CAES Vol. 2, № 1 (February 2016)

An analysis of the distributions of linguistic distances

Simon Brown

PhD in biochemistry, Deviot Institute, Deviot, Tasmania 7275, Australia
e-mail: Simon.Brown@deviotinstitute.org

Abstract

Standard measures of language relatedness such as the proportion of cognates or lexical distance that are commonly used are averaged over the pairs in word lists. Underlying these are distributions of data that have characteristics that convey information about the language pairs. A simple model of the distribution of lexical distance (D) based on a mixture of the beta distribution and discrete probabilities has been devised. Expressions based on this model are given for the expected value and variance of D that agree well with the values obtained from 1225 pairs of Indo-European languages.

Key words: distribution; lexical distance; mean; variance

Introduction

There is an interesting dichotomy in the literature concerning lexical distance. Frequently, word lists are used to calculate the average distance between pairs of languages (Gray & Atkinson 2003, Serva & Petroni 2008), but the other use to which the lists are put is to measure the proportion of cognates or the replacement or conservation of words (Dyen et al. 1967, Lees 1953, Lieberman et al. 2007). One might imagine that the loss of a cognate or the replacement of a word could be measured by an increase in distance, whereas conservation is associated with little or no change. However, the situation is not as clear as this might imply because two forms are cognate if they are descended from the same ancestor. For example, Dyen et al. (1992) considered the German *ich* and the Spanish *yo*, 'I', to be cognate although the lexical distance between them is 1.

In each case, however, the value reported is an average, whether it is the percentage of cognate words or the mean distance. The natural inference is that each measurement reflects an underlying distribution (Brown 2015). For example, the proportion of cognates (p_c) could be considered to be the result of a sequence of measurements in which each of n word pairs is tested and the result is 'cognate'/'not cognate' or 1/0. Then the sequence of results might be (0, 0, 1, 0, 1, …) and

$$p_c = \frac{1}{n}\left(n_0 \cdot 0 + n_1 \cdot 1\right), \qquad (1)$$

where n_0 and n_1 are the numbers of measurements in which the result is 0 or 1, respectively. Similarly, the lexical distance (D) is a representation of a sequence of results (d_1, d_2, ...), where d_k is the distance between the kth pair of words, and

$$D = \frac{1}{n}\left(d_1 + d_2 + \cdots + d_n\right). \qquad (2)$$

In these cases, p_c or D is reported, but these values represent just one property of the distribution of the results. The distributions have characteristics other than the mean (p_c or D), such as measures of the extent or symmetry of dispersion, that convey information about the data.

One simple way of looking at this is to consider that among the pairs of words used in the lexical comparison of two languages there are three broad possibilities (Figure 1). First, the pair might be identical (such as the French and English *long*) in which case the distance for this word (d) is 0 and the word might be considered to have been completely conserved between the two languages. This is a relatively rare occurrence in most language comparisons. Second, the pair might have nothing in common (such as the French *penser* and the English *think*), so that $d = 1$. In some comparisons, this is a frequent outcome. Third, between these extremes are the many examples for which d is between 0 and 1. Given that these possibilities apply to many word pairs, there is a distribution underlying every estimate of lexical distance or assessment of word replacement. The probability of observing d in this simple view is

$$P(d) = \begin{cases} p_0 & d = 0 \\ 1 - p_0 - p_1 & 0 < d < 1, \\ p_1 & d = 1 \end{cases} \quad (3)$$

where p_0 and p_1 are the probability that $d = 0$ and $d = 1$, respectively, and so the limiting case for closely related languages is

$$P(d) = \begin{cases} 0 & d = 0 \\ 1 & d > 0 \end{cases}, \quad (4)$$

the corresponding limiting case for unrelated languages is

$$P(d) = \begin{cases} 0 & d < 1 \\ 1 & d = 1 \end{cases} \quad (5)$$

(Figure 1). It will be apparent that (4) and (5) are special cases of (3) and that the intermediate case (3) shown in Figure 1 can have a variety of forms.

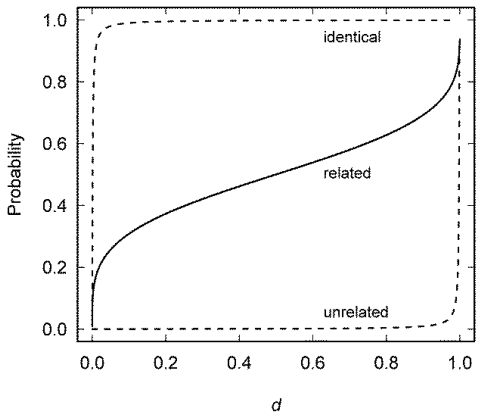

Figure 1. Approximation of the cumulative distribution functions corresponding to closely (4) and distantly (5) related languages (dashed curves), and of that of a more usual situation (solid curve) in which some words are replaced, others are conserved and many are altered (3).

Here I show that the distance between two languages can be expressed in terms of the distribution implied by (3). Using the Indo-European word lists maintained by Serva and Petroni[1] (2008) as an example, I show that (i) the expected value and the variance of D can be estimated reliably, and (ii) p_0 and p_1 can be used to generate approximate expressions for the expected value and the variance of D.

Background

A simple model based on (3) is outlined in the Appendix. This yields an expression for the expected value of the lexical distance (D) for a language pair

$$\mathrm{E}(D) = (1 - p_0)\frac{\alpha}{\alpha + \beta} + p_1 \frac{\beta}{\alpha + \beta} \tag{6}$$

that depends on the proportion of words in a list for which $d = 0$ (p_0) or $d = 1$ (p_1) and on the parameters (α and β) of the beta distribution fitted to the distribution of d for $0 < d < 1$. An expression for the variance of D

$$\mathrm{Var}(D) = \frac{\alpha^2(1 - p_0)p_0 + (2\alpha p_0 + \beta)\beta p_1 - \beta^2 p_1^2}{(\alpha + \beta)^2} + \frac{\alpha\beta(1 - p_0 - p_1)}{(\alpha + \beta + 1)(\alpha + \beta)^2} \tag{7}$$

is also obtained, although this can be written in other forms that may be helpful (Appendix).

Application to Indo-European languages

In order to demonstrate its usefulness, this analysis (3) was applied to the Indo-European word lists (n = 200) maintained by Serva and Petroni (2008). These are based on those originally generated by Dyen et al. (1992). All calculations were carried out in R (Ihaka & Gentleman 1996). The (unweighted) Levenshtein distance was calculated using the stringdist package (van der Loo 2014) and was then normalised to the length of the longer of each pair of words to yield the normalised Levenshtein distance (LDN) that is commonly used in this context (Serva & Petroni 2008). The value of D for each language pair was then determined using (A1). To characterise the distribution of d the number of words for which $d = 0$ or $d = 1$ were counted (n_0 and n_1, respectively) and p_0 and p_1 were determined from these. The cumulative distribution function (CDF) of the beta distribution was fitted to the remaining values of d ($0 < d < 1$) by nonlinear least squares[2] to obtain estimates of α and β.

Does the model fit the data?

All 1225 language pairs derived from the 50 Indo-European languages represented in the word lists have been analysed and in no case is the mean squared error (MSE) greater than 7.5×10^{-3}, which indicates that the model fits the data quite well. Moreover, there is no clear relationship between MSE and D, p_0, p_1, α or β, although in each case there is weak correlation ($R^2 < 0.11$). The distribution of d, the quality of the fit of (3) and some issues arising from this approach are illustrated in several examples. In each case both the frequency distribution of d and the CDF for $0 < d < 1$ are shown (the curves are fitted to $0 < d < 1$ only, even where n_0 and n_1 are shown).

[1] These are available at http://univaq.it/~serva/languages/languages.html.

[2] Similar results were obtained when α and β were determined by the method of moments from the mean and variance of d (A7).

The MSE ranged from 7.5×10^{-3} for Sardinian-Ukrainian (Figure 2, A and B) to 4.0×10^{-4} for Czech-Armenian (Figure 2, C and D), but in each case the fit is reasonable. Each of these language pairs is not closely related ($D = 0.873$ and 0.899) and, consistent with the concept underlying the analysis (Figure 1), the CDF is small for $d < 0.4$ and rises steeply as d increases above about 0.7 (Figure 2, B and D). There is some evidence of a small, but systematic discrepancy between the data and the beta CDF for $0.4 < d < 0.7$ and that may influence the quality of the fit at higher d (Figure 2B).

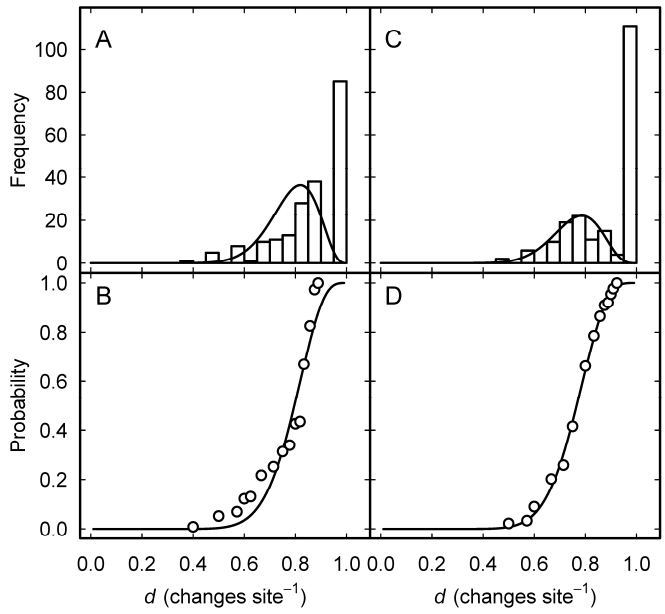

Figure 2. Distributions of d for Sardinian-Ukrainian (A and B) and Czech-Armenian (C and D). The histograms (A and C) show all of the data (that is n_0 and n_1 are shown) and the curves are the densities corresponding to the fits of the CDF (B and D) to the data excluding those for which $d = 1$. In (A) and (B) $\alpha = 15.10$, $\beta = 4.11$, $p_0 = 0$, $p_1 = 0.425$, $D = 0.873$ and MSE $= 7.5 \times 10^{-3}$, and in (C) and (D) $\alpha = 15.19$, $\beta = 4.87$, $p_0 = 0$, $p_1 = 0.555$, $D = 0.899$ and MSE $= 4.0 \times 10^{-4}$.

The distributions of d for more closely related language pairs can be almost symmetrically distributed ($\alpha \approx \beta$), as it is for Dutch-Swedish (Figure 3A), or may be positively skewed ($\alpha < \beta$), as it is for Icelandic-Faroese (Figure 3C), consistent with Figure 1. The model accounts adequately for these forms of the distribution, but there is some indication of a systematic discrepancy for Icelandic-Faroese up to about $d = 0.65$ (Figure 3D).

The differences between the data shown in Figures 2 and 3 are not just that the skewness is negative (Figure 2, A and C), approximately zero (Figure 3A) or positive (Figure 3C), but also in the values of p_0 and D. For both examples in Figure 2, $p_0 = 0$ and $D > 0.85$, whereas $p_0 > 0$ and $D < 0.65$ for those in Figure 3. Moreover, as p_0 increases in these examples, p_1 decreases from more than 0.4 (Figure 2) to less than 0.25 (Figure 3).

Reliability of the estimated mean and variance

The theoretical expressions for the expected value (6) and the variance (7) of D based on (3) perform well for the 1225 Indo-European language pairs (Figure 4). The agreement between the calculated and observed values is slightly better for the expected value of D (E(D) =

$(1.027 \pm 0.004)D - (0.031 \pm 0.003)$, $R^2 = 0.983$, $F_{1,\ 1223} = 72440$, $p < 0.001$), but it is clear that a small systematic deviation is not accounted for (Figure 4A), as is apparent from some of the cumulative distribution plots (Figure 2B). The estimated variance of D shows slightly more dispersion than E(D), but there is a very strong correlation with the observed variance (E(Var(D)) = (1.052 ± 0.003)Var(D) $- (0.0043 \pm 0.0001)$, $R^2 = 0.989$, $F_{1,\ 1223} = 112800$, $p < 0.001$) despite a small systematic deviation at higher variance (Figure 4B). This means that a simple vector (p_0, p_1, α, β) naturally summarises the distribution for each language pair.

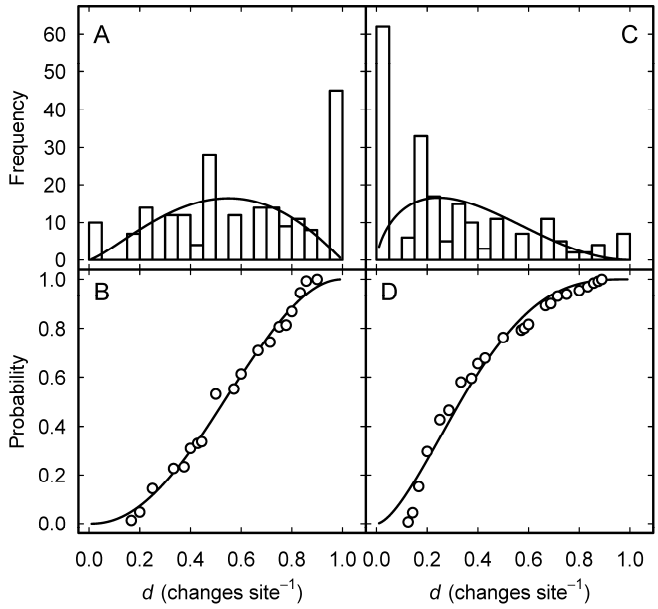

Figure 3. Distributions of d for Dutch-Swedish (A and B) and Icelandic-Faroese (C and D). The histograms (A and C) show all of the data (that is n_0 and n_1 are shown) and the curves are the densities corresponding to the fits of the CDF (B and D) to the data excluding those for which $d = 0$ and $d = 1$. In (A) and (B) $\alpha = 2.31$, $\beta = 2.07$, $p_0 = 0.05$, $p_1 = 0.225$, $D = 0.622$ and MSE = 1.0×10^{-3}, and in (C) and (D) $\alpha = 1.66$, $\beta = 3.03$, $p_0 = 0.31$, $p_1 = 0.035$, $D = 0.286$ and MSE = 2.6×10^{-3}.

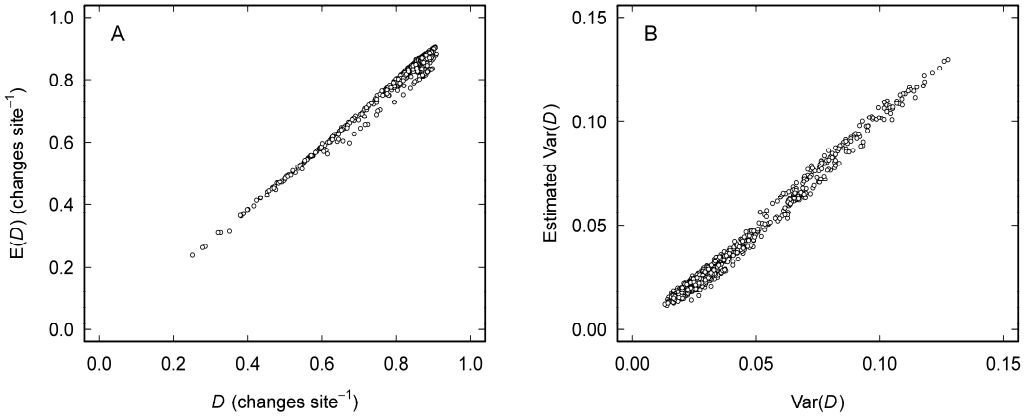

Figure 4. Performance of the model (3) in estimating D (A) and Var(D) (B). The expected value of D and the estimated variance of D were calculated using (6) and (7), respectively. In (A) least squares linear regression of E(D) on the actual values yielded E(D) = $(1.027 \pm 0.004)D - (0.031 \pm 0.003)$ ($R^2 = 0.983$, $F_{1,\ 1223} = 72440$, $p < 0.001$). In (B) least squares linear regression of these estimates on the actual values yielded E(Var(D)) = (1.052 ± 0.003)Var(D) $- (0.0043 \pm 0.0001)$ ($R^2 = 0.989$, $F_{1,\ 1223} = 112800$, $p < 0.001$).

CAES Vol. 2, № 1 (February 2016)

Parameter estimates and D

The values of p_0 and p_1 obtained from the 1225 language pairs are weakly negatively correlated with one another, but p_0 decreases and p_1 increases strongly with D (Figure 5). Moreover, it follows from (6) that E(D) must be between p_1 and $1 - p_0$. Given this, it is unsurprising that reasonable empirical estimates of D can be obtained from p_0 and p_1 alone

$$\hat{D} = (0.621 \pm 0.004) - (1.34 \pm 0.03)p_0 + (0.548 \pm 0.009)p_1 + (2.9 \pm 0.2)p_0 p_1 \qquad (8)$$

(Figure 6A) for which the adjusted R^2 = 0.922 ($F_{3,\ 1221}$ = 4823, $p < 0.001$). Similarly, reasonable estimates of the variance of D can be obtained in the same way

$$\mathrm{Var}(D) = (0.066 \pm 0.001) + (0.070 \pm 0.009)p_0 - (0.099 \pm 0.002)p_1 + (1.62 \pm 0.04)p_0 p_1 \quad (9)$$

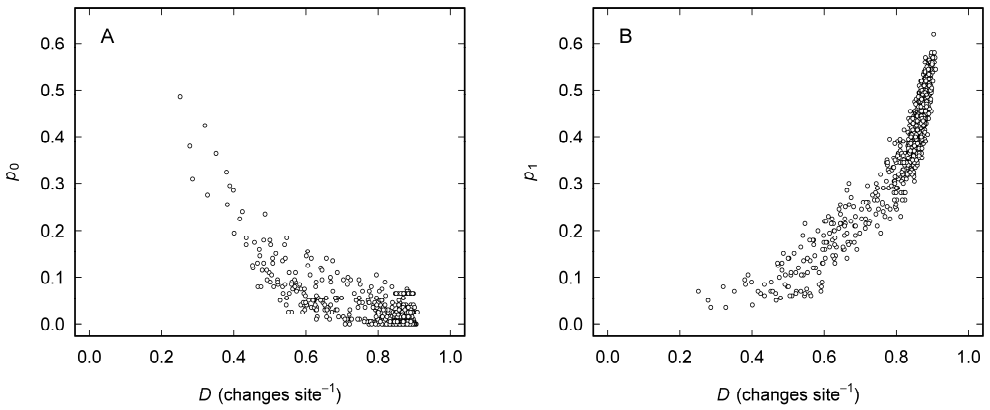

Figure 5. Relationship between D and p_0 (A) and D and p_1 (B) for 1225 pairs of Indo-European languages.

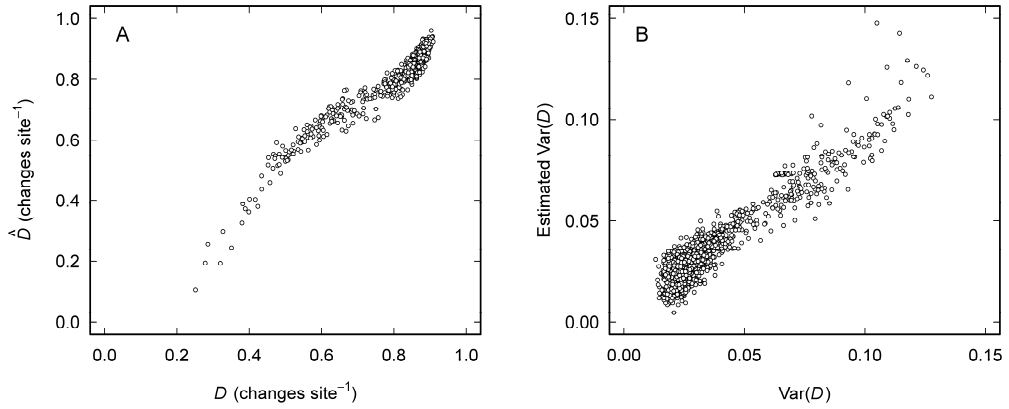

Figure 6. Estimation of D (A) and Var(D) (B) by least squares multiple regression on p_0 and p_1 (B). In (A) an estimate of D was obtained by least squares multiple regression which gave an expression (8) for which the adjusted R^2 = 0.922, $F_{3,\ 1221}$ = 4823, $p < 0.001$. In (B) an estimate of Var(D) was obtained by least squares multiple regression which gave an expression (9) for which the adjusted R^2 = 0.895, $F_{3,\ 1221}$ = 3476, $p < 0.001$.

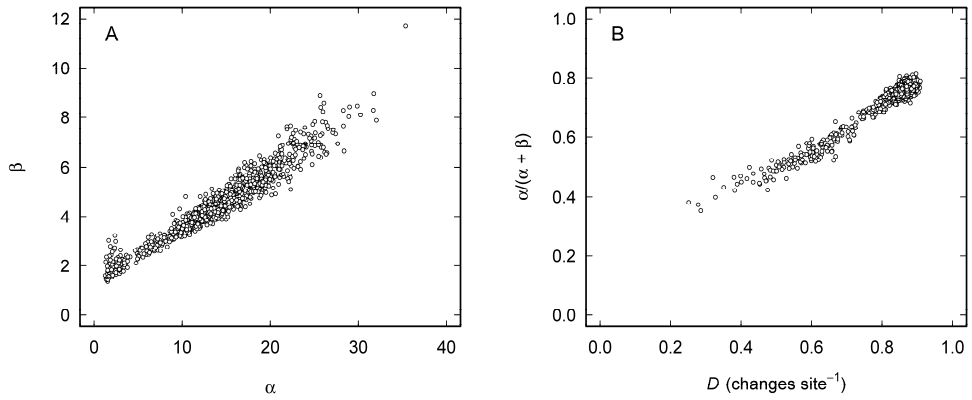

Figure 7. Correlation between α and β (A) and between $\alpha/(\alpha + \beta)$ and D (B). Note that $\alpha/(\alpha + \beta)$ is the expected value of d for $0 < d < 1$ (A7) rather than the expected value of D. The least squares linear regression of the values in (A) is $\beta = (0.226 \pm 0.002)\alpha + (1.32 \pm 0.02)$ ($R^2 = 0.940$, $F_{1, 1223} = 19270$, $p < 0.001$) and in (B) is $\alpha/(\alpha + \beta) = (0.731 \pm 0.005)D + (0.128 \pm 0.004)$ ($R^2 = 0.944$, $F_{1, 1223} = 20590$, $p < 0.001$).

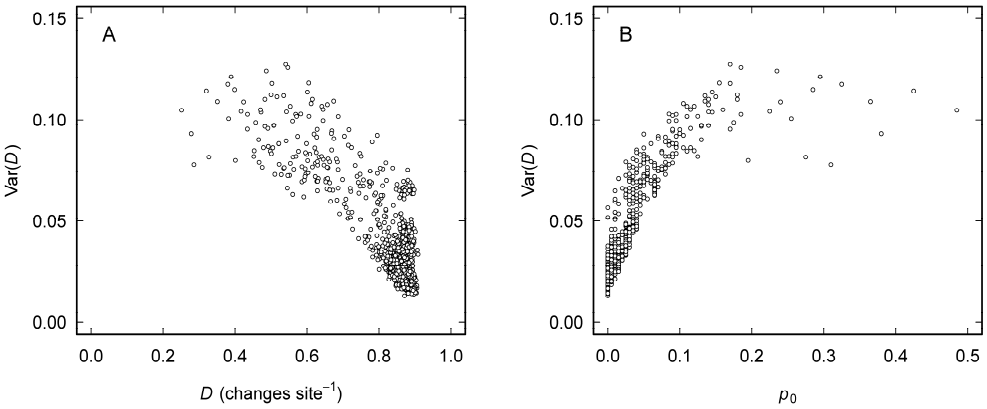

Figure 8. Relationship between the variance of D and D (A) or p_0 (B) for 1225 pairs of Indo-European languages.

(Figure 6B) for which the adjusted $R^2 = 0.895$ ($F_{3, 1221} = 3476$, $p < 0.001$). While these are not as good as those obtained using (p_0, p_1, α, β) and the theoretical expressions ((6) and (7), Figure 4), both (8) and (9) have an adjusted R^2 of about 0.9, which is conventionally interpreted to mean that p_0 and p_1 can account for about 90% of the variance in each case (Figure 6).

In contrast, α and β are strongly correlated with one another (Figure 7A), the relationships between D and α or β are consistent with (6), and $\alpha/(\alpha + \beta)$ is strongly correlated with D (Figure 7B). As $\alpha/(\alpha + \beta)$ is E(d) for $0 < d < 1$ (A7), it follows from (A5) that the correlation with D is to be expected.

Properties of the variance

This analysis also provides some insight into the variance of D (Figure 8), which is sometimes neglected (Brown 2015). In practice, Var(D) for this sample of languages is no more than about 0.12 (Figure 8), which is about half of the theoretical maximum (Appendix). Generally, the variance of D falls as D increases (Figure 8A), this is associated with a general

tendency for Var(D) to decline as p_1, α or β rises, although Var(D) does vary considerably for all values of D. However, Var(D) increases hyperbolically with increasing p_0 in such a way that it becomes relatively independent of $p_0 > 0.15$ (Figure 8B).

Conclusions

The model based on (3) that is outlined in the Appendix yields a reasonable description of the distribution of d for Indo-European language pairs. The distributions can be summarised using the vector of parameters (p_0, p_1, α, β). In particular, the model enables estimation (Figure 4) of the expected value of D (6) and the variance of D (7) directly from (p_0, p_1, α, β). It could be extended to the calculation of other measures if it were desirable. It is also possible to estimate D and Var(D) from a combination of p_0 and p_1 (Figure 6) or D from E(d) = $\alpha/(\alpha + \beta)$ (Figure 7B), although neither of these approaches is as reliable as the full model. Nevertheless, given that a subset of (p_0, p_1, α, β), either (p_0, p_1) or (α, β), can be almost as effective as the full model, it is reasonable to infer that the calculation of LDN may be unnecessary in some contexts. For example, it may be that p_0 and p_1 might provide an adequate approximation of D (Figure 6A) if the calculation of LDN happens to be impracticable for some reason.

It has become common practice to relate D to the time since two languages diverged from a common ancestor (Gray & Atkinson 2003, Petroni & Serva 2008, Serva & Petroni 2008). However, it is uncommon for the uncertainty of D (7) to be considered in this process, even when it is acknowledged that there is uncertainty in the dates used in such calibrations (Gray & Atkinson 2003). While the variance of D declines with D, it also varies considerably between language pairs with similar values of D (Figure 8A). These are important considerations in any attempt to relate D to divergence time and the issue warrants careful analysis.

The fundamental assumption on which the model is based is that the distribution of d ($0 < d < 1$) consists of a single component that can be represented using the beta distribution (Appendix). It is apparent from the examples (Figures 2 and 3) that this assumption is reasonable, but it is also clear that a second, small component is present in the distribution of at least some language pairs (Figure 2B). While no analysis of either its properties or its distribution has yet been undertaken, it does appear that this component is most prominent among the more distantly related language pairs. Nevertheless, this putative second component appears to have a fairly consistent form in the CDF, in that it is relatively small (contributing no more than about 20% of $n - n_0 - n_1$) and is located at $d \approx 0.4$-0.7 adjacent to the base of the major increase in the CDF. It is interesting to observe that the distribution of divergence times of Indo-European languages has a similar feature. For example, Serva and Petroni (2008) reported that the distribution of estimated divergence times for the same set of Indo-European languages was bimodal with a minor component ($p_{\max} \approx 0.01$) at about 1800 and a much larger one ($p_{\max} \approx 0.06$) at about 4800 y. This does not appear to be a characteristic common to all languages because the distribution of D for Austronesian languages is unimodal (Petroni & Serva 2008). Using their calibration coefficients, the divergence times for Indo-European languages (1800 y and 4800 y) correspond to $D \approx 0.59$ and 0.86, respectively[3]. This is approximately consistent with the putative minor component and the major component apparent in Figure 2B, for example, as are their relative

[3] Calculated from the expression for the divergence time $t = -\varepsilon\ln(1 - \gamma D)$, where $\varepsilon = 1750$ y and $\gamma = 1.09$, specified by Serva and Petroni (2008).

contributions to the distribution. This may be coincidental, but an alternative interpretation is that the bimodal distribution of distance may be intrinsic to at least some language pairs rather than being a property of the entire sample of languages. Work is underway to clarify the significance of this observation.

Appendix. Background to equations (6) and (7)

The measurement of lexical distance (D) using the Levenshtein distance normalised (LDN) is the average of the distances (d) between n pairs of words and represents just one means of summarising the distribution of d. Specifically,

$$D = \frac{1}{n}\sum_{k=1}^{n} d_k \tag{A1}$$

and $0 \le d_k \le 1$, where $d_k = 0$ means that the kth pair of words is identical, $d_k = 1$ means that the kth words are completely different, and values of $0 < d_k < 1$ indicate various degrees of difference between the kth pair of words. Writing the number of words for which d_k is 0 or 1 as n_0 and n_1, respectively, and, without loss of generality, assuming that the d_k are sorted by magnitude, (A1) can be written as

$$D = \frac{1}{n}\left(n_0 \cdot 0 + \sum_{k=n_0+1}^{n-n_1} d_k + n_1 \cdot 1 \right) = \frac{1}{n}\left(\sum_{k=n_0+1}^{n-n_1} d_k + n_1 \right). \tag{A2}$$

Rearranging (A2) and dividing by $n - n_0 - n_1$

$$\frac{1}{n-n_0-n_1}\sum_{k=n_0+1}^{n-n_1} d_k = \frac{nD - n_1}{n-n_0-n_1}, \tag{A3}$$

gives an expression for the the average of those of the $(n - n_0 - n_1)$ values of d that are neither 0 nor 1. Letting $p_0 = n_0/n$ and $p_1 = n_1/n$, and writing $\mathrm{E}(d)$ for (A3) yields an expression for D that depends only on p_0, p_1 and $\mathrm{E}(d)$

$$D = (1 - p_0 - p_1)\mathrm{E}(d) + p_1 = (1 - p_0)\mathrm{E}(d) + p_1(1 - \mathrm{E}(d)), \tag{A4}$$

where

$$\mathrm{E}(d) = \frac{D - p_1}{1 - p_0 - p_1} \tag{A5}$$

is simply the expected value of that subset of d (A3).

It remains to specify the distribution for $0 < d < 1$. In this case it is natural to use the beta distribution which is defined on this domain[4] and provides great flexibility, in that it can be positively or negatively skewed. The probability density function of the beta distribution is

[4] It is often assumed that $0 \le x \le 1$, but, to be consistent with (3), the definition of Feller (1971: 50) is used. This also minimises the constraints on α and β because (A6) is undefined at $x = 0$ if $0 < \alpha < 1$ and at $x = 1$ if $0 < \beta < 1$.

$$f(x;\alpha,\beta)=\frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \tag{A6}$$

where $\mathrm{B}(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ is the beta function, $\alpha, \beta > 0$ and it is assumed that $0 < x < 1$ (3). The mean, variance and skewness of this distribution are

$$\mathrm{E}(x)=\frac{\alpha}{\alpha+\beta},\quad \mathrm{Var}(x)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\quad\text{and}\quad \mathrm{Sk}(x)=\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}, \tag{A7}$$

respectively. If d has the beta distribution, then the special case in which $\alpha = \beta = 1$ is the uniform distribution, $\beta > \alpha$ for those cases that are positively skewed and $\alpha > \beta$ for those that are negatively skewed (A7). The limiting cases of (3) are $\beta >> \alpha$ for closely related languages (4), so $\mathrm{E}(d) \approx 0$, and $\alpha >> \beta$ for completely unrelated languages (5), for which $\mathrm{E}(d) \approx 1$ (Figure 1).

Substituting (A7) into (A4) yields an expression for the expected value of D

$$\mathrm{E}(D)=(1-p_0)\frac{\alpha}{\alpha+\beta}+p_1\frac{\beta}{\alpha+\beta} \tag{A8}$$

from which $0 \le \mathrm{E}(D) \le 1$ because $\mathrm{E}(D) = 0$ if $p_0 = 1$, $\mathrm{E}(D) = 1$ if $p_1 = 1$ and for all other values of p_0, p_1 $0 < \mathrm{E}(D) < 1$. The variance of D is

$$\mathrm{Var}(D)=\frac{\alpha^2(1-p_0)p_0+(2\alpha p_0+\beta)\beta p_1-\beta^2p_1^2}{(\alpha+\beta)^2}+\frac{\alpha\beta(1-p_0-p_1)}{(\alpha+\beta+1)(\alpha+\beta)^2}, \tag{A9}$$

which is 0 if $p_0 = 1$ or $p_1 = 1$, but is positive otherwise, and reduces to that of the beta distribution (A7) if $p_0 = p_1 = 0$. For the two special cases where $p_0 = 0$ or $p_1 = 0$, (A9) reduces to

$$\mathrm{Var}(D;p_0=0)=(1-p_1)\left(\frac{\beta^2p_1}{(\alpha+\beta)^2}+\mathrm{Var}(d)\right) \tag{A10}$$

and

$$\mathrm{Var}(D;p_1=0)=(1-p_0)\left(\frac{\alpha^2p_0}{(\alpha+\beta)^2}+\mathrm{Var}(d)\right), \tag{A11}$$

respectively, where $\mathrm{Var}(d)$ is given by (A7). This leads to a simpler version of (A9)

$$\mathrm{Var}(D)=\frac{\alpha^2(1-p_0)p_0}{(\alpha+\beta)^2}+\frac{2\alpha\beta}{(\alpha+\beta)^2}p_0p_1+\frac{\beta^2(1-p_1)p_1}{(\alpha+\beta)^2}+(1-p_0-p_1)\mathrm{Var}(d), \tag{A12}$$

from which it follows that $0 \le \mathrm{Var}(D) \le 0.25$. Equations (A8) and (A9) (or (6) and (7), respectively) provide a means of estimating D and its variance from the vector $(p_0, p_1, \alpha, \beta)$ that characterises the distribution of d defined by (3).

References

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

Dyen I., James A. T., Cole J. W. T. 1967. Language divergence and estimated word retention rate. *Language*, vol. 43 no. 1; pp.: 150 – 171

Dyen I., Kruskal J. B., Black B. 1992. An Indoeuropean classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society*, vol. 82 no. 5; pp.: 1 – 132

Feller W. 1971. *An introduction to probability theory and its applications II*. John Wiley & Sons, Inc., New York

Gray R. D., Atkinson Q. D. 2003. Language-tree divergence times support Anatolian theory of Indo-European origin. *Nature*, vol. 426; pp.: 435 – 439

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

Lees R. B. 1953. The basis of glottochronology. *Language*, vol. 29 no. 2; pp.: 113 – 127

Lieberman E., Michel J. –B., Jackson J., Tang T., Nowak M. A. 2007. Quantifying the evolutionary dynamics of language. *Nature*, vol. 449; pp.: 713 – 716

Petroni F., Serva M. 2008. Language distance and tree reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008; pp.: P08012

Serva M., Petroni F. 2008. Indo-European languages tree by Levenshtein distance. *Europhysics Letters,* vol. 81: pp.: 68005

van der Loo M. P. J. 2014. The stringdist package for approximate string matching. *R Journal*, vol. 6 no. 1: pp.: 111 – 122