

The distribution of phoneme inventory and language evolution

Simon Brown

Deviot Institute, Deviot, Tasmania, Australia;
College of Public Health, Medical and Veterinary Sciences, James Cook University,
Queensland, Australia;
e-mail: Simon.Brown@deviotinstitute.org

Abstract

The number of phonemes varies widely among languages, ranging from about 10 to more than 140. The overall distribution of phonemes is positively skewed and can be reasonably approximated by the lognormal distribution. This distribution is generally associated with multiplicative processes from which model-dependent estimates of the rates of growth of language families and the phoneme inventory of individual languages can be obtained. Using the inventories in UPSID, one simple model is used to show that vowel and consonant inventories differ in their rates of growth. While other appropriate models should be considered, this general approach provides a means of learning more about phoneme change.

Key words: distribution; growth; language evolution; phoneme inventory

Introduction

Among many the many differences between languages, one of particular note is the substantial variation in phoneme inventory. For example, the 451 languages represented in the UCLA Phonological Segment Inventory Database (UPSID) include 919 distinct phonemes and the number employed in a single language ranges from just 11 in Pirahã and Rotokas to 141 in !Xun (Maddieson, 1984). The presence of specific classes of phoneme in an inventory has been associated with geographical features and climate¹ (Everett 2013, Everett *et al.* 2015, Everett *et al.* 2016), and physiological rationalisations for these associations have been elaborated, although no supporting evidence has been provided. Phoneme inventory has also been used to date languages (Perreault & Mathew 2012) and to analyse the geographical dispersion and evolution of language (Atkinson 2011, Creanza *et al.* 2015). The correlation between phoneme inventory and population has been investigated on several occasions (Donohue & Nichols 2011, Hay & Bauer 2007, Pericliev 2004), but the effect size tends to be small even when the correlation is statistically significant. These associations are interesting for at least two reasons. First, because it has long been known that phonology changes rapidly, continuously (although not necessarily at a constant rate) and that the process is constrained and driven by social, cultural and demographic forces (Labov *et al.* 2013, Davletshin 2014). Second, statistically

¹ This is somewhat reminiscent of the old idea that is summarised by Farrar (1860, pp. 50-51) as “[t]he languages of the South are limpid, euphonic, and harmonious, as though they had received an impress from the transparency of their heaven, and the soft, sweet sounds of the winds that sigh among their woods. On the other hand, in the hirrients and gutturals, the burr and roughness of the Northern tongues, we catch an echo of the breaker bursting on their crags, and the crashing of the pine-branch over the cataract”. Compare the ideas of Munroe *et al.* (1996, 2009) and the associated literature.

significant associations can often be established in a complex system if only enough of the possibilities are tested (Roberts & Winters 2013).

Implicit in the UPSID statistics quoted above is the well known observation that the distribution of the number of phonemes employed in each language is positively skewed (Maddieson 1984). The implications of this are that (i) a large number of phonemes is observed in a small number of languages, (ii) a small number of phonemes is observed in some languages and (iii) that the probability of these states is actually quite high because of the skewed distribution. The well known inference is that a significant number of phonemes are used in only one language, which, in part, has motivated attempts to identify language based on uncommon phonemes (Hombert & Maddieson 1999).

However, the UPSID inventories are an approximately instantaneous cross-section through the evolutionary trajectories of the languages represented. If the database had been populated at any other time some of the inventories would have been different. This prompts speculation about the effect on the evolution of the overall distribution because it may imply more about the process of language evolution than appears to have been appreciated hitherto. Specifically, known processes give rise to specific standard distributions that are, or can be, positively skewed. In general, the gamma distribution arises from chain reactions and the operation of queues, the Weibull distribution is associated with fragmentation processes and the lognormal distribution is linked with multiplicative and diffusive processes. While such associations are not the only possibilities, they are well established. Here, the distribution of the UPSID phoneme inventories is considered and one possible interpretation is examined.

The distribution of phonemes

Several phoneme inventory databases are available through PHOIBLE Online (Moran *et al.* 2014). Of these, UPSID is particularly valuable because the sampling strategy employed in its construction was to include only one language from each “moderately distant genetic grouping” (Maddieson 1984: 158). While not a random sample, this strategy does eliminate the possibility of the inclusion of variants of a language and provides a reasonably representative sample of languages. This provides a useful resource for an examination of the distribution of phoneme inventories.

