Nettle D (1999) Is the rate of linguistic change constant? *Lingua* 108(2):119–136.
5. Bowern C (2010) Correlates of language change in hunter-gatherer and other 'small' languages. *Lang Linguist Compass* 4(8):665–679.
6. Wichmann S, Holman EW (2009) Population size and rates of language change. *Hum Biol* 81(2-3):259–274.
7. Nettle D (2012) Social scale and structural complexity in human languages. *Philos Trans R Soc Lond B Biol Sci* 367(1597):1829–1836.
8. Henrich J (2004) Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses: The Tasmanian case. *Am Antiq* 69(2):197–214.
9. Kline MA, Boyd R (2010) Population size predicts technological complexity in Oceania. *Proc Biol Sci* 277(1693):2559–2564.
10. Collard M, Ruttle A, Buchanan B, O'Brien MJ (2013) Population size and cultural evolution in nonindustrial food-producing societies. *PLoS ONE* 8(9):e72628.
11. Trudgill P (2004) Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguist Typol* 8(3):305–320.
12. Atkinson QD (2011) Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332(6027):346–349.
13. Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M (2008) Languages evolve in punctuational bursts. *Science* 319(5863):588.
14. Lupyan G, Dale R (2010) Language structure is partly determined by social structure. *PLoS ONE* 5(1):e8559.
15. Swadesh M (1952) Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proc Am Philos Soc* 96(4):452–463.
16. Rama T (2013) Phonotactic diversity predicts the time depth of the world's language families. *PLoS ONE* 8(5):e63238.
17. Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. *Int J Am Linguist* 21(2):121–137.
18. Thomason SG, Kaufman T (1988) *Language Contact, Creolization, and Genetic Linguistics* (Univ of California Press, Oakland, CA).
19. Pagel M, Atkinson QD, Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163):717–720.
20. Mesoudi A, Whiten A, Laland KN (2006) Towards a unified science of cultural evolution. *Behav Brain Sci* 29(4):329–347, discussion 347–383.
21. Gray RD, Atkinson QD, Greenhill SJ (2011) Language evolution and human history: What a difference a date makes. *Philos Trans R Soc Lond B Biol Sci* 366(1567):1090–1100.
22. Wichmann S, Stauffer D, Schulze C, Holman EW (2008) Do language change rates depend on population size? *Adv Complex Syst* 11(3):357–369.
23. Roberts S, Winters J (2013) Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLoS ONE* 8(8):e70902.
24. Ladd DR, Roberts SG, Dediu D (2015) Correlational studies in typological and historical linguistics. *Annu Rev Linguistics* 1(1):4.1–4.21.
25. Lanfear R, Welch JJ, Bromham L (2010) Watching the clock: Studying variation in rates of molecular evolution between species. *Trends Ecol Evol* 25(9):495–503.
26. Welch JJ, Waxman D (2008) Calculating independent contrasts for the comparative study of substitution rates. *J Theor Biol* 251(4):667–678.
27. Greenhill SJ, Blust R, Gray RD (2008) The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evol Bioinform Online* 4:271–283.
28. Greenhill SJ, Clark R (2011) POLLEX-Online: The Polynesian lexicon project online. *Oceanic Linguist* 50(2):551–559.
29. Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913):479–483.
30. Wilmshurst JM, Hunt TL, Lipo CP, Anderson AJ (2011) High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proc Natl Acad Sci USA* 108(5):1815–1820.
31. Marck JC (2000) *Topics in Polynesian Language and Culture History* (Pacific Linguistics, Canberra, Australia).
32. Kirch PV (1984) *The Evolution of the Polynesian Chiefdoms* (Cambridge Univ Press, Cambridge, UK).
33. Pawley A (1967) The relationships of Polynesian Outlier languages. *J Polyn Soc* 76(3): 259–296.
