An examination of the calibration of linguistic distance.

II. Covariates

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 assessment of linguistic relatedness is often based on the analysis of word lists to estimate the distance between languages. To relate the distance to a divergence time requires established calibration data and a model representing the underlying processes of language change. However, the available calibration data may be based on different sorts of evidence, come from different geographical regions and represent different language families, for example. Covariates such as these can impact significantly on the estimated divergence time. This reinforces concerns about the selection of calibration data, especially when just a small number of values are chosen.

Key words: dating evidence; geographical region; linguistic distance; model dependence

Introduction

Despite the difficulties associated with the use of lexical distance in the measurement of relatedness (Akulov 2015b, Brown 2016b, Hoijer 1956, Matisoff 1990, Rea 1958, Teeter 1963), two common measures are based on the analysis of word lists. The first is the normalised Levenshtein distance (LDN) from which the second, the LDN divided (LDND), is derived (Bakker et al. 2009). For two languages (α and β) LDN is the average Levenshtein distance \( d_{L_i} \) (Levenshtein 1966) between \( n \) pairs of words normalised to the length \( L_{ii} \) of the longer of the two words

\[
D_o(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_{L_i}(\alpha_i, \beta_i)}{L_{ii}},
\]

where \( \alpha_i \) and \( \beta_i \) represent word \( i \) in languages \( \alpha \) and \( \beta \), respectively. The LDND is

\[
D_s(\alpha, \beta) = \frac{D(\alpha, \beta)}{\Gamma(\alpha, \beta)},
\]

where LDN is normalised to the ‘global distance’

\[
\Gamma(\alpha, \beta) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{d_{L_i}(\alpha_i, \beta_j)}{L_{ij}}
\]

(Petroni & Serva 2010) between the \( n(n-1) \) pairs of words of different meanings in the word lists of \( \alpha \) and \( \beta \) (Brown 2016c).

Part 1 was published in CAES vol. 2 no. 2
To relate either of these measures of distance to the divergence time \((t)\) between languages requires established calibration values. While this process is model-dependent (Brown 2016b), one simple expression for this relationship is

\[
d_x = \phi \left( 1 - \exp \left(-k, t \right) \right),
\]

where \(D_x\) may be \(D_o\) (1) or \(D_s\) (2) depending on whether the distance measure is LDN or LDND, respectively, and that the values of \(\phi\) and \(k_x\) also depend on whether \(D_x\) is \(D_o\) (in which case \(x = o\)) or \(D_s\) (\(x = s\)). However, (4) is also based on values of the rate constant \((k_x)\) for the divergence process and the upper limit \((\phi_s)\) of \(D_x\) that are estimated from appropriate calibration data (Brown 2016b). Given (4), it is unsurprising that the calibration of \(D_o\) is relatively insensitive to differences between large divergence times. For example, this is the case for \(t\) in excess of about 4500 years for Indo-European languages, although this estimate was based on values of \(k_o\) and \(\phi_o\) derived from only a few calibration values (Brown 2016b).

It has been shown that estimates of \(D_o\) for Indo-European languages are intrinsically uncertain (Brown 2016a). Moreover, it is mathematically inevitable that LDND is greater than LDN, as has been confirmed for Indo-European languages (Brown 2016c), and, as LDND includes the global distance (3), it is likely that it has at least the same uncertainty as LDN with which it is highly correlated (Brown 2016c). This means that the uncertainty in the calibration of \(D_s\) has several sources. Of these, the distribution of \(D_o(a, \beta, \phi)\) has been considered (Brown 2016a) and some thought has been given to the estimation of \(k\) and \(\phi\) using LDN (Brown 2016b). However, the calibration data on which this is based have not been considered. Here, some of the factors intrinsic to the data (the covariates) that contribute to the calibration and its uncertainty and their impact on the estimated divergence time are considered.

**Calibration data**

Some recent work has been based on just a small number of calibration values (Brown 2016b). Specifically, the LDN values given by Gray and Atkinson (2003) and Petroni and Serva (2008) provided a convenient basis for analysing some of the issues associated with the sensitivity of the relationship between \(D_o\) and \(t\). However, having shown that LDND is highly correlated with LDN and that \(D_s \geq D_o\) when the same Indo-European word lists are used (Brown 2016c), it is reasonable to make use of the much larger set of LDND-based calibration values given by Holman et al. (2011). These data are expressed in terms of similarity \((S)\), expressed as a percentage, which is related to \(D_s\) by

\[
S = 100(1 - D_s).
\]

While the implicit assumption that \(D_s \leq 1\) is theoretically incorrect (Brown 2016c), it is a reasonable approximation and, using (5), the original \(D_s\) values can be calculated from \(S\) (Table 1). Holman et al. (2011) analysed their calibration data by considering \(\log(S)\) as a function of \(t\)

\[
\log(S) = 2\log(r) + \log(S_0),
\]

where \(S_0\) is the initial value of \(S\) and \(r\) is the average proportion of lexical similarity retained after 1000 y (Swadesh 1950). Using (5), (6) is equivalent to

\[
D_s = 1 - 0.01S_0 e^{-2t} = (1 - 0.01S_0) + 0.01S_0 (1 - \exp[2\ln(r) \cdot t]),
\]

which is related to (4), but differs from it in both the initial \((D_s(t = 0) = 0, D_o(t = 0) > 0)\) and final values \((D_o(t \to \infty) = \phi_o, D_s(t \to \infty) = 1)\) because (4) and (6) are based different models (Brown 2016b).
Table 1. The 52 calibration values given by Holman et al. (2011) grouped according to the dating evidence. Values of $D_s$ were calculated from the reported similarity ($S$) using (5).

<table>
<thead>
<tr>
<th>Archaeological</th>
<th>Time (y)</th>
<th>$D_s$ (site$^{-1}$)</th>
<th>Time (y)</th>
<th>$D_s$ (site$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benue-Congo</td>
<td>6500</td>
<td>0.964</td>
<td>Malayo-Chamic</td>
<td>2400</td>
</tr>
<tr>
<td>Central S. African Khoisan</td>
<td>2000</td>
<td>0.883</td>
<td>Malayo-Polynesian</td>
<td>4250</td>
</tr>
<tr>
<td>Dardic</td>
<td>3550</td>
<td>0.730</td>
<td>Mississippi Valley Siouan</td>
<td>2475</td>
</tr>
<tr>
<td>E. Polynesian</td>
<td>1050</td>
<td>0.517</td>
<td>Ongamo-Maa</td>
<td>1150</td>
</tr>
<tr>
<td>E. Malayo-Polynesian</td>
<td>3350</td>
<td>0.924</td>
<td>Pama-Nyungan</td>
<td>4500</td>
</tr>
<tr>
<td>Ga-Dangme*</td>
<td>600</td>
<td>0.509</td>
<td>Saami</td>
<td>1750</td>
</tr>
<tr>
<td>Indo-Aryan</td>
<td>3900</td>
<td>0.752</td>
<td>Southern Nilotic</td>
<td>2500</td>
</tr>
<tr>
<td>Indo-European</td>
<td>5500</td>
<td>0.947</td>
<td>Swahili*</td>
<td>1200</td>
</tr>
<tr>
<td>Indo-Iranian</td>
<td>4400</td>
<td>0.917</td>
<td>Temotu</td>
<td>4300</td>
</tr>
<tr>
<td>Inuit</td>
<td>800</td>
<td>0.396</td>
<td>Tupi-Guarani (coastal)*</td>
<td>1750</td>
</tr>
<tr>
<td>Iranian</td>
<td>3900</td>
<td>0.859</td>
<td>Turkic*</td>
<td>2500</td>
</tr>
<tr>
<td>Ma’anyan-Malagasy</td>
<td>1350</td>
<td>0.697</td>
<td>Wakashan</td>
<td>2500</td>
</tr>
</tbody>
</table>

| Epigraphic     |         |                     |         |                     |
| Cholan         | 1600    | 0.567               | Hmong-Mien | 2500 | 0.943         |
| Czech-Slovak   | 1050    | 0.328               | Oromo     | 460  | 0.365         |
| Ethiopian Semitic | 2450 | 0.811              | Scandinavian | 1100 | 0.672       |
| Goidelic       | 1050    | 0.725               | Sorbian (Lusatian) | 450  | 0.312         |

| Historical     |         |                     |         |                     |
| Breton-Welsh   | 1450    | 0.574               | Maa      | 600  | 0.325         |
| Cham           | 529     | 0.449               | Maltese-Maghreb Arabic | 910  | 0.664       |
| Chamic         | 1550    | 0.841               | Mongolic | 750  | 0.792         |
| Chinese        | 2000    | 0.870               | Northern Roglai-Tsat | 1000 | 0.742       |
| Common Turkic  | 1419    | 0.621               | Romance  | 1729 | 0.710         |
| E. Slavic      | 760     | 0.605               | Romani   | 650  | 0.381         |
| English-Frisian| 1550    | 0.694               | Slavic   | 1450 | 0.570         |
| Germanic       | 2100    | 0.708               | Southern Songhai | 550  | 0.371       |
| Italo-Western Romance | 1524 | 0.678              | Southwest Tungusic | 236  | 0.467       |
| Ket-Yugh       | 1300    | 0.513               | Western Turkic | 900  | 0.546       |

* Dates for which both archaeological and historical evidence were available.

In addition to the larger number of values, these data have at least three other properties that were recognised by Holman et al. (2011). First, they represent estimates based on three different sorts of dating evidence: archaeological, epigraphic and historical (although four of the 52 values are based on both historical and archaeological data and, therefore, they are included in each of these categories in the analysis). In contrast, the LDN values given by Gray and Atkinson (2003) were predominantly historical and epigraphic, whereas those used by Petroni and Serva (2008) were not well defined. Second, they include representatives of 17 different language families (Holman et al. 2011), rather than being restricted to Indo-European languages. However, of these, only the Indo-European and Austronesian families
have more than five\(^1\) values among the data. Third, they include values relating to four geographical regions (Holman et al. 2011), although there are only five values from the Americas.

For at least two reasons these calibration data are not directly comparable to those used previously (Brown 2016b). First, the words in the lists are represented phonetically using the 41 symbol ASJP code rather than the 26 letter Latin alphabet (Holman et al. 2011). Second, even if this were not the case, Holman et al. (2011) report similarity based on LDND rather than on LDN. While it might be expected that these two measures are correlated when the same word lists are used (Brown 2016c), there is no reason to expect that this would necessarily extend to a situation in which both the lists and the representation of the words in those lists differ.

Nevertheless, even if the data themselves are not strictly comparable, it does not preclude the possibility that the estimates of \(k_x\) and \(\phi_s\) obtained from the two sets of calibration data might be compared. Here it is assumed that \(\phi_s\) tends to be greater than \(\phi_o\), because \(D_s \geq D_o\) when the same Indo-European word lists are used (Brown 2016c), and that comparison is of limited value.

**Influence of data selection**

Holman et al. (2011) reported that the type of dating evidence, the language family and the geographical region were not significant factors in the discrepancy (expressed as a percentage) between the calibration dates and the Automated Similarity Judgment Program dates. This does not necessarily mean that these factors are insignificant in the relationship between \(t\) and \(D\) (or, equivalently, \(S(5)\)).

In fact, the type of dating evidence, the language family and the geographical region do have some impact on the estimates of the parameters of (4). Moreover, the relationship between these and the estimation of divergence time using (4) means that even statistically insignificant differences among these estimates can have a considerable effect on the estimated \(t\) (Brown 2016b). One of the inferences to be drawn from this is that more and improved calibration data are desirable.

Considering the data by dating evidence, and remembering that \(\phi_s\) is not bounded from above (Brown 2016c), the estimated upper limit of LDND (\(\phi_s\)) ranges from about 1.0 for the epigraphic data (Figure 1A(i)) to 0.77 for the historical data (Figure 1A(ii)), whereas the archaeological data and the entire set of calibration data yield values of about 0.9 (Figure 1A, (iii) and (iv)). As might be anticipated, the range of dates over which the data are available is a significant consideration. For example, the lack for data for longer divergence times precludes reliable estimates using epigraphic or historical data alone, although these data are useful for smaller \(t\) (Figure 1A, (i) and (ii)). The archaeological data are sufficient to yield a reasonable estimate of \(\phi_s\), which is not significantly different from that obtained using all 52 values (Table 2). The rate constants range from \(0.8 \times 10^{-3} \text{ y}^{-1}\) to \(1.5 \times 10^{-3} \text{ y}^{-1}\) and the estimates obtained from epigraphic or historical data alone have considerably greater

\(^1\) Five is the minimum number of data points required to obtain even a very poor estimate of the parameters of (4) by nonlinear regression. This is apparent from the fit and the parameter estimates obtained using the calibration data from the Americas (Figure 2A(iv) and Table 2).
uncertainty than that \((0.9 \times 10^{-3} \text{ y}^{-1})\) obtained using archaeological data (Table 2). The considerable differences in the uncertainty of these parameter estimates and their interdependence are illustrated by the 95% confidence regions in \((\phi_s, k_s)\)-space (Figure 1B). The confidence regions estimated for the epigraphic and historical data are considerably larger than that obtained using the archaeological data, consistent with their much greater uncertainty, but the latter lies predominantly within those of the others (Figure 1B).

![Figure 1](image-url)

Figure 1. The calibration data based on the nature of the dating evidence. In (A) the epigraphic (i), historical (ii) or archaeological (iii) evidence or the complete set of values given in Table 1 (iv) are shown. In panels (ii) and (iii) the grey squares represent those four values for which there is both archaeological and historical evidence (E. W. Holman et al. 2011). In each panel, the solid curves represent least squares fits of (4) to the data and the dashed curves represent extrapolations of these. In (B) the 95% confidence regions of the parameter estimates (Beale 1960) for the epigraphic (E, \(\cdots\cdots\)), historical (H, \(-\cdots\)) and archaeological (A, \(\cdots\)) calibration data are shown. The grey, white and black circles are the coordinates of the parameter estimates for epigraphic, historical and archaeological evidence, respectively, used in the curves in panel (A) and these are given in Table 2.

Only three of the four geographical regions are represented by more than five values, which means that the parameter estimates for the fourth (the Americas) have limited reliability. However, for these subsets of the data \(\phi_s\) ranges from 0.84 to about 1 and \(k_s\) ranges from about \(0.6 \times 10^{-3}\) y to \(1.2 \times 10^{-3}\) y (Figure 2A). Despite the uncertainty associated with these estimates (Table 2), the range of dates for which values are available are reasonable for the African, Eurasian and Pacific data (Figures 2A, (i), (ii) and (iii)), consistent with the confidence regions shown in Figure 2B, but this is not the case for the data from the Americas (Figure 2A(iv)) for which no reliable estimates of the parameter uncertainty could be obtained (Table 2).

Of the 17 language families represented in Table 1 (Holman et al. 2011), only two contribute more than five values. The data for Indo-European languages yield \(\phi_s = 0.85\) and \(k_s = 1 \times\)
$10^{-3}$ y and for the Austronesian languages $\phi_s = 0.91$ and $k_s = 1.2 \times 10^{-3}$ y (Table 2). It is notable that the estimate of $k_s$ for Indo-European languages is larger than the values of $k_o$ reported previously (Brown 2016b, Serva & Petroni 2008).

Table 2. Properties of the curves (4) for the calibration data classified by dating evidence (Figure 1), geographical region (Figure 2) or selected language families. Parameter estimates ($\pm$ SE) and the mean square of the residuals (MSR) were obtained using R (Ihaka & Gentleman 1996).

<table>
<thead>
<tr>
<th></th>
<th>$\phi_s$</th>
<th>$k_s \times 10^{-3} \text{y}^{-1}$</th>
<th>MSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data</td>
<td>0.89 ± 0.04</td>
<td>1.0 ± 0.1</td>
<td>0.012</td>
</tr>
<tr>
<td>Evidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archaeological</td>
<td>0.92 ± 0.03</td>
<td>0.9 ± 0.1</td>
<td>0.006</td>
</tr>
<tr>
<td>Epigraphic</td>
<td>1.0 ± 0.3</td>
<td>0.8 ± 0.5</td>
<td>0.021</td>
</tr>
<tr>
<td>Historical</td>
<td>0.77 ± 0.08</td>
<td>1.5 ± 0.5</td>
<td>0.015</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africa</td>
<td>0.97 ± 0.07</td>
<td>0.9 ± 0.1</td>
<td>0.005</td>
</tr>
<tr>
<td>Americas</td>
<td>1</td>
<td>0.6</td>
<td>0.007</td>
</tr>
<tr>
<td>Eurasia</td>
<td>0.84 ± 0.06</td>
<td>1.2 ± 0.3</td>
<td>0.017</td>
</tr>
<tr>
<td>Pacific</td>
<td>0.91 ± 0.04</td>
<td>1.2 ± 0.2</td>
<td>0.007</td>
</tr>
<tr>
<td>Family*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Austronesian</td>
<td>0.91 ± 0.04</td>
<td>1.2 ± 0.2</td>
<td>0.006</td>
</tr>
<tr>
<td>Indo-European</td>
<td>0.85 ± 0.06</td>
<td>1.0 ± 0.2</td>
<td>0.011</td>
</tr>
</tbody>
</table>

*Only those language families with more than five values are included.

Figure 2. The calibration data based on the geographical region. In (A) the Eurasian (i), African (ii), Pacific (iii) or American (iv) calibration data given in Table 1 are shown. In each panel, the solid curves represent least squares fits of (4) to the data and the dashed curves represent extrapolations of these. In (B) the 95% confidence regions of the parameter estimates (Beale 1960) for the Pacific (· · · ·), African (····) and Eurasian (– – –) calibration data are shown. The grey, white and black circles are the coordinates of the parameter estimates for the Pacific, African and Eurasian regions, respectively, used in the curves in panel (A) and these are given in Table 2.
Model-dependence

The model used to implement the calibration is a significant factor in determining the relationship between $t$ and $D_s$ (Brown 2016b). Holman et al. (2011) used ordinary least squares regression to fit (6) to the calibration data. This yielded a correlation between $t$ and $\log(S)$ of $-0.84$ and the coefficients of (6) were $r = 0.72$ and $S_0 = 92\%$ for the complete set of data. Acknowledging that both $t$ and $S$ are inherently uncertain, they also fitted the inverse function for which they reported the equivalent of $r = 0.79$ and $S_0 = 62\%$. If estimates of the “analytical” variances of both $t$ and $\log(S)$ were available, a regression line, based on (6), obtained by taking into account the uncertainties of both variables would lie between these two extremes (Francq & Govaerts 2014).

The fits of (6) to the calibration values obtained by Holman et al. (2011) have a mean square of the residuals (MSR) of 0.013 for $(r, S_0) = (0.72, 92\%)$ and 0.012 for $(0.79, 62\%)$, and that shown in Figure 1A(iv), obtained by nonlinear regression using (4), has an MSR of 0.012. There is, therefore, no significant difference in the quality of the fits, but they are clearly different (Figure 3). Specifically, (4) passes though the origin ($t = 0, D_x = 0$), whereas (7) passes through ($t = 0, D_x = 1 - 0.01S_0$), which implies that LDND can never be smaller than 0.08 or 0.38. The latter lower limit, corresponding to $(r, S_0) = (0.79, 62\%)$, is clearly inconsistent with the five calibration values that are smaller, rendering this fit unusable. A second difference between (4) and (7) is that the latter asymptotically approaches 1, whereas the former approaches $\phi_s$, as $t$ increases. This difference reflects the consideration of the possibility of convergent evolution in the model on which (4) is based, but not in that underlying (7) (Brown 2016b). Which of these fits best represents the data presumably depends largely on which model is preferred.

Figure 3. Comparison of the curves fitted to all the calibration data (○) given in Table 1. The solid line (——) is the fit of (4) shown in Figure 1A(iv), the dashed (– – – ; $(r, S_0) = (0.72, 92\%)$) and dotted lines (·····; $(r, S_0) = (0.79, 62\%)$) are the fits of (6) reported by Holman et al. (2011) expressed in the form of (7).
While (4) and (7) are different because the latter has an extra constant term, parameter values of (7) can be used to estimate those corresponding to a parallel curve described by (4), and vice versa (Appendix). In the case of the curves shown in Figure 3, \((r, S_0) = (0.72, 92\%)\) and \((0.79, 62\%)\) correspond to \((k_s, \phi_s) = (0.66 \times 10^{-3}\text{ y}^{-1}, 0.92)\) and \((0.47 \times 10^{-3}\text{ y}^{-1}, 0.62)\), respectively. For the first of these, the estimated \(\phi_s\) is not significantly different from that in Table 2, but the estimated rate constant \((k_s)\) is about 34\% smaller than the value of \(1 \times 10^{-3}\text{ y}^{-1}\) given in Table 2, similar to the values of \(k_o\) for Indo-European languages reported previously (Brown 2016b, Serva & Petroni 2008), and is smaller than any of the other rate constants in Table 2.

**The impact on the estimated divergence time**

The parameter estimates summarised in Table 2 do not differ greatly from one another. However, as has been shown previously (Brown 2016b), small differences in \(\phi_x\) or \(k_x\) can be associated with significant differences in the estimated divergence time. In order to examine this effect for the subsets of the calibration data considered the \(t\) values for as many of the entries in Table 1 as possible were calculated using each of the parameter sets in Table 2. This required the inverse of (4)

\[
t = -\frac{1}{k_x} \ln \left(1 - \frac{D_x}{\phi_x}\right),
\]

which is undefined unless \(D_x < \phi_x\). For consistency, this constraint was interpreted as requiring that \(D_x\) was smaller than any of the \(\phi_x\) in the subsets being considered. This meant that the number of values varied because, for example, it required that \(D_x < 0.77\) for the comparison of the dating evidence (so \(n = 35\) in Figure 4A), but \(D_x < 0.84\) for the comparison by geographic region (so \(n = 37\) in Figure 4B).

![Figure 4](image_url)

Figure 4. The impact of the selection of calibration data by dating evidence (A) or geographic region (B) on the estimated divergence time. Values of \(D_x\) in Table 1 were used to calculate \(t\) (8) using the parameter estimates obtained from the indicated subsets of the data (Table 2). In each panel only those values of \(D_x < \min(\phi_x)\) were used and the actual \(t\) for these values are also shown.
Consistent with the regression approach taken (Figures 1A(iv) and 2A(iv)), the range of \( t \) values estimated using the parameters derived from the entire set of calibration data (the ‘full model’) was smaller than that of the actual values used in the dating evidence (Figure 4A) and the regional (Figure 4B) comparisons. Those obtained using the archaeological or epigraphic parameters were very similar to that obtained from the full model, although the epigraphic estimates were slightly decreased at larger \( D_s \). In contrast, the historical estimates were almost 1000 y larger at high \( D_s \) and were smaller at low \( D_s \) (Figure 4A). The effect of the parameters for different geographical regions is more complex (Figure 4B). At small \( D_s \), the estimates obtained using the African, Eurasian and Pacific data were similar to that of the full model, but the Eurasian model increased the estimates at large \( D_s \) compared with the African and Pacific models and the full model. The estimates derived from the parameters based on data representing the Americas gave consistently larger estimates than the other models (Figure 4B), but this should be treated with caution because the parameter estimates were based on only a very small number of values (Figure 2A(iv)).

**Some general observations**

Two issues arise from the analyses described here. First, that the selection of calibration data is a critical step. Second, that the choice of the model on which to base the calibration is also important. The divergence time estimate depends on both of these irrespective of whether LDN, LDND or some other measure is used.

The data used to calibrate the estimation of \( t \) certainly matter. The values of \( t \) calculated using parameters based on calibrations founded on historical dating evidence can be quite different from those based on archaeological or epigraphic evidence (Figure 4A). It may be that this reflects the relatively restricted range of dates (236-2100 y) of the historical data (Figure 1A(ii)), but the epigraphic data in Table 1 are similarly restricted (450-2500 y, Figure 1A(i)) and they yield estimates comparable with those based on the archaeological evidence (Figure 4A). Similarly, calibrations based on data from different geographical regions (even discounting the few data representing the Americas) can also yield different estimates of \( t \) (Figure 4B). Whether this reflects real regional variation or underlying differences between language families is not clear. One of the weaknesses of LDN or LDND as measures of relatedness is that \( k_x \) is too large for them to be useful at longer \( t \) (Brown 2016b), especially if it is not constant in time (Atkinson et al. 2008, Nettle 1999) or space (Atkinson 2011). Other measures, that change more slowly, are necessary to overcome this difficulty (Brown 2016b), which requires the further development of different approaches (Akulov 2015a, Grimes & Agard 1959, Kroeber & Chrétien 1937). However, it is not inevitable that more sophisticated measures will yield a significant improvement unless great care is taken (Brown 2015). Nevertheless, the choice of calibration data is a critical step irrespective of whether lexical distance or some better measure is employed.

The model-dependence of the calibration of linguistic distance illustrated in Figure 3 has been considered previously (Brown 2016b). The choice of model influences the size of the estimated divergence time whether the data are based on LDN, LDND or some other measure. This is because the model encodes a particular view of the processes underlying language change and so the choice deserves careful consideration.
Conclusions

The calibration data of Holman et al. (2011) provides a means of assessing the intrinsic variation due to the dating evidence, geographic region and, to a lesser extent, language family. The selection of the data used in calibration affects the relationship between $D_s$ and $t$ (Figures 1 and 2, Table 2) and can strongly influence the estimates of divergence time derived from it (Figure 4). This reinforces the concern previously expressed (Brown 2016b) about the use of a small number of apparently arbitrarily chosen calibration values. This issue merits further investigation. Similarly, the choice of the model that underpins the analysis is an important consideration (Brown 2016b) because it also strongly influences the apparent relationship between $D_s$ and $t$ (Figure 3).

While these observations were drawn from data based on LDND, there is no justification for thinking that data based on LDN would have been very different.

Appendix

The relationship between the parameters of (4) and (7) can be clarified by writing the latter as

$$D_s = 1 - 0.01S_0 + 0.01S_0\left[1 - \exp\left(-2 \times 10^{-3} \ln\frac{1}{r} t\right)\right], \quad (A1)$$

in which it is recognised that $D_s(t = 0) = 1 - 0.01S_0$ and that $r$ is dimensionless, but is based on $t$ expressed in thousands of years. While (A1) is not identical to (4), the last term on the right of (A1) does have the same form as (4) so it is possible to match the shapes of the curves by matching these terms. Comparing the last term on the right of (A1) with (4) gives $k_s = -2 \times 10^{-3} \ln(r)$ and $\phi_s = 0.01S_0$. The parameter value estimates of Holman et al. (2011) are $(r, S_0) = (0.72, 92\%)$ and $(0.79, 62\%)$, which correspond to $(k_s, \phi_s) = (0.66 \times 10^{-3} \, y^{-1}, 0.92)$ and $(0.47 \times 10^{-3} \, y^{-1}, 0.62)$, respectively. The curves described by $(r, S_0)-(k_s, \phi_s)$ pairs are parallel and offset by the constant in (A1).

References


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


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

Brown S. 2016a. An analysis of the distributions of linguistic distances. *Cultural Anthropology and Ethnosemiotics*, vol. 2 no. 1; pp.: 2 – 12


Brown S. 2016c. Two measures of linguistic distance. *Cultural Anthropology and Ethnosemiotics*, vol. 2 no. 3; pp.: 2 – 12