As Maddieson (1984) observed, the distribution of the number of phonemes among the languages in UPSID is positively skewed (Figure 1A) so that the mean (31) is greater than the median (29) which is greater than the mode (26). This, and the fact that the number of phonemes can not be negative, limits the range of distributions that might be used as approximations. As described in the Appendix, the gamma, Weibull and lognormal distributions were tested and, of these, the lognormal distribution provided the best fit to the distribution. The lognormal probability density function (PDF) of x is

$$f_{LN}(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right), \quad (1)$$

where μ and σ are the mean and standard deviation, respectively, of $\ln(x_i)$ and the cumulative distribution function (CDF) is

$$F_{LN}(x; \mu, \sigma) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{\ln(x) - \mu}{\sqrt{2}\sigma}\right), \quad (2)$$

where $\text{erf}(\cdot)$ is the error function. Equation (2) is a reasonable approximation of the CDF of the number of phonemes (Figure 1B). The distributions of both vowel (Figure 1, C and D) and consonant inventories (Figure 1, E and F) are also positively skewed and, in each case, the lognormal distribution is a reasonable approximation.

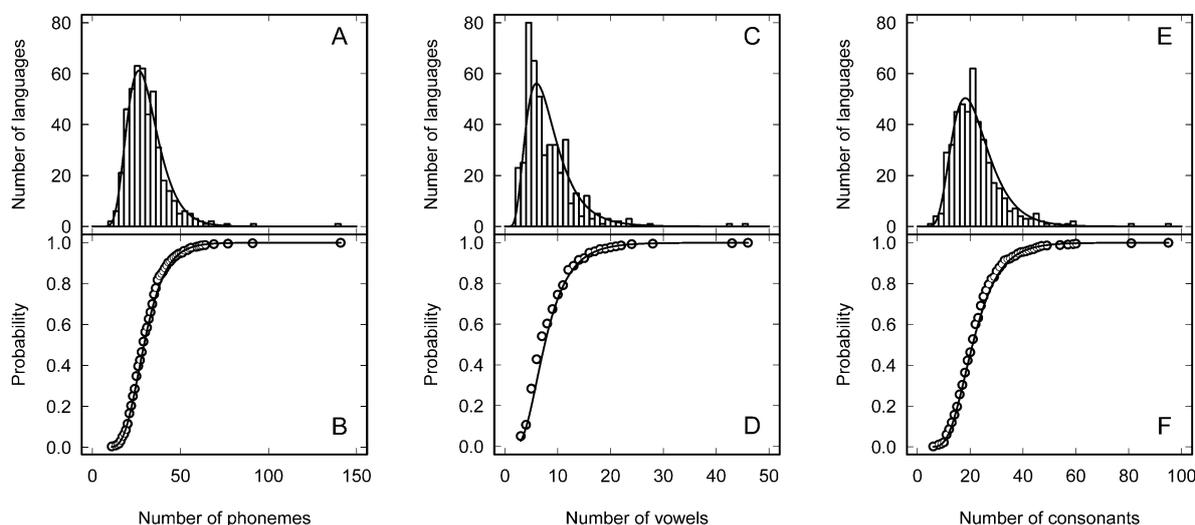


Figure 1. Distribution of the inventories of the 451 languages in UPSID of phonemes (A and B), vowels (C and D) and consonants (E and F). In each case distributions are shown as both the PDF (A, C, E) and the CDF (B, D, F). In (A), (C) and (E) each bar has the same relative width (2% of the range of the abscissa). In each panel the curve is the appropriate form of the lognormal distribution (1, 2) based on the maximum likelihood fit. Details of the fits are given in Table 1 and details of the model selection for the phoneme distribution are given in the Appendix.

Table 1. Properties of the distributions of consonants, vowels and phonemes among the 451 languages in UPSID and estimates of the mean and standard deviation obtained from the maximum likelihood fit of the lognormal distribution (1).

	Consonants		Vowels		Phonemes	
mean (SD)	22	(10)	9	(5)	31	(12)
median	21		7		29	
range	6 -	95	3 -	46	11 -	141
skewness	2.33		2.72		2.88	
kurtosis excess	11.22		14.42		19.82	
CV	0.423		0.563		0.373	
$\hat{\mu}$ (SD)	3.04	(0.02)	2.02	(0.02)	3.38	(0.02)
$\hat{\sigma}$ (SD)	0.38	(0.01)	0.48	(0.02)	0.32	(0.01)
p	>0.999		>0.999		>0.999	

Maddieson (1984: 9) also reported a weak, but significant, correlation between the number of consonants and vowels in the original sample² ($r = 0.38$). For the expanded database used here the correlation coefficient is somewhat smaller ($r = 0.22$ [95% CI: 0.13, 0.31], $p < 0.001$). However, it is significantly influenced by the two languages (!Xun and Parauk) each of which has more than 40 vowels (Figure 2A). If these are excluded the correlation is not significant ($r = 0.09$ [95% CI: -0.01, 0.18], $p = 0.064$). While each inventory is simply the sum of the numbers of its consonants and vowels, it is interesting that the correlation between the numbers of vowels and phonemes ($r = 0.598$ [95% CI: 0.535, 0.654], $p < 0.001$, Figure 2B) is weaker than that of the numbers of consonants and phonemes ($r = 0.92$ [95% CI: 0.90, 0.93], $p < 0.001$, Figure 2C). There is no significant change in this when the same two vowel-rich languages are excluded ($r = 0.49$ and $r = 0.91$).

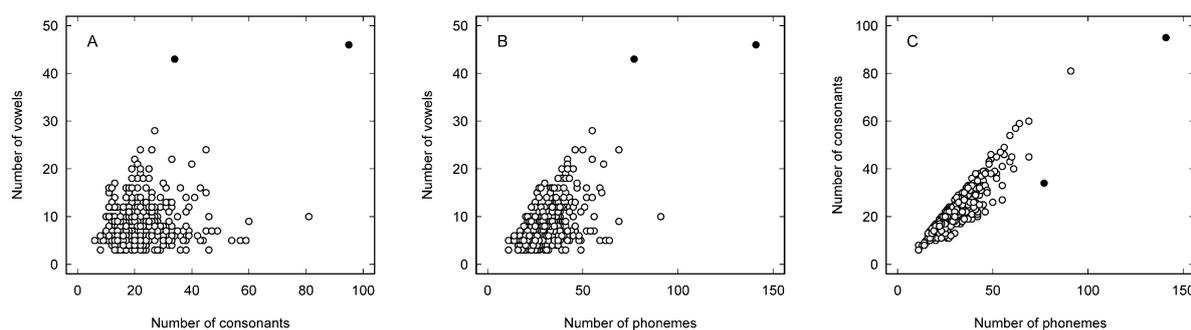


Figure 2. Correlation between the numbers of consonants and vowels (A), vowels and phonemes (B) and consonants and phonemes (C) for the 451 languages in UPSID. In each panel the black symbols (●) represent the two languages (!Xun and Parauk) with more than 40 vowels. The correlation coefficients are given in the text.

The apparent difference between vowels and consonants (Figure 2, B and C) might indicate that vowels and consonants contribute differently to the evolution of phoneme inventory. If this is the case, then it should be apparent from an analysis of the distributions which are approximately instantaneous cross-sections through the evolutionary trajectories of the languages represented. This question is among those addressed in the following analysis of the distributions.

One simple interpretation of the distribution

Given that the phoneme, vowel and consonant inventories in UPSID are approximately lognormally distributed (Figure 1), the natural question is what this might signify. It is well known that the product of many variables (x_i , $i = 1, 2, \dots, n$) is asymptotically lognormally distributed³. This can be seen simply by considering that if f is the product of n variables

$$f = \prod_{i=1}^n x_i, \quad (3)$$

then the log of f is the sum of the log of the x_i

² The original database of 317 languages was later expanded to the current 451 languages (Maddieson & Precoda 1989).

³ In some circumstances a lognormal distribution can be obtained transiently from additive processes (Mouri 2013) or from ‘disordered’ kinetics (Brown 2008).

$$\ln f = \ln \left(\prod_{i=1}^n x_i \right) = \sum_{i=1}^n \ln(x_i) \quad (4)$$

and, by the central limit theorem, $\ln(f)$ is normally distributed, in which case f has the lognormal distribution (1). While this usually taken to apply when n is large, Ioka and Nakamura (2002) showed that f can have the lognormal distribution for as few as $n = 3$ variables. However, if the x_i are correlated the sum can have the Weibull distribution (Bertin & Clusel, 2006). Choi *et al.* (2009) have shown that the characteristics of the growth in the variable of interest determine whether the distribution is Weibull or lognormal.

Remembering that the interpretation of such data depends on the model used (Brown 2016a, Brown 2016b), one simple description of growth can be based on an estimate of the number of families at time t (N_t) as a multiple (k_t) of that at time $t - 1$ (N_{t-1}), so

$$N_t = k_t N_{t-1} = k_t k_{t-1} N_{t-2} = \dots = k_t k_{t-1} \dots k_2 k_1 N_0, \quad (5)$$

which is clearly multiplicative and $\ln(N_t)$ would have the normal distribution, so N_t would have the lognormal distribution (4). This simple model is another way of writing $dN/dt = rN$, where r is a function of the rate constants (k_i). If the distribution of the phonemes is $f(x, t)$ and there are N families⁴, then

$$\frac{d}{dt} Nf(x, t) = \frac{dN}{dt} f(x, t) + N \frac{\partial f(x, t)}{\partial t} = rNf(x, t) + N \frac{\partial f(x, t)}{\partial t}. \quad (6)$$

Based on a Markovian approximation of the left hand side of (6), the evolution of the distribution derived from this generalised multiplicative model can be summarised by

$$\frac{\partial}{\partial t} f(x, t) = -(r + \lambda)f(x, t) + \lambda f(x - \delta, t), \quad (7)$$

where $f(x, t)$ is the distribution of the number of phonemes (x) as a function of time (t), r is the rate at which new families appear and λ and b are the mean rate and quantity by which the number of phonemes grows, respectively, and $\delta = xb/(1+b)$. If $f(x, t)$ is the lognormal distribution, then $r > \lambda b$ and

$$\mu = At \text{ and } \sigma = \sigma_0 \exp(Bt), \quad (8)$$

where

$$A \approx \lambda b + \left(\frac{\lambda}{2}\right)b^2 \text{ and } B \approx r - \lambda b \quad (9)$$

(Choi *et al.* 2009).

At least two predictions follow from (8) and both can be tested, to some extent, using the data in UPSID by comparing the distributions of inventories of language families. First, not only does μ tend to increase over time, so that the distribution of the number of phonemes moves to higher values, but the distribution also broadens⁵. Second, these changes are linked, so eliminating t from (8) yields

$$\ln \sigma = \ln \sigma_0 + \frac{B}{A} \mu, \quad (10)$$

which indicates that $\ln \sigma$ should be linearly related to μ and that the slope and intercept of the line are B/A and the initial value of $\ln \sigma$ ($\ln \sigma_0$), respectively.

⁴ This interpretation of N and x is based on UPSID in which there is only one representative of a language so the necessary statistics can only be estimated for a family rather than a language.

⁵ Which means that the slope of the CDF at the median ($\exp(-\mu)(2\pi\sigma^2)^{-1/2}$) declines with t (the rate constant for this is $A + B = r + 0.5\lambda b^2$ (9)).

Of the 85 language families represented in UPSID, 30 have more than four representatives and seven have more than 20 representatives. The distributions of the phoneme inventories of each of these families can be analysed using (8). The seven highly represented families can be used to examine the general prediction that as the median number of phonemes ($= \exp \hat{\mu}$) rises the distribution should broaden (8). These distributions are roughly consistent with the prediction, but the pattern is strongest for the vowel inventories. For clarity only three of these families are shown in Figure 3. In the case of the phoneme inventories, the gradient of the CDF at the median declines only slightly as the median number of vowels increases (Figure 3A), which corresponds to more obvious changes in the PDF which broadens and becomes more asymmetric as the median rises (Figure 3B). In contrast, the gradient of the CDF of the vowels increases significantly as the median increases (Figure 3C), so the PDF broadens considerably as the median rises (Figure 3D). The distributions of the phoneme inventories are combinations of the corresponding vowel and consonant distributions so the relationship is weaker for the consonant distributions (Figure 3, E and F).

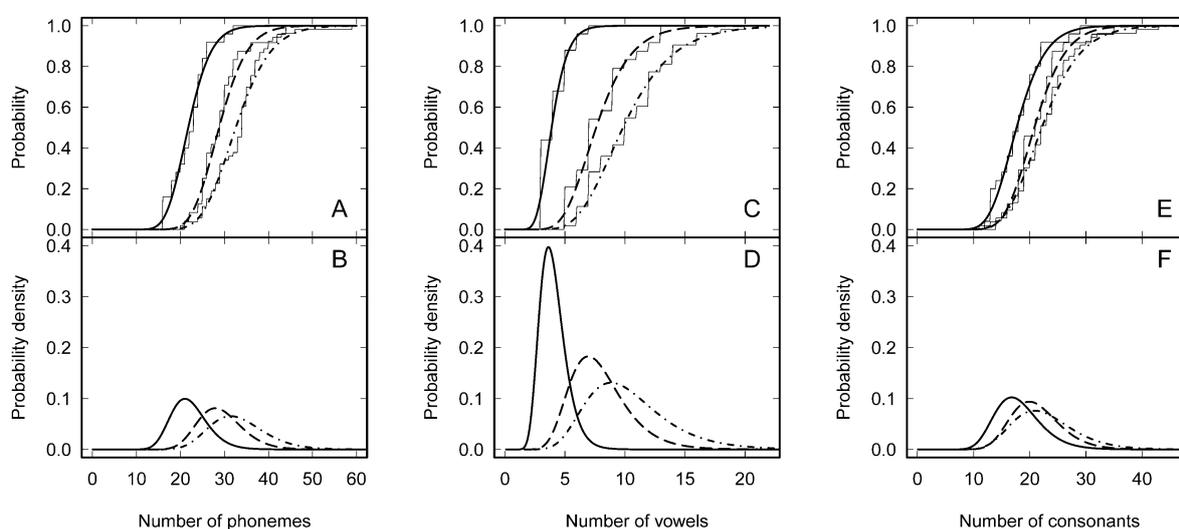


Figure 3. Distribution of the inventories of the languages in three language families of phonemes (A and B), vowels (C and D) and consonants (E and F). Distributions are shown as the CDF (A, C, E) and the PDF (B, D, F) for languages in the Australian (—), Nilo-Saharan (---) and Niger-Congo (- · - · -) families. In each panel the smooth curve is the appropriate form of the lognormal distribution (1, 2) based on the maximum likelihood fit and in the upper panels (A, C and E) the stepped lines are the empirical CDFs. Model selection was carried out as described in the Appendix.

There is some weak evidence that $\ln \sigma$ is linearly related to μ (10), which is the second testable prediction arising from (8). This is particularly evident for the vowels (Figure 4), for which $B/A \approx 0.6$ and $\ln \sigma_0 \approx -2.1$ (Table 2). While this is statistically significant ($p = 0.008$), the model accounts for only slightly more than 20% of the variance, so it is clearly not an adequate description of a very complex process, but it is an indication that this approach might have merit. Moreover, it is apparent from Figure 4 that two families are likely to influence the regression disproportionately. These are those families for which values of $\ln \sigma$ for the vowels were largest (Khoisan, $n = 4$) and smallest (Uto-Aztecan, $n = 6$). Eliminating just these two families from the analysis did reduce the significance ($p = 0.058$), but did not significantly alter the estimates of B/A or $\ln \sigma_0$ for the vowels. Given this, the analysis was repeated using just those families with more than 12 representatives, for which the parameter estimates (μ and σ) have smaller

uncertainty. Once again, the estimates of B/A and $\ln \sigma_0$ for the vowels were not significantly different from those given in Table 2 and $p = 0.012$. For this reason all 30 families are included in the analysis of vowel and consonant distributions given in Figure 4 and Table 2.

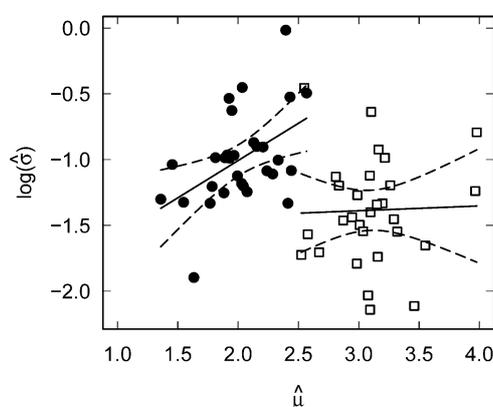


Figure 4. Relationship between estimates of $\log \sigma$ and μ for vowels (\bullet) and consonants (\square). The variables were estimated for each of the 30 language families with at least four representatives in UPSID. The solid lines are the least squares fits of (10) and the dashed curves define the corresponding 95% confidence bands. Details of the regressions are given in Table 2.

Table 2. Fit of (10) to the phoneme inventories of the those language families in UPSID represented by at least four members.

	Consonants	Vowels
B/A	0.0 (0.2)	0.6 (0.2)**
$\ln \sigma_0$	-1.5 (0.7)*	-2.1 (0.4)***
RSE	0.408	0.317
n	28	28
r^2	0.001	0.228
F	0.028	8.276
p	0.868	0.008
	* $p < 0.05$	** $p < 0.01$
		*** $p < 0.001$

The situation is different for consonants (Figure 4) for which B/A is much smaller, although $\ln \sigma_0 \approx -1.5$ is not significantly different from the estimate for vowels (Table 2). The estimate of $B/A \approx 0$ indicates that $B \approx 0$ and that $r \approx \lambda b$, which is the point of intersection between the lognormal ($r > \lambda b$) and Weibull distribution ($r < \lambda b$) regimes of the model (Choi *et al.* 2009). However, the distributions of both the consonant inventories of all the UPSID languages (Figure 1, E and F) and those of specific families (Figure 3E) are much more likely to be lognormal than Weibull (Table 1). This apparent discrepancy prompts the inferences that (a) another model might be more suitable for this particular purpose and, given the scatter of the consonant values in Figure 4, (b) more data, such as those in the other sources available in PHOIBLE Online (Moran *et al.* 2014), might be helpful.

This analysis indicates at least two things. First, that there is a difference in the contribution of consonants and vowels to the evolution of phoneme inventory. This is consistent with the difference between the correlations between vowel and phoneme inventories and consonant and phoneme inventories (Figure 2, B and C). Second, it provides a means of estimating the relative rates of growth of families and inventories. To see the latter, it follows from (9) that

$$B = \frac{1}{2}(2+b)a\lambda b \text{ and } r = \left(1 + a + \frac{1}{2}ab\right)\lambda b, \quad (11)$$

where $a = B/A$, from which

$$\frac{3}{2+3a} \leq \frac{A}{r} < \frac{1}{a}. \quad (12)$$

So A/r might be taken to indicate that vowel and consonant inventories grow at about 0.8-1.7 and at least 1.5, respectively, times the rate of growth of language families.

More can be inferred from these parameter estimates if it can be assumed that $b \ll 1$, which is likely to be the case if r and λ (and therefore A and B) have dimensions of y^{-1} . In this case, $A \approx \lambda b$, $B \approx a\lambda b$ and $r \approx (1+a)\lambda b$, which means that the rate constant for the increase in families (r) is larger than that for the increase in the number of phonemes (λ) if $a > 0$. This is the case for vowels ($A \approx \lambda b$, $B \approx 0.6\lambda b$, $r \approx 1.6\lambda b$), but is not the case for consonants ($A \approx \lambda b$, $B \approx 0$, $r \approx \lambda b$, Figure 4). Based on these estimates, the vowel and consonant inventories of families increase at the same rate, but the rate constant for the growth of σ is greater for vowels.

Some general observations

One concern arising from any comparison of phoneme inventories is that phonemes themselves are not necessarily uniquely determined (Chao 1934) and have been described as a ‘convenient fiction’ (Twaddell 1935, Drescher 2011). This is especially significant when phonemes are compared between languages, for example Trubetzkoy noted that

... since phonemic systems are structured differently in every language and even in every dialect, it is relatively rare to find a phoneme with exactly the same phonemic content in two different languages. One must not be misguided by the use of common international symbols of transcription. These symbols are only useful expedients. If the same letters should only be used for phonemes with fully equivalent phonemic content, a separate alphabet would have to be used for every language. (Trubetzkoy 1969: 74)

In developing UPSID, Maddieson (1984: 6) recognised difficulties of this sort. He observed that the decisions about phonemic status and phonetic description made in the compilation of UPSID did not necessarily correspond with those made by the compilers of the Stanford Phonology Archive which included 192 of the 317 languages in the original UPSID. While this may be an issue for the direct comparison of the phonemes themselves, it is less significant when just the numbers of phonemes are compared.

The interpretation of data is often model-dependent (Brown 2016a, Brown 2016b). There are two senses in which this applies to the distribution of phoneme inventories. First, slightly different expressions for μ and σ (8) can be obtained using other models (Hosoda *et al.* 2011, Goh *et al.* 2014). In these cases, as in (8), both μ and σ increase with time. These models will yield alternative interpretations of the data. However, it should be remembered that more complex analyses are not always helpful (Brown 2015). Second, model-dependence also applies to the distribution used to model the phoneme inventories. While the lognormal is the most suitable of the distributions considered here (Appendix), it is conceivable that some other distribution could be better. If this were the case it is also likely that different expressions for μ and σ (8) would be obtained.

The lognormal distribution (1-2) is intimately linked with growth processes (Kobayashi *et al.* 2011, Hosoda *et al.* 2011), but the model used here (7) is based on a very particular growth model: the assumption that $dN/dt = rN$. This implies that growth is effectively unlimited, indeed it indicates that the rate of increase in N only gets larger as N grows. This is distinct from other expressions used to model language evolution (Brown 2016a) and other processes (Brown 2007) in which growth tends towards an upper limit and the growth rate declines when N exceeds a specific threshold. Which of these approaches is most appropriate to phoneme inventory change warrants further investigation.

The model employed here represents one approach to the analysis of language evolution. As has been argued previously (Brown 2016a, Brown 2016b), if this is to be effective measures of variables that change slowly are required. Potential measures are being developed (Akulov 2015a, Akulov 2015b), but it seems unlikely that phoneme inventories change sufficiently slowly to meet this requirement (Davletshin 2014, Labov *et al.* 2013). Nevertheless, the evolution of phoneme inventory is of intrinsic interest and the methods developed to analyse it may eventually be of value in other contexts.

Conclusions

The analysis of UPSID phoneme inventory distribution described here is one of several possible similar approaches based on the lognormal approximation. It indicates that the evolution of vowels and consonants may differ (Figures 3 and 4), although the estimate of $B \approx 0$ for the consonant inventories (Table 2) may be an indication that another model might be more suitable. This general approach warrants further consideration and would benefit from the analysis of the other data available in PHOIBLE Online.

Appendix

At least five suitable distributions are defined for positive variables ($x > 0$) and can provide reasonable fits to the distribution: the lognormal (1), gamma

$$f_{\Gamma}(x; \beta, \gamma) = \frac{1}{\beta \Gamma(\gamma)} \left(\frac{x}{\beta} \right)^{\gamma-1} \exp\left(-\frac{x}{\beta} \right), \quad (\text{A1})$$

Weibull

$$f_w(x; \alpha, \gamma) = \frac{\gamma}{\alpha} \left(\frac{x}{\alpha} \right)^{\gamma-1} \exp\left(-\left(\frac{x}{\alpha} \right)^{\gamma} \right), \quad (\text{A2})$$

log-logistic and inverse beta distributions. The latter (also known as the beta distribution of the second kind) is indistinguishable from the lognormal distribution when σ is small, as it is here (Table A1), and the log-logistic distribution can also be difficult to distinguish from the lognormal distribution so neither of these was considered further. In contrast, the normal distribution is defined for positive and negative values ($x \in (-\infty, \infty)$) and is, therefore, logically inconsistent with the distribution of the number of phonemes, nevertheless the normal distribution was also fitted to the data. To identify the most suitable of these, the distributions were fitted to the inventories by maximum likelihood (Chakraborty 2015, Ihaka & Gentleman 1996) and the Akaike information criterion (AIC, (Akaike 1974, Sugiura 1978)) was calculated for each. The probability that the i th distribution is the most suitable (p_i) was estimated from the

AIC excess ($\Delta AIC_i = AIC_i - AIC_{\min}$, where AIC_{\min} is the lowest AIC which corresponds to the most likely choice) using

$$p_i = \frac{\exp(-\Delta AIC_i/2)}{\sum_{\text{all } j} \exp(-\Delta AIC_j/2)} \quad (\text{A3})$$

(Akaike 1978). Note that this method merely identifies the most appropriate of the models tested.

Taking the UPSID phoneme inventory data as an example, the AIC_{\min} is obtained for the lognormal distribution (Table A1), which indicates that this is the most suitable of those tested (Figure A1). Conversely, the largest AIC is that of the normal distribution and those of the gamma and Weibull distributions lie between these extremes. The probabilities estimated from (A3) indicate that the lognormal distribution is much more likely ($p > 0.999$) than the other distributions ($p < 0.001$).

Table A1. Properties of the fits of six distributions to the UPSID phoneme inventories.

Distribution	Parameters (SD)				AIC	ΔAIC	p
	location (μ, α, β)		shape (σ, γ)				
gamma	9.2	(0.6)	0.30	(0.02)	3347	31	<0.001
lognormal	3.38	(0.02)	0.32	(0.01)	3316	0	>0.999
normal	31.0	(0.5)	11.5	(0.4)	3490	174	<0.001
Weibull	2.58	(0.08)	34.6	(0.7)	3485	169	<0.001

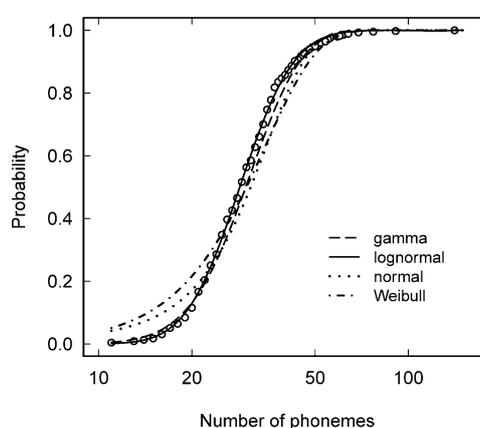


Figure A1. Fits of the gamma (---), lognormal (—), normal (····) and Weibull (- · - · -) distributions to that of the phoneme inventories of 451 languages in UPSID (o). Details are given in Table A1. Note that the abscissa is plotted on a log scale to make the differences clearer.

References

Akaike H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, vol. 19 no. 6; pp.: 716 – 723

Akaike H. 1978. On the likelihood of a time series model. *Journal of the Royal Statistical Society*, vol. 27D no. 3-4; pp.: 217 – 235

Akulov A. 2015a. Verbal grammar correlation index (VGCI) method: a detailed description. *Cultural Anthropology and Ethnosemiotics*, vol. 1 no. 4; pp.: 19 – 42

Akulov A. 2015b. Why conclusions about genetic affiliation of certain language should be based on comparison of grammar but not on comparison of lexis? *Cultural Anthropology and Ethnosemiotics*, vol. 1 no. 3; pp.: 5 – 9

Atkinson Q. D. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, vol. 332; pp.: 346 – 349

Bertin E., Clusel M. 2006. Generalized extreme value statistics and sum of correlated variables. *Journal of Physics A: Mathematical and General*, vol. 39; pp.: 7607 – 7619

Brown S. 2007. Two implications of common models of microbial growth. *ANZIAM Journal*, vol. 49; pp.: C230 – C242

Brown S. 2008. Estimating the distribution of forms of cytochrome oxidase from the kinetics of cyanide binding. *ICFAI Journal of Biotechnology*, vol. 2 no. 4; pp.: 51 – 60

Brown S. 2015. A bioinformatic perspective on linguistic relatedness. *Cultural Anthropology and Ethnosemiotics*, vol. 1 no. 4; pp.: 43 – 52

Brown S. 2016a. An examination of the calibration of linguistic distance. I. Sensitivity. *Cultural Anthropology and Ethnosemiotics*, vol. 2 no. 2; pp.: 1 – 9

Brown S. 2016b. An examination of the calibration of linguistic distance. II. Covariates. *Cultural Anthropology and Ethnosemiotics*, vol. 2 no. 4; pp.: 2 – 13

Chakraborty S. 2015. Generating discrete analogues of continuous probability distributions - a survey of methods and constructions. *Journal of Statistical Distributions and Applications*, vol. 2; pp.: 6

Chao, Y. –R. 1934. The non-uniqueness of phonemic solutions of phonetic systems. *Bulletin of the Institute of Phonetic Systems, Academia Sinica*, vol. 4 no. 4; pp.: 363 – 397 [reprinted in Joos M. 1958. *Readings in linguistics. The development of descriptive linguistics in America since 1925*, American Council of Learned Societies, New York, pp.: 38 – 54]

Choi M. Y., Choi H., Fortin J.-Y., Choi J. 2009. How skew distributions emerge in evolving systems. *Europhysics Letters*, vol. 85; pp.: 30006

Creanza N., Ruhlen M., Pemberton T. J., Rosenberg N. A., Feldman M. W., Ramachandran S. 2015. A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences of the USA*, vol. 112 no. 5; pp.: 1265 – 1272

Davletshin A. 2014. A seemingly on-going sound change in Takuu language of Papua New Guinea: historical and theoretical implications. *Journal of Language Relationship*, vol. 12 no. 1; pp.: 1 – 20

Donohue M., Nichols J. 2011. Does phoneme inventory size correlate with population size? *Linguistic Typology*, vol. 15; pp.: 161 – 170

Dresher B. E. 2011. The phoneme, in van Oostendorp M., Ewen C. J., Rice K. (eds) *The Blackwell companion to phonology: general issues and segmental phonology*. Blackwell Publishing Ltd, Oxford; pp.: 241 – 266

Everett C. 2013. Evidence for direct geographic influences on linguistic sounds: the case of ejectives. *PLoS ONE*, vol. 8 no. 6; pp.: e65275

Everett C., Blasi D. E., Roberts S. G. 2015. Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences of the USA*, vol. 112 no. 5; pp.: 1322 – 1327

Everett C., Blasi D. E., Roberts S. G. 2016. Language evolution and climate: the case of desiccation and tone. *Journal of Language Evolution*, vol. 1 no. 1; pp.: 33 – 46

Farrar F. W. 1860. *An essay on the origin of language*. John Murray, London

Goh S., Kwon H. W., Choi M. Y. 2014. Discriminating between Weibull distributions and log-normal distributions emerging in branching processes. *Journal of Physics A: Mathematical and Theoretical*, vol. 47 no. 225101

Hay J., Bauer L. 2007. Phoneme inventory size and population size. *Language*, vol. 83 no. 2; pp.: 388 – 400

Hombert J.-M., Maddieson I. (1999). The use of ‘rare’ segments for language identification. In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999*, vol. 1; pp.: 379 – 382. International Speech Communication Association

Hosoda K., Matsuura T., Suzuki H., Yomo T. 2011. Origin of lognormal-like distributions with a common width in a growth and division process. *Physical Review*, vol. 83E; pp.: 031118

Ihaka R., Gentleman R. 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, vol. 5; pp.: 299 – 314

Ioka K., Nakamura T. 2002. A possible origin of lognormal distributions in gamma-ray bursts. *Astrophysical Journal*, vol. 570; pp.: L21 – L24

Kobayashi N., Kuninaka H., Wakita J., Matsushita M. 2011. Statistical features of complex systems. Toward establishing sociological physics. *Journal of the Physical Society of Japan*, vol. 80; pp.: 072001

Labov W., Rosenfelder I., Fruehwald J. 2013. One hundred years of sound change in Philadelphia: linear incrementation, reversal, and reanalysis. *Language*, vol. 89 no. 1; pp.: 30 – 65

Maddieson I. 1984. *Patterns of sounds*. Cambridge University Press, Cambridge

Maddieson I., Precoda K. 1989. Updating UPSID. *Journal of the Acoustical Society of America* vol. 86 suppl 1; pp.: S19

Moran S., McCloy D., Wright R. 2014. PHOIBLE Online. Leipzig: Max Planck Institute for Evolutionary Anthropology (<http://phoible.org>, last accessed 27 January 2017)

Mouri H. 2013. Log-normal distribution from a process that is not multiplicative but is additive. *Physical Review*, vol. 88E; pp.: 042124

Munroe R. L., Fought J. G., Macauley R. K. S. 2009. Warm climates and sonority classes: not simply more vowels and fewer consonants. *Cross-Cultural Research*, vol. 43 no. 2; pp.: 123 – 133

Munroe R. L., Munroe R. H., Winters S. 1996. Cross-cultural correlates of the consonant-vowel (CV) syllable. *Cross-Cultural Research*, vol. 30 no. 1; pp.: 60 – 83

Pericliev V. 2004. There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology*, vol. 8; pp.: 376 – 383

Perreault C., Mathew S. 2012. Dating the origin of language using phonemic diversity. *PLoS ONE*, vol. 7 no. 4; pp.: e35289

Roberts S., Winters J. 2013. Linguistic diversity and traffic accidents: lessons from statistical studies of cultural traits. *PLoS ONE*, vol. 8 no. 8; pp.: e70902

Sugiura N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, vol. A7; pp.: 13 – 26

Trubetzkoy N. S. 1969. *Principles of phonology*. University of California Press, Berkeley

Twaddell W. F. 1935. On defining the phoneme. *Language*, vol. 11 no. 1; pp.: 5 – 62