34. Pawley A, Ross M (1995) The prehistory of Oceanic languages: A current view. *The Austronesians: Historical and Comparative Perspectives*, eds Bellwood P, Fox JJ, Tryon D (Australian National University, Canberra, Australia), pp 43–80.
35. Tadmor U, Haspelmath M, Taylor B (2010) Borrowability and the notion of basic vocabulary. *Diachronica* 27(2):226–246.
36. Brown BW, Lovato J, Russel K, Halvorsen KB (2012) Asypow: Calculate power utilizing asymptotic likelihood ratio methods. R package version 2012.04-1. Available at CRAN. R-project.org/package=asypow.
37. Kirch PV, Rallu J-L (2007) *The Growth and Collapse of Pacific Island Societies: Archaeological and Demographic Perspectives* (Univ of Hawaii Press, Honolulu).
38. Trudgill P (2011) Social structure and phoneme inventories. *Linguist Typol* 15(2): 155–160.
39. Reali F, Griffiths TL (2010) Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. *Proc Biol Sci* 277(1680):429–436.
40. Hunley K, Bowern C, Healy M (2012) Rejection of a serial founder effects model of genetic and linguistic coevolution. *Proc Biol Sci* 279(1736):2281–2288.
41. Pawley A (1966) Polynesian languages: A subgrouping based on shared innovations in morphology. *J Polyn Soc* 75(1):39–64.
42. Durie M, Ross M, eds (1996) *The Comparative Method Reviewed: Regularity and Irregularity in Language Change* (Oxford Univ Press, Oxford).
43. Lewis MP, Simons GF, Fennig CD, eds (2013) *Ethnologue: Languages of the World* (SIL International, Dallas), 17th Ed. Available at www.ethnologue.com.
44. Næss Å, Hovdhaugen E (2011) *A Grammar of Vaeakau-Taumako* (Walter de Gruyter, Göttingen, Germany).
45. Nunn PD (2007) *Climate, Environment and Society in the Pacific in the Last Millennium* (Elsevier, Oxford).
46. Kirch PV (1984) The Polynesian outliers: Continuity, change, and replacement. *J Pac Hist* 19(4):224–238.
47. Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011) Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473(7345):79–82.
48. Blust R (2000) Why lexicostatistics doesn't work: The 'universal constant' hypothesis and the Austronesian languages. *Time Depth in Historical Linguistics*, eds Renfrew C, McMahon A, Trask L (McDonald Institute for Archaeological Research, Cambridge, UK), pp 311–331.
49. Greenhill SJ (2011) Levenshtein distances fail to identify language relationships accurately. *Comput Linguist* 37:689–698.
50. Pagel M, Meade A (2006) Estimating rates of lexical replacement on phylogenetic trees of languages. *Phylogenetic Methods and the Prehistory of Languages*, eds Forster P, Renfrew C (McDonald Institute for Archaeological Research, Cambridge, UK).
51. Greenhill SJ, Atkinson QD, Meade A, Gray RD (2010) The shape and tempo of language evolution. *Proc Biol Sci* 277(1693):2443–2450.
52. Greenhill SJ, Gray RD (2012) Basic vocabulary and Bayesian phylolinguistics: Issues of understanding and representation. *Diachronica* 29(4):523–537.
53. Greenhill SJ, Currie TE, Gray RD (2009) Does horizontal transmission invalidate cultural phylogenies? *Proc Biol Sci* 276(1665):2299–2306.
54. Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135(2):599–607.
55. Bromham L, Penny D, Rambaut A, Hendy MD (2000) The power of relative rates tests depends on the data. *J Mol Evol* 50(3):296–301.
56. Mace R, Pagel M (1994) The comparative method in anthropology. *Curr Anthropol* 35(5):549–564.
57. Murray-McIntosh RP, Scrimshaw BJ, Hatfield PJ, Penny D (1998) Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. *Proc Natl Acad Sci USA* 95(15):9047–9052.
58. Waldhör T, Haidinger G, Schober E (1998) Comparison of R^2 measures for Poisson regression by simulation. *J Epidemiol Biostat* 3:209–215.

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:

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

$\lambda_1(t)$ 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).

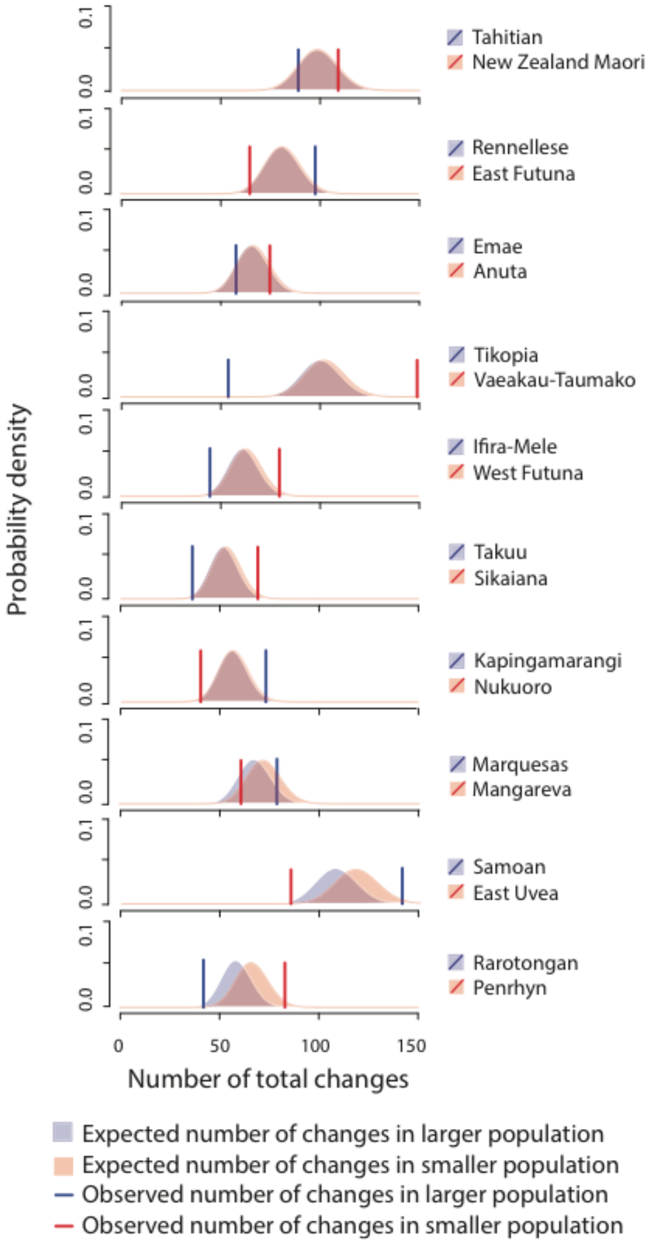


Figure S2: Log (ln) population size and number of loan words identified in the Austronesian Basic Vocabulary Database for the languages included in this study. There is no evidence of an association between population size and identified loan words, with or without the point on the extreme right of the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

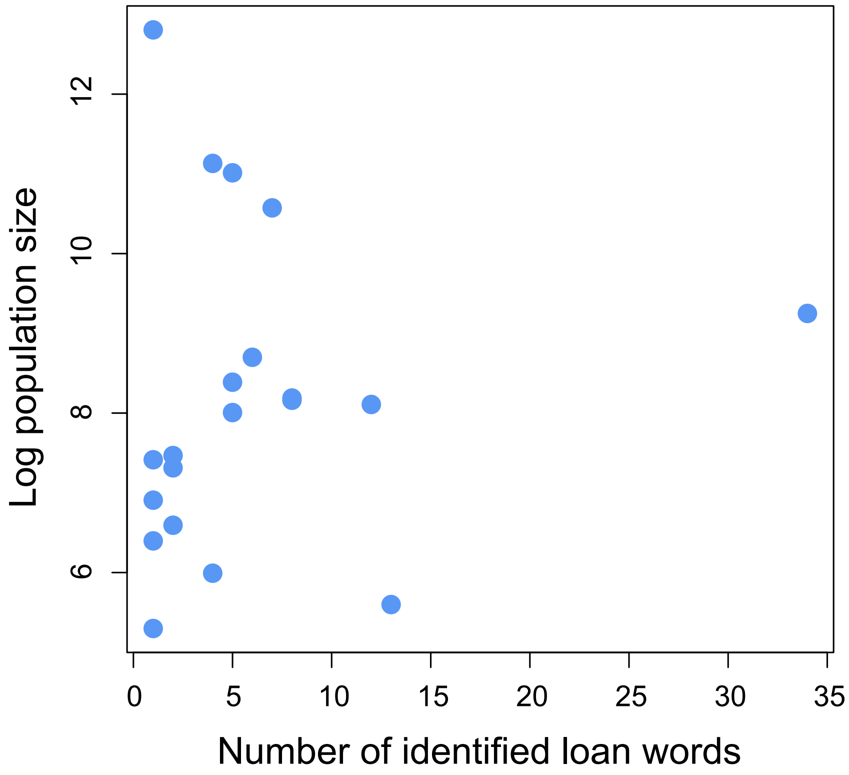
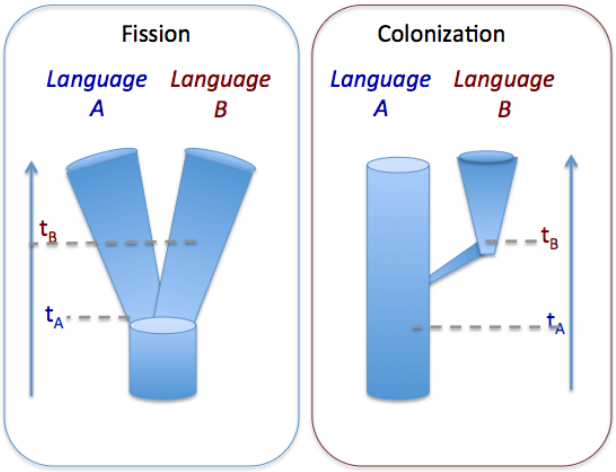


Figure S3: Illustration of the two modes of language origin modelled, and their relationship to the establishment dates of the two languages of the pair (t_A and t_B). For the fission model, the older date (t_A) provides the best estimate of date of divergence of the two languages in the pair. For the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


Fission
Language A
Language B
t_B
t_A
Colonization
Language A
Language B
t_B
t_A