

Two measures of linguistic distance

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

Linguistic relatedness is often assessed on the basis of lexical analysis using the normalised Levenshtein distance (LDN), but a variant of this is the LDN divided (LDND) has been used in a similar way. As LDND is the LDN normalised by the ‘global distance’ ($\Gamma(\alpha, \beta)$) between two languages, it is useful to consider both the properties of both LDND and $\Gamma(\alpha, \beta)$. Because $\Gamma(\alpha, \beta)$, like LDN, can not be greater than 1, LDND is ‘almost always’ greater than LDN and has no upper limit. However, for Indo-European word lists LDND is linearly related to LDN ($p < 0.001$) because $\Gamma(\alpha, \beta)$ has a very narrow distribution. Similar $\Gamma(\alpha, \beta)$ of were obtained in two numerical experiments based on (i) randomly generated ‘words’ or (ii) English words. This indicates that LDN may a better measure of lexical distance.

Key words: distribution; global distance; Levenshtein distance; lexical distance

Introduction

A common measure of linguistic relatedness is the normalised Levenshtein distance (LDN) derived from the lexical analysis of word lists. A related measure is the LDN divided (LDND) which was introduced to account for the possibility of accidental similarities¹ between languages (Bakker et al. 2009). Difficulties associated with the use of lexical distance to measure relatedness have been discussed (Akulov 2015b, Brown 2016b, Hoijer 1956, Matisoff 1990, Rea 1958, Teeter 1963), but until alternatives are established it is desirable that the best measure is employed.

The LDN between words i and j is

$$d(i, j) = \frac{d_L(i, j)}{L_{ij}}, \quad (1)$$

where d_L is the Levenshtein distance (Levenshtein 1966) between the two words and L_{ij} is the number of characters in the longer of the two words. The distance between two languages (α and β) is measured as the average LDN between n pairs of words

$$D(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^n d(\alpha_i, \beta_i), \quad (2)$$

where α_i and β_i represent word i in languages α and β , respectively.

¹ Originally these were phonological similarities as LDND is used in the ASJP database (<http://asjp.clld.org/>), but there is no particular reason why the same calculation should not be applied to standard word lists as Petroni and Serva (2010) have done.

The LDN divided (LDND), introduced to account for the possibility of accidental similarities between languages (Bakker et al. 2009), involves normalisation of LDN by a factor that has been called the ‘global distance’ (Petroni & Serva 2010)

$$\Gamma(\alpha, \beta) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n d(\alpha_i, \beta_j), \quad (3)$$

in which the distance between $n(n-1)$ pairs of words of different meanings are averaged². The expression for LDND is then

$$D_s(\alpha, \beta) = \frac{D(\alpha, \beta)}{\Gamma(\alpha, \beta)}, \quad (4)$$

from which it follows that $D_s(\alpha, \beta) \geq D(\alpha, \beta)$, because given (1), $0 \leq \Gamma(\alpha, \beta) \leq 1$.

The logical relationship between LDN and LDND is that the former is generated from pairs of words of the same meaning (2), whereas the latter is LDN normalised (4) by the average value of all the pairs of words that have different meanings (Figure 1). Petroni and Serva (2010) compared these measures, concentrating on their relationship with $\Gamma(\alpha, \beta)$, but they did not consider completely the relationship between the measures themselves. However, it may be that LDND is a better measure than LDN and it may overcome some of the difficulties associated with lexical distance (Akulov 2015b, Brown 2016b, Hoijer 1956, Matisoff 1990, Rea 1958, Teeter 1963). Here, the effect of the use of $\Gamma(\alpha, \beta)$ in the normalisation of lexical distance estimates based on LDN is considered.

		Language α			
		word 1	word 2	...	word n
Language β	word 1	$d(\alpha_1, \beta_1)$			$d(\alpha_n, \beta_1)$
	word 2				
	⋮				
	word n	$d(\alpha_1, \beta_n)$			$d(\alpha_n, \beta_n)$

Figure 1. Graphical representation of the relationship between LDN (= average of the values of the n black cells) and LDND, which is the ratio of LDN and $\Gamma(\alpha, \beta)$ (= average of the values of the $n(n-1)$ grey cells). Each shaded cell represents a specific pair of words between which the distance is $d(\alpha_i, \beta_j)$, as indicated in some of the cells.

Properties of LDN and LDND

Both LDN and the global distance ($\Gamma(\alpha, \beta)$) vary within a specific range. By definition $0 \leq d_L(i, j) \leq L_{ij}$, from which it follows from (1) that $0 \leq d(i, j) \leq 1$. As LDN is an average of n different $d(i, j)$ (2), the distance between languages α and β lies between 0 and 1 (or $0 \leq D(\alpha, \beta) \leq 1$). Similarly, as $\Gamma(\alpha, \beta)$ is an average of $n(n-1)$ different $d(i, j)$ (3), the global distance between languages α and β also lies between 0 and 1 ($0 \leq \Gamma(\alpha, \beta) \leq 1$). However, the global

² In the original statement of this normalisation, Bakker *et al.* (2009) suggest that the average is over $0.5n(n-1)$ pairs, but as the $n \times n$ matrix of $d(i, j)$ represented in Figure 1 is not symmetrical (it is almost always the case that $d(\alpha_i, \beta_j) \neq d(\alpha_j, \beta_i)$ for $i \neq j$), I follow Petroni and Serva (2010).

distance is actually less than 1 ($\Gamma(\alpha, \beta) < 1$) unless none of the word pairs have anything in common (in which case $d(i, j) = 1$ for all i and j where $i \neq j$)³.

In contrast, LDND has no specific upper bound. As LDND is the ratio of LDN and the global distance (4), it follows that the lower bound must be 0 ($D_s(\alpha, \beta) \geq 0$). However, the largest observed value of LDND depends on the specific values of LDN and the global distance. Moreover, that largest value applies only to the specific languages and word lists from which it is estimated. In this respect it is not generalisable.

While there it has no upper bound, it can be inferred from (4) that the value of LDND is almost always greater than that of the LDN on which it is based (2) simply because the global distance is almost always less than 1 ($\Gamma(\alpha, \beta) < 1$). The value of LDND is never less than that of LDN (4) because $\Gamma(\alpha, \beta)$ can not exceed 1 (3).

Interpreting LDND

No matter which languages are considered, how many words there might be in the word lists or how those words were chosen, both LDN and $\Gamma(\alpha, \beta)$ have a defined range. In contrast, the range of values of LDND depends on the set of languages considered and on the words in the list used. For example, the four pairs of Icelandic and English words in Set 1 (Table 1) yield a high LDN ($D = 0.958$) and a higher LDND ($D_s = 0.996$). For Set 2, in which just one word differs from Set 1, LDN is the same, but LDND is increased by about 10% ($D_s = 1.102$). For Set 3, which has no words in common with the other two sets, LDN is actually slightly smaller ($D = 0.938$), but LDND is increased by almost 30% ($D_s = 1.288$).

This example is based on a very small number of words ($n = 4$), but for larger numbers the effect is still apparent. The distribution of LDN for $n = 200$ words in 50 Indo-European languages ranges from about 0.25 to 0.91 (Figure 2A), whereas that for LDND ranges from about 0.29 to 1.01 (Figure 2B). In each case the distribution is bimodal and can be modelled as a weighted (p) sum of two normal distribution functions (Φ)

$$p\Phi(\mu_1, s_1) + (1 - p)\Phi(\mu_2, s_2), \quad (5)$$

where μ_i and s_i are the mean and standard deviation of component i (Figure 2), the parameters of which are given in Table 2. The means of the component distributions of LDN and LDND differ by 0.1, indicating that the distribution of LDND is located 0.1 units higher than that of LDN. The standard deviation of component 1 is 0.02 for both LDN and LDND, but for component 2 the standard deviation is greater for LDND than it is for LDN. This indicates that the lower peak has been broadened by normalising to the global distance, but that the main peak has been unaffected. The fraction of the density in component 1 (p) is essentially the same for LDN and LDND (Table 2).

The values of the global distance range from about 0.827 to 0.928 and the distribution is approximately normal (Figure 3). It is notable that the standard deviation is 0.01 (a coefficient of variation of 1.1%), consistent with the very narrow range of variation.

The relationship between the distributions shown in Figures 2 and 3 is more clearly seen from the cumulative distributions (Figure 4). The two components of LDN are apparent as the

³ It is also the case that LDN is actually less than 1 ($D(\alpha, \beta) < 1$) unless $d(i, i) = 1$ for all i (Brown 2015).

relatively broad shoulder rising to less than 0.2 and the steeply rising feature rising to 1. The components of LDND are similar, but the shoulder is broader than that of LDN, consistent with the larger standard deviation (Table 2). The cumulative distribution of the global distance lacks the shoulder of LDN and LDND, and consist of a single steeply increasing component.

Table 1. Comparison of measures derived from three sets of words selected from the English and Icelandic⁴ word lists of Serva and Petroni⁵ (2008). The (unweighted) Levenshtein distance was calculated using the stringdist package (van der Loo 2014) as described previously (Brown 2016a) in R (Ihaka & Gentleman 1996).

	$d(\alpha_i, \beta_j)$				$D(\alpha, \beta)$	$\Gamma(\alpha, \beta)$	$D_s(\alpha, \beta)$
Set 1	<i>dyr</i>	<i>lelegur</i>	<i>svartr</i>	<i>stor</i>			
<i>animal</i>	1.00	1.00	1.00	1.00	0.958	0.962	0.996
<i>bad</i>	1.00	1.00	0.83	1.00			
<i>black</i>	1.00	0.86	0.83	1.00			
<i>big</i>	1.00	0.86	1.00	1.00			
Set 2	<i>dyr</i>	<i>lelegur</i>	<i>svartr</i>	<i>mold</i>			
<i>animal</i>	1.00	1.00	1.00	0.83	0.958	0.869	1.102
<i>bad</i>	1.00	1.00	0.83	0.75			
<i>black</i>	1.00	0.86	0.83	1.00			
<i>earth</i>	0.80	0.86	0.50	1.00			
Set 3	<i>hann</i>	<i>hver</i>	<i>veioa</i>	<i>huo</i>			
<i>he</i>	0.75	0.50	0.80	0.67	0.938	0.728	1.288
<i>who</i>	1.00	1.00	0.80	0.67			
<i>hunt</i>	0.50	0.75	1.00	0.50			
<i>skin</i>	0.75	1.00	0.80	1.00			

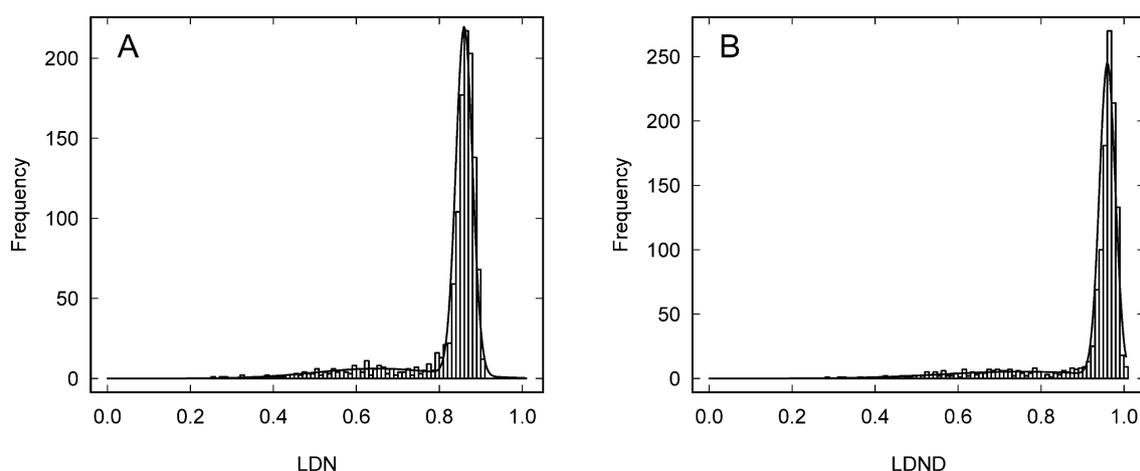


Figure 2. Distributions of LDN (A) and LDND (B) for 50 Indo-European languages. Values were calculated using the word lists ($n = 200$) of Serva and Petroni (2008). The curves are the sum of two normal distribution functions (the parameters are given in Table 2).

⁴ The Icelandic words are written as they are in the word list for consistency with the $d(i, j)$ specified.

⁵ Available at <http://univaq.it/~serva/languages/languages.html>.

Table 2. Parameters describing the approximations of the distributions of LDN, LDND and $\Gamma(\alpha, \beta)$ shown in Figures 2 and 3.

Component	Parameter	LDN	LDND	$\Gamma(\alpha, \beta)$
1	mean (μ_1)	0.86	0.96	0.90
	standard deviation (s_1)	0.02	0.02	0.01
	proportion (p)	0.825	0.82	1.00
2	mean (μ_2)	0.65	0.75	—
	standard deviation (s_2)	0.15	0.19	—

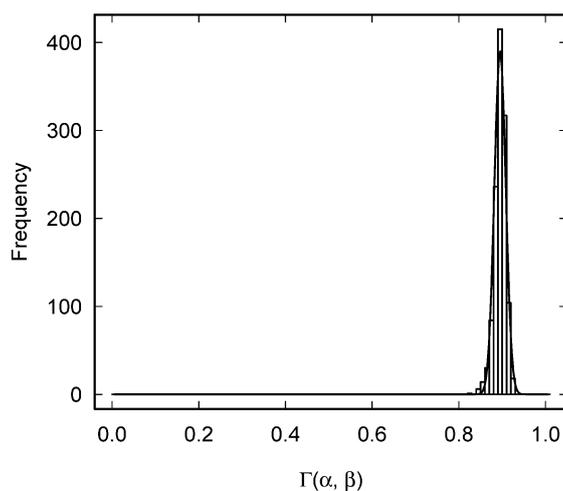


Figure 3. The distribution of the global distance for 50 Indo-European languages. Values were calculated using the word lists ($n = 200$) of Serva and Petroni (2008). The curve is the normal distribution function (the parameters are given in Table 2).

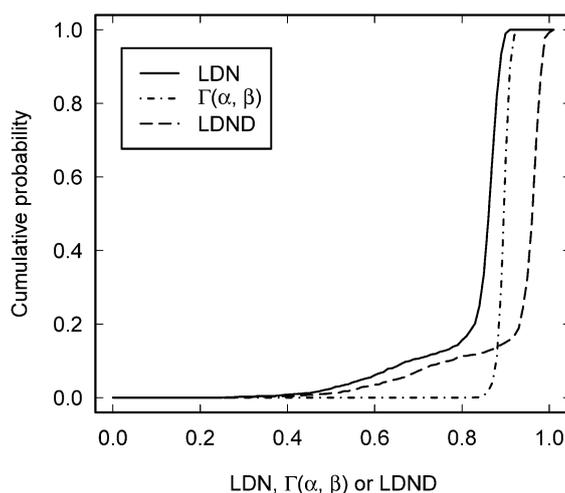


Figure 4. Cumulative distributions of LDN (—), the global distance (· · · ·) and LDND (— — —) for 50 Indo-European languages. Values were calculated from the corresponding distributions shown in Figures 2 and 3.

While it is apparent from Figure 2 that the distribution of LDND is located 0.1 unit higher than that of LDN, this does not make clear whether there is any consistent pattern in the extent to which these variables differ for individual language pairs. However, a plot of LDND against LDN for 50 Indo-European languages ($n = 1225$) is linear ($r^2 = 0.987$, $p < 0.001$) and ordinary least squares regression⁶ yields estimates of the gradient and intercept of 1.075 ± 0.007 (95% CI) and 0.034 ± 0.006 (95% CI), respectively (Figure 5). The diagonal line in Figure 5 represents equality between LDN and LDND, and it is apparent that the data lie above this line and are not quite parallel to it, consistent with the least squares regression coefficients. The mean deviation from equality estimated from the regression is 0.08 ± 0.01 , roughly consistent with the difference between LDN and LDND estimated from their distributions (Figure 2). The mean $\Gamma(\alpha, \beta) = 0.90$ (Table 2) is consistent with this because the expected difference ($\approx (1/0.9) - 1$, based on (4)) is about 0.1.

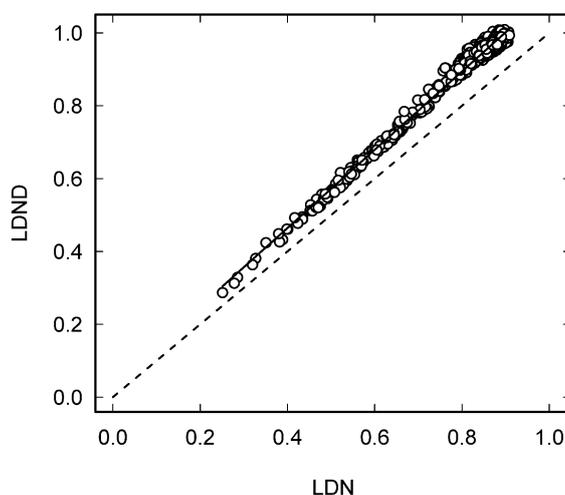


Figure 5. The correlation between LDND and LDN (\circ) for 50 Indo-European languages. Values were calculated from the corresponding distributions shown in Figure 2. The dashed line represents equality ($\text{LDND} = \text{LDN}$) and the solid line is the least squares regression (intercept = 0.034 ± 0.006 (95% CI), gradient = 1.075 ± 0.007 (95% CI)).

Two numerical experiments

The distribution of the global distance derived from 50 Indo-European languages is interesting for two reasons. First, the distribution is narrow (Figure 4 and Table 2). Specifically, the standard deviation is only 0.01, which is 1% of the possible range ($0 \leq \Gamma(\alpha, \beta) \leq 1$). Second, the values are relatively high (Figure 4 and Table 2). The data indicate that $\Gamma(\alpha, \beta)$ ranges from about 0.827 to 0.928, which, based on LDN calibration data (Brown 2016b), represents relatively little similarity. As the normalisation by $\Gamma(\alpha, \beta)$ is intended to address the issue of accidental similarities between languages (Bakker et al. 2009), it might be anticipated that some similarities might have been identified among the 50 Indo-European languages considered here.

⁶ Both LDN and LDND have associated uncertainties (Brown 2016a), whereas ordinary least squares regression assumes that this is true of only the dependent variable. Taking the uncertainty in both LDN and LDND into account (Kummell 1879) yields a gradient and an intercept of 1.083 ± 0.007 (95% CI) and 0.028 ± 0.005 (95% CI), respectively.

To clarify these issues two numerical experiments were performed in R (Ihaka & Gentleman 1996). The first involved the generation of random words of different lengths was used to examine the range of values of distance. The second, intended to contrast similarity and difference, involved a list of English homophones. The point is that members of a set of homophones are often very similar to one another (such as *aisle/isle*), but are different from the members of other sets (compare *aisle/isle* and *buy/by*). This provides a means of contrasting similarity and difference.

In the first experiment ‘words’ of random lengths were generated in pairs using all 26 letters of the Latin alphabet. To constrain the simulation the number of characters in each word was similar to that in the Indo-European word list used previously. In that list the number of characters per word averages 4.9 ± 1.9 (SD) and ranges from 1 to 26, similar to other reports (Eroglu 2013, Miller et al. 1958). No attempt was made to match the distribution of individual letters which is relatively complex (Bourne & Ford 1961). Based on this 100000 pairs of words were randomly generated using (i) the uniform and (ii) the discrete gamma distributions (Chakraborty & Chakravarty 2012) to determine the lengths and each character was similarly generated using the uniform distribution so that all letters were equally likely.

The distribution used to determine word length made little difference to the distribution of the distances between words, so the results given are those obtained using the uniform distribution. The $d(i, j)$ ranged from about 0.5 to 1, but for about 15% of the word pairs $d(i, j) = 1$ and $d(i, j) < 0.65$ for less than 1% of pairs. The average $d(i, j)$ was 0.91 ± 0.06 (SD) and it made no significant difference if those pairs for which $d(i, j) = 1$ were omitted (0.89 ± 0.05 (SD)). The average distance obtained from this simulation is not significantly different from $\Gamma(\alpha, \beta)$ estimated from the Indo-European word lists (Table 2). On the other hand, the standard deviation is somewhat larger (0.06 rather than 0.01). This might reflect the neglect of the distribution of letters in the simulation, but it is much more likely that it reflects the fact that each of the 1225 $\Gamma(\alpha, \beta)$ represented in Figure 3 is an average of 39800 ($= 200 \times 199$) estimates of distance between word pairs and the standard deviation in Table 2 is the average of these. Irrespective of this, the mean distance obtained from the simulation is not significantly different from $\Gamma(\alpha, \beta)$ estimated from the word lists (Table 2).

The second experiment used a list of English homophones⁷ which was edited to remove any entry giving more than two words (such as *air/ere/err/heir*), to avoid similarities within each of the final lists, or involving a contraction (such as *I'll*), and to eliminate any duplication of words. The result of this procedure was a list of 607 pairs of words from which two lists were generated using one member of each pair. Using these lists, a measure of similarity was obtained using $d(i, i)$ and a corresponding measure of difference was obtained using $d(i, j)$ for $i \neq j$ (in each case $i = 1, 2, \dots, 607$).

The distributions of the distances obtained from this experiment are quite different for homophones and non-homophones (Figure 6). The average distance between the homophone pairs was 0.36 ± 0.18 (SD), whereas it was 0.51 ± 0.01 (95% CI) greater ($p < 0.001$) for the non-homophone pairs (Table 1). Despite this, the range of the distances was similar for homophones and non-homophones (Table 3), although this is not apparent from the distributions because small distances were very much less likely for the non-homophones

⁷ The list (available at <http://www-01.sil.org/linguistics/wordlists/english/>) is reportedly based on the work of Townsend (1975).

(Figure 6B) than for homophones (Figure 6A). The distance between the non-homophone pairs (0.88 ± 0.13) was not significantly different from the $\Gamma(\alpha, \beta)$ estimated from the Indo-European word lists (Table 2).

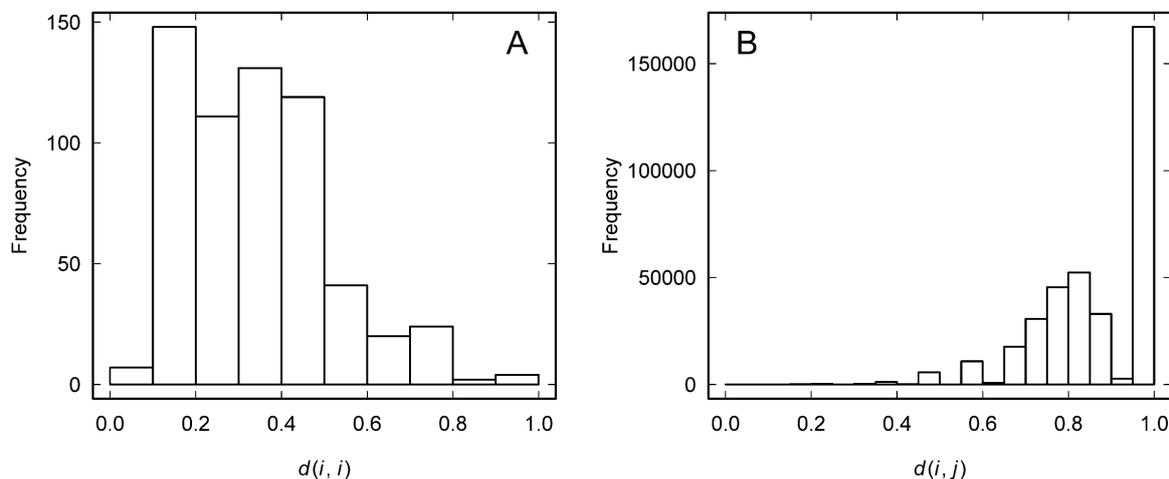


Figure 6. Distribution of distances between pairs ($d(i, i)$) of American English homophones (A) and between non-homophones ($d(i, j)$, $i \neq j$) in the same lists (B). In (A) $n = 607$ and in (B) $n = 367842$, and so the width of the bars is greater in (A) than in (B). Details of the distributions are given in Table 3.

Table 3. Details of the distributions of distances between American English homophones and non-homophones shown in Figure 6.

	Homophones ($d(i, i)$)	Non-homophones ($d(i, j)$, $i \neq j$)
mean	0.36	0.88
standard deviation	0.18	0.13
range	0.08-1.00	0.14-1.00
n	607	367842

Some observations and implications

Neither LDN nor the global distance can exceed 1 or be smaller than 0 (2-3), but LDND has no upper bound, so it can, and does (Table 1 and Figure 2B), exceed 1. Consistent with this, LDND is almost always greater than LDN. The natural inference is that LDND is not an absolute measure, it depends on the composition of the word lists (Table 1). Given this, care is required when comparing values obtained with different lists or in using LDND to calibrate linguistic distance (Brown 2016b). For example, it has been argued that ‘unrelatedness’ is not restricted to language pairs for which $D \approx 1$, but appears to be closer to $D \approx 0.9$ (Brown 2016b). As LDND is has no upper limit, this does prompt one to ask how it might be used to assess relatedness. In contrast, LDN is, at least notionally, a more general measure, albeit one with deficiencies (Akulov 2015b, Brown 2015, Brown 2016b).

The difference between LDN and LDND appears to be limited for the word lists ($n = 200$) for each of 50 Indo-European languages examined, although the relationship between them can depend to some extent on the composition of the word lists (Table 1). The distribution of LDND is similar to that of LDN, but is located 0.1 unit higher (Figure 2 and Table 2). This is consistent with the very narrow distribution of the global distance ($\Gamma(\alpha, \beta)$) around an average of 0.9 (Figure 3 and Table 2) because (4) implies that the ratio of LDND and LDN is given by $1/\Gamma(\alpha, \beta) \approx 1.11$. However, while LDND is linearly related to LDN ($p < 0.001$), the difference between them is slightly larger for higher values of LDN, because the gradient is statistically significantly greater than 1 (Figure 4). Nevertheless, based on this regression the average difference between LDND and LDN is 0.08, similar to the difference between their distributions (Figure 2).

The distribution of the global distance ($\Gamma(\alpha, \beta)$) obtained from 50 Indo-European language word lists is not only relatively large, but the range is surprisingly narrow (Table 2). It is interesting that the mean value is about 0.9, close to the upper limit of LDN previously identified (Brown 2016b). This might be taken to indicate that the level of similarity between the word pairs used to calculate $\Gamma(\alpha, \beta)$ (Figure 1) is low. Two numerical experiments were carried out to examine this issue. In the first of these the average distance between randomly generated words was about 0.9 and, in the second, the average distance between homophones was about 0.36 and that between non-homophones was close to 0.9, although the range over which the distances were distributed was similar for both homophones and non-homophones (Table 3). This indicates that the average global distance for Indo-European languages does not necessarily provide a sensitive quantitative distinction between random variation (in the first experiment) and lexical differences (in the second experiment). It may be that the global distance does provide a means of quantifying chance phonological resemblances between languages.

Conclusions

Some doubt has been expressed about the use of LDN to assess linguistic relatedness (Akulov 2015b, Brown 2016b), but it is unlikely that LDND is a better alternative lexical measure. It may be that LDND is a more effective measure in the context for which it was designed (Bakker et al. 2009). However, not only does LDND lack the upper bound that is intrinsic to LDN (2), but the normalising global distance does not appear to be useful in distinguishing between randomly generated words (in the first numerical experiment), English words (in the second experiment) and other Indo-European languages (Figures 3 and 6B). Until more robust measures of linguistic relatedness are established, as argued elsewhere (Akulov 2015a, Brown 2016b), LDN appears to be the better measure of lexical distance.

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

- Bakker D., Müller A., Vellupillai V., Wichmann S., Brown C. H., Brown P., Egorov, D., Mailhammer R., Grant A., Holman E. W. 2009. Adding typology to lexicostatistics: a combined approach to language classification. *Linguistic Typology*, vol. 13; pp.: 169 – 181
- Bourne C. P., Ford D. F. 1961. A study of the statistics of letters in English words. *Information and Control*, vol. 4; pp.: 48 – 67
- 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. 2016b. An analysis of the calibration of linguistic distance. I. Sensitivity. *Cultural Anthropology and Ethnosemiotics*, vol. 2 no. 2; pp.: 1 – 9
- Chakraborty S., Chakravarty D. 2012. Discrete gamma distributions: properties and parameter estimations. *Communications in Statistics – Theory and Methods*, vol. 41 no. 18; pp.: 3301-3324
- Eroglu S. 2013. Menzerath-Altmann law for distinct word distribution analysis in a large text. *Physica A*, vol. 392; pp.: 2775 – 2780
- Hoiyer 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
- Kummell C. H. 1879. Reduction of observation equations which contain more than one observed quantity. *Analyst*, vol. 6 no. 4; pp.: 97 – 105
- Levenshