CAES Vol. 2, № 2 (June 2016)

An examination of the calibration of linguistic distance. I. Sensitivity

Simon Brown

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

Abstract

The interpretation of the calibration of linguistic distance depends on both the model used and on the implementation of the calibration. While these are important considerations and can greatly affect the estimated divergence times, the calibration data themselves are clear evidence that lexical distance is at best of limited value as a measure of language change for divergence times in excess of about 4500 y. Specifically, the data indicate that it is likely that at least 50% of lexical divergence takes place over only 1000 y and that by about 4500 y a 1% change in lexical distance corresponds to a 10% change in estimated divergence time. A measure of linguistic distance that changes more slowly than lexical distance is required if longer divergence times are to be estimated reliably.

Key words: calibration; divergence time; lexical distance; sensitivity

Introduction

Among the simplest measures of linguistic relatedness are those based on the analysis of word lists. These are simple for at least two reasons. First, the data are encoded in an inherently systematic manner. While the composition of the lists themselves is an endless cause for discussion, the words in the lists are treated as directly comparable. Second, the necessary computational tools are readily available and easy to use, and new methods can be developed without too much difficulty given the powerful string manipulation tools that are incorporated into Python, Perl and other programming languages. In contrast, typological, morphological, semantic and other measures are not necessarily as naturally encoded in way that facilitates analysis, although progress is being made (Akulov 2015a). It may be for these reasons that lexical distance is widely used, despite being problematic (Akulov 2015b, Hoijer 1956, Matisoff 1990, Rea 1958, Teeter 1963).

Irrespective of how linguistic distance (D) is measured, it is common to relate it to the divergence time (t). This calibration is usually based on a relatively small set of known coordinates ((t, D), Table 1) and a model. The latter might be as simple as the assumption that the rate of change in a measure of language similarity (N) is proportional to N

$$N' = -bN \qquad N(0) = N_0 \tag{1}$$

where b is the rate constant for the process[1] and N_0 is the initial value of N. This is the basis of glottochronology in which N is the number cognates in two word lists, and has the solution

[1] There is confusion in the literature concerning the difference between the rate of this process (N'), which changes continuously, and the rate constant (b), which does not change. Clearly, N' never equals b (although

$$N = N_0 \exp(-bt) \tag{2}$$

(Dobson 1978, Lees 1953). In this case[2]

$$D = 1 - \frac{N}{N_0} = 1 - \exp(-bt) \tag{3}$$

from which the divergence time is

$$t = -\varepsilon \ln(1 - D) \tag{4}$$

and $b = \varepsilon^{-1}$ can be estimated for each calibration coordinate (t, D) (Lees 1953).

Table 1. Calibration values for Indo-European languages. Values of D were calculated in R (Ihaka & Gentleman 1996) from the word lists (n = 200) of Serva and Petroni[3] (2008). 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. The divergence times were calculated from the calibration dates given by Gray and Atkinson (2003) and those used by Serva and Petroni (2008) are also listed. The upper values in brackets are the estimated divergence times obtained by Gray and Atkinson (2003).

	D	Divergence time*
	(changes site^{-1})	(y)
Icelandic-Norwegian (—)	0.492	1100
Iberian-French (450-800 CE)	0.591	1200 — 1600
Slavic (pre-700 CE)	0.658	>1300
Italian-French (—)	0.560	1600
Welsh-Breton (400-550 CE)	0.681	1600 — 1750
Italic-Romanian (150-300 CE)	0.652	1700 — 1900
Germanic (50-250 CE)	0.645	1800 — 2000
Balto-Slavic (1400 BCE-100 CE)	0.823	1900 — 3400
Indic (pre-200 BCE)	0.820	>2200
Irish-Welsh (pre-300 BCE)	0.830	>2300 (2900)
Iranian (pre-500 BCE)	0.909	>2500
Indo-Iranian (pre-1000 BCE)	0.821	>3000 (4600)
Greek (pre-1500 BCE)	0.901	>3500 (7300)

*Each value is rounded to the nearest century.

One other assumption underlying (1) is that there is no linguistic convergence. This possibility was considered by Serva and Petroni who assumed that the average lexical distance (D), measured using the normalised Levenshtein distance (so $0 \leq D \leq 1$), could increase as well as decrease (Petroni & Serva 2008, Serva & Petroni 2008): for example, the distance between words increases as a result of random changes and declines as words become more similar due to language borrowings or just by accident. This was expressed as a diffential equation in D

$$D' = -a(1 - D) - bD \qquad D(0) = 0, \tag{5}$$

and the solution given was written as

$$t = -\varepsilon \ln(1 - \gamma D) \tag{6}$$

$|N'(0)|/N_0 = b$), because $N \geq 0$ (1), and the dimensions of N' and b differ: if N is a measurement of conserved words, N' might be expressed in units of conserved words y^{-1} and b in units of y^{-1}.

[2] Extended models have been proposed in which there is a rate constant for each word in a list (van der Merwe 1966). While this is inevitably correct, it has not proved especially helpful to date, although see Brown (2016).

[3] Available at http://univaq.it/~serva/languages/languages.html.

(Petroni & Serva 2008, Serva & Petroni 2008), which is a natural generalisation of (4). Substituting two calibration coordinates (Italian-French and Icelandic-Norwegian in Table 1) into (6), they estimated $\varepsilon = (a + b)^{-1} = 1750$ y and $\gamma = (a + b)/a = 1.09$, from which they calculated[4] that $a \approx 5 \times 10^{-4}$ y^{-1} and $b \approx 6 \times 10^{-5}$ y^{-1} (Petroni & Serva 2008, Serva & Petroni 2008). There are three significant problems with (5) and (6): (i) $D' \leq 0$ for D in the appropriate range ($0 \leq D \leq 1$), and so (ii) D is always negative for the initial condition used and the values of a and b specified, and (iii) (5) and (6) are inconsistent with positive values of both a and b. Ignoring these, the values of a and b obtained from $\varepsilon = (a + b)^{-1} = 1750$ y and $\gamma = (a + b)/a = 1.09$ are also inconsistent and should actually be 5.2×10^{-4} y^{-1} and 4.7×10^{-5} y^{-1}, respectively. This means that the smaller rate constant for convergence (b) was over-estimated by about 25% and a was only slightly under-estimated (by about 4%), so that both $(a + b)$ and $(a + b)/a$ were approximately correct (Petroni & Serva 2008, Serva & Petroni 2008).

I consider the implications of these models and the significance of the sensitivity of the estimation of divergence time to the linguistic distance. In particular, I argue that the calibration data used for Indo-European languages (Gray & Atkinson 2003, Petroni & Serva 2008, Serva & Petroni 2008) are likely to be unreliable for t greater than about 4500 y. It follows that a measure of linguistic distance with a smaller rate constant is necessary to overcome this limitation.

Two modifications of the model

To rectify the inconsistency between (5) and (6), the least change required is to write (5) as

$$D' = a(1-D) - bD = a - (a+b)D \qquad D(0) = 0\,, \tag{7}$$

taking D as nondimensional, to which the solution is (6). This model has two implications. First, the more similar two languages are the greater is the tendency for them to diverge (that is $D' \propto (1 - D)$) and the more different they are the more likely it is that they might converge (so $D' \propto D$). In other words, the rates of divergence and convergence in (7) are $a(1 - D)$ and bD, respectively. If $a \approx 5 \times 10^{-4}$ y^{-1} and $b \approx 6 \times 10^{-5}$ y^{-1} (Petroni & Serva 2008, Serva & Petroni 2008), then the rate of divergence is approximately $8.3/D$ times the rate of convergence for small $D > 0$. The second implication of (7) is that at equilibrium, when the rates of convergence and divergence are equal (or $D' = 0$), $D_{\text{eq}} = a/(a + b)$. As both a and b are positive, this means that D_{eq} must be less than 1 and, in ordinary circumstances, D might vary from 0 to D_{eq}. Using Serva and Petroni's values ($a \approx 5 \times 10^{-4}$ y^{-1}, $b \approx 6 \times 10^{-5}$ y^{-1}) once again, the upper limit of D is $D_{\text{eq}} \approx 0.893$, although it follows from (2) and (7) that $D_{\text{eq}} = \gamma^{-1} \approx 0.917$, so the approximation of a and b does lead to a slight underestimation of D_{eq}.

An alternative approach[5] is to assume a constant divergence rate (f) and a convergence rate that increases with D

$$D' = f - cD \qquad D(0) = 0 \tag{8}$$

[4] These approximations are given as they are reported, although the units have been added.

[5] To ensure that $D'(D = 0) > 0$ and $D'(D = 1) < 0$ it might be thought sufficient to introduce a constant rate of divergence (g) into (5), to yield $D' = g - a(1 - D) - bD,$ and assume that $b > g > a$. The extra term (g) represents a background rate of divergence that is not dependent on D. If $f = g - a$ and $c = b - a$ (9) reduces to (8), but in this case it is not possible to determine a because it is confounded. For this reason (8) is to be preferred.

then the solution is (6) with $\varepsilon = c$ and $\gamma = c/f$. Using the same values of ε and γ yields $c \approx 5.7 \times 10^{-4}$ y^{-1} and $f \approx 5.2 \times 10^{-4}$ y^{-1}, from which (i) the rate of divergence is about $0.9/D$ times the rate of convergence for $D > 0$ and (ii) $D_{\mathrm{eq}} = f/c = 0.912$. What is interesting about (8) is that is identical to (7) if $f = a$ and $c = a + b$, which is, within error, the case (Table 2). The most significant difference between (7) and (8) is how they are interpreted, but that leads to quite different views of important aspects of linguistic divergence (Table 2).

Table 2. Model-dependence of the equilibrium linguistic distance (D_{eq}) and the relative rates of divergence and convergence using using $\varepsilon^{-1} = 1750$ y and $\gamma = 1.09$.

Model	ε^{-1} (y)	γ	Parameters ($\times 10^4$ y^{-1})	D_{eq}	relative rates of divergence and convergence
eqn (1)	1750	—	$b = 5.7$	1.000	∞
eqn (7)	1750	1.09	$a = 5.2$ $b = 0.47$	0.917	$11.1/D$
eqn (8)	1750	1.09	$f = 5.2$ $c = 5.7$	0.917	$0.9/D$

[*] The slight difference between these estimates of D_{eq} is due to differences in the rounding of the corresponding parameter estimates.

How does this relate to the calibration data?

The calibration values listed in Table 1 are reasonably consistent, but they do not conform to a single curve (Figure 1A). Moreover, the calibrations obtained from lower and upper values of the ranges specified in Table 1 are also roughly consistent over much of the range (Figure 1A). This is, perhaps, unsurprising given that three of the upper values (those in brackets in Table 1) were obtained using the lower values to calibrate the *D-t* relationship (Gray & Atkinson 2003).

However, the two-point calibration (Petroni & Serva 2008, Serva & Petroni 2008) is based on coordinates that are relatively closely-spaced compared with the range of values available (Table 1) and they correspond to short divergence times. No reason is given for the selection of these particular coordinates and no source is cited for them. Even assuming they were considered to be the most reliable of the values in Table 1, the extrapolation of a curve derived from points at 1100 y and 1600 y to almost 8000 y (Petroni & Serva 2008, Serva & Petroni 2008) is surely questionable.

In the absence of any indication of differences in the reliability of these coordinates, it seems reasonable to use them all. Rearranging (6) and letting $\phi = \gamma^{-1}$ and $k = \varepsilon^{-1}$ yields a general form of (3)

$$D = \phi\left(1 - \exp\left(-kt\right)\right), \qquad (9)$$

which can be fitted to the values in Table 1 to obtain $\phi = 1.0 \pm 0.1$ changes site^{-1} and $k = (7 \pm 2) \times 10^{-4}$ y^{-1} for the lower values (the solid circles in Figure 1A) and $\phi = 0.90 \pm 0.05$ changes site^{-1} and $k = (8 \pm 1) \times 10^{-4}$ y^{-1} for the upper values (the open circles in Figure 1A). Because of the 'noise' in the calibration data (Table 1) that is apparent in Figure 1A, these estimates are not statistically different from those reported previously (0.917 changes site^{-1} and 5.7 ×

10^{-4} y^{-1} for ϕ and k, respectively (Petroni & Serva 2008, Serva & Petroni 2008)), but t is so great that even these small differences in ε have significant impact on the calibration (Figure 1A).

There is a significant change in the *D*-*t* relationship (9) as D increases: at small D the relationship is almost linear, but at larger D a small change corresponds to a large change in the estimated divergence time (Figure 1A). This can be expressed conveniently using the sensitivity (S) of the calibration (9)

$$S = \frac{d\ln D}{d\ln t} = \frac{t}{D}D' = \frac{kt\exp(-kt)}{1-\exp(-kt)} \qquad (10)$$

which measures the relative change in D corresponding to a relative change in t. For example, if $S = 0.5$, a 0.5% change in D corresponds to a 1% change in t. In the case of the *D*-*t* relationship, $S \approx 1$ when D is small and declines as divergence time increases (Figure 1B). This means, for example, that by about 4500 y, when $S \approx 0.1$, a 2% change in D (from $D = 0.87$ to $D \approx 0.89$, perhaps) would correspond to a 20% change in t (from 4500 y to 5400 y). By $t = 7300$ y (the greatest time in Table 1), $S \approx 0.03$ so a 0.1% change in D (from 0.895 to 0.896) would correspond to a tripling of t (from 7300 y to 24300 y).

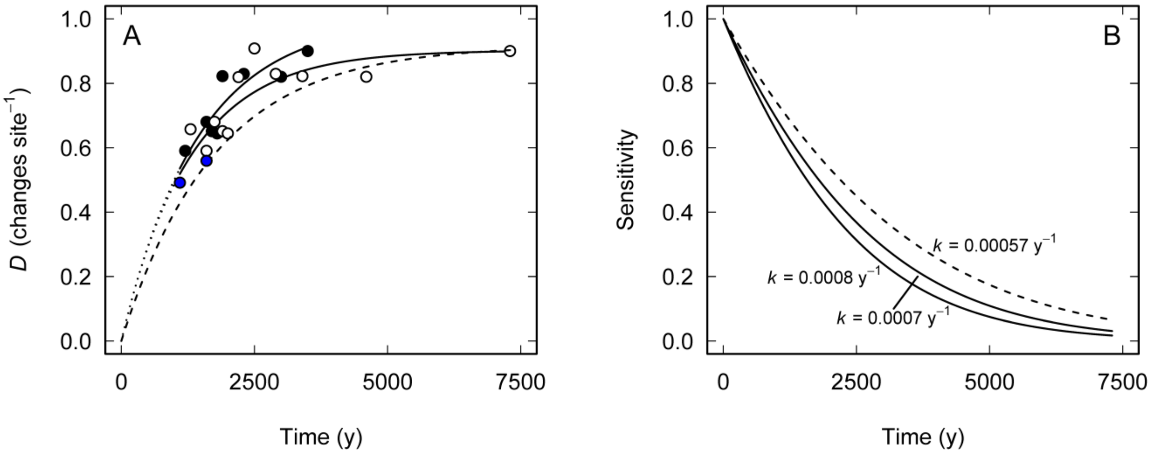

Figure 1. Relationship between D (A) or sensitivity (B) and the time since the divergence of Indo-European languages (based on the values in Table 1). The solid curves in (A) are nonlinear least squares fits of (9) to the data and the dotted curves are intended to emphasise the extrapolation of the relationships beyond the range of the data [● – lower values ($\phi = 1.0 \pm 0.1$ changes site^{-1}, $k = (7 \pm 2) \times 10^{-4}$ y^{-1}); ○ – higher values ($\phi = 0.90 \pm 0.05$ changes site^{-1}, $k = (8 \pm 1) \times 10^{-4}$ y^{-1})]. The dashed curve is that obtained from the two-point (●) calibration ($\phi = 0.917$ changes site^{-1}, $k = 5.7 \times 10^{-4}$ y^{-1} (Petroni & Serva 2008, Serva & Petroni 2008)). In (B) the time-dependence of the sensitivity corresponding to each of the three curves shown in (A) were calculated using (10).

This analysis illustrates three points. First, the extent to which the uncertainty in the measurement of D (Brown 2016) is propagated in the estimation of divergence time can be considerable. Second, the calibration obtained from data such as those in Table 1 is useful over a much more restricted range of D and t than appears to have been appreciated hitherto (Gray & Atkinson 2003, Petroni & Serva 2008, Petroni & Serva 2010). Such calibrations are unlikely to be reliable when $S < 0.1$ and even this may introduce unacceptable uncertainty into the estimated divergence time. Based on Figure 1B, a lower limit of $S \approx 0.1$ corresponds to a practical limit of $t \approx 4500$ y. Third, the sensitivity remains high for longer if the rate constant is smaller (Figure 1B) and this is independent of γ (10). This is not unexpected, but it invites the speculation that calibration based on lexical distance are suitable for short divergence times, and that other measures might be more suitable for longer timescales. It is also clear that the smaller rate constant estimated from only two calibration coordinates (Petroni & Serva 2008, Serva & Petroni 2008) is inconsistent with much of the calibration data[6] (Figure 1A).

The impact on the estimated divergence times of the differences in k shown in Figure 1A is very significant. For each ϕ and k (Figure 1A) the corresponding distribution of t can be determined from (9) using the 1225 values of D obtained from the pairwise comparison of the 50 Indo-European languages in the word lists (Petroni & Serva 2008, Serva & Petroni 2008). Using $\phi = 0.917$ changes site^{-1} and $k = 5.7 \times 10^{-4}$ y^{-1} yields a bimodal distribution, similar to that reported previously for these values (Petroni & Serva 2008, Serva & Petroni 2008), that extends to a maximum divergence time of about 8000 y (Figure 2A). The larger rate constants in Figure 1A progressively and substantially decrease the maximum divergence time without altering the overall form of the distribution (Figures 2B and 2C). These differences do not affect the structure of the phylogram which is calculated using D rather than t.

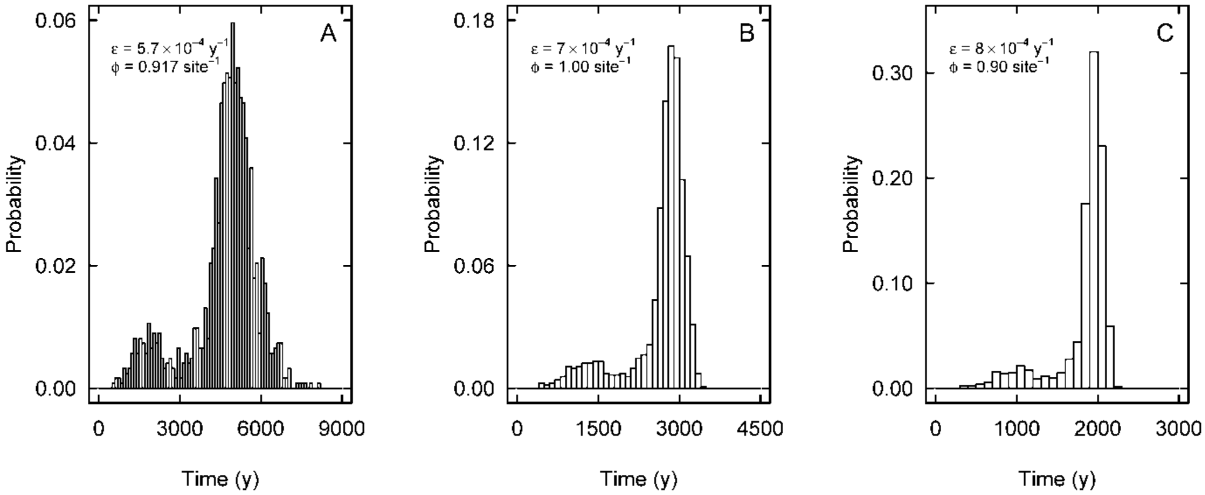

Figure 2. Distributions of divergence times estimated for Indo-European languages using the three calibration curves shown in Figure 1A. Values of D were calculated in R (Ihaka & Gentleman 1996) from the word lists (n = 200) of Serva and Petroni (2008). The (unweighted) Levenshtein distance was calculated using the stringdist package (van der Loo 2014) and was then normalised to the length of the longer of each pair of words. The divergence time was the calculated from D using (6) and the indicated values of ε and γ. In each panel the probabilities are represented in 100 y periods.

[6] It is possible that the estimates of D differ from those obtained by Serva and Petroni, despite using the same word lists. However, the distribution of D shown in Figure 2A agrees well with that previously reported (Petroni & Serva 2008, Serva & Petroni 2008).

A constraint on the rate constant

Usefully, the symmetry of (10) means that the sensitivity of D to changes in k is just

$$S = \frac{d\ln D}{d\ln t} = \frac{d\ln D}{d\ln k}, \tag{11}$$

so it is a simple matter to estimate the ideal value of k (Figure 3). For example, if it is considered desirable that $S \geq 0.1$ at 10000 y, then $k \leq 3.62 \times 10^{-4}$ y^{-1}. Interestingly, this limit is consistent with values of the rate constant for word replacement recently estimated using word use frequency for the three language pairs formed from Hittite, Homeric Greek and modern Greek[7] (Altschuler *et al.* 2013) and for about 70% of the words in the lists (n = 200) used by Pagel *et al.* (2007). Similarly, Lees (1953) estimated an average rate constant (b in (1)) equivalent to about 2 × 10^{-4} y^{-1} (the range was 1.5-3.3 × 10^{-4} y^{-1}). This contrasts with Sankoff's (1970) analysis of Indo-European which suggests that the rate constant was less than this limit for only about 35% of 1077 meanings.

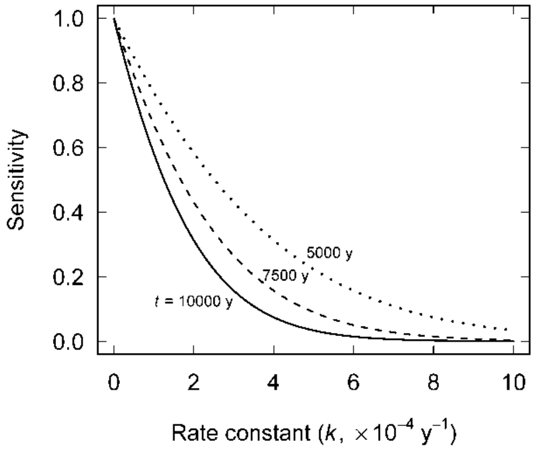

Figure 3. Sensitivity of distance (D) to the rate constant (k) for selected values of the divergence time (t). Sensitivity was calculated using (10) given the equivalence relation (11).

Some observations and implications

One assumption implicit in this analysis and in others like it (Petroni & Serva 2008, Serva & Petroni 2008) is that the rate constant for divergence (k) does not change. While the calibration data used here (Table 1) might be roughly consistent with this assumption (Figure 1A), it is clearly a matter of convenience because small differences in k can have very significant effects (Figures 2 and 3). Of course, the rate constant need not be uniform in time (Atkinson *et al.* 2008, Nettle 1999) or space (Atkinson 2011) or between words (Pagel *et al.* 2007) or meanings (Sankoff 1970). This should be considered in the context of the distribution of lexical distance, and the uncertainty associated with it (Brown 2016).

Three other observations are prompted by Figure 1A. First, the curves fitted to the smallest and largest divergence estimates are more similar to one another than either is to that used by Serva and Petroni (Petroni & Serva 2008, Serva & Petroni 2008). Moreover, the calibration

[7] Unfortunately, confusion between rate and rate constant combined with an absence of units (I assume that the values they give should be multiplied by 10^{-4} y^{-1}) complicates the interpretation of these data.

points on which the latter is based are are (i) relatively closely-spaced compared with the range of values available (Table 1) and (ii) correspond to short divergence times. This raises questions about the implementation of the calibration, such as whether it is desirable to base the calibration on just two values when several are available. This issue merits further consideration. The second observation is that there are no calibration values within the last millennium, although the data indicate that about 50% of the divergence occurs over the course of only 1000 y (Figure 1A). More recent calibration data would probably reduce the uncertainty of the parameter estimates, especially k. The third observation is that the fits (and the models) imply that the upper limit of D is less than 1. If this is correct, then 'unrelatedness' is not restricted to language pairs for which $D \approx 1$, but for which $D \approx D_{\mathrm{eq}}$ which appears to be closer to 0.9 than to 1 (Akulov 2015b, Brown 2015).

Conclusions

The sensitivity (S) of the calibration of lexical distance is such that a very small change in D corresponds to a large change in t over much of the range of the data (Figure 1A). This is undesirable, but is especially problematic when the measurement of D is questioned (Akulov 2015b, Hoijer 1956, Matisoff 1990, Rea 1958, Teeter 1963). The inference is that lexical distance is not a very robust measure of the timescale of language development, especially for times in excess of 4500 y. It would be better to base the measure of divergence on something that changes less rapidly than lexical distance. For example, if it is considered that a more robust measure would be characterised by values of S greater than 0.1 at 10000 y, then k should be less than about 3.6×10^{-4} y^{-1}, which is about half of the rate constants obtained from the lexical distance calibration data (Figure 1A).

References

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

Altschuler E. L., Calude A. S., Meade A., Pagel M. 2013. Linguistic evidence supports date for Homeric epics. *BioEssays*, vol. 35; pp.: 417-420

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

Atkinson Q. D., Meade A., Venditti C., Greenhill S. J., Pagel M. 2008. Languages evolve in punctuational bursts. *Science*, vol. 319; pp.: 588

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

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

Dobson A. J. 1978. Evolution times of languages. *Journal of the American Statistical Association*, vol. 73 no. 361; pp.: 58-64

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

Hoijer H. 1956. Lexicostatistics: a critique. *Language*, vol. 32 no. 1; pp.: 49-60

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

Matisoff J. A. 1990. On megalocomparison. *Language*, vol. 66 no. 1; pp.: 106-120

Nettle D. 1999. Is the rate of linguistic change constant? *Lingua*, vol. 108; pp.: 119-136

Pagel M., Atkinson Q. D., Meade A. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, vol. 449; pp.: 717-720

Petroni F., Serva M. 2008. Language distance and tree reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008 no. 8; pp.: P08012

Petroni F., Serva M. 2010. Measures of lexical distance between languages. *Physica A*, vol. 389; pp.: 2280-2283

Rea J. A. 1958. Concerning the validity of lexicostatistics. *International Journal of American Linguistics*, vol. 24 no. 2; pp.: 145-150

Sankoff D. 1970. On the rate of replacement of word-meaning relationships. *Language*, vol. 46 no. 3; pp.: 564-569

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

Teeter K. V. 1963. Lexicostatistics and genetic relationships. *Language*, vol. 39 no. 4; pp.: 638-648

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

van der Merwe N. J. 1966. New mathematics for glottochronology. *Current Anthropology*, vol. 7 no. 4; pp.: 485-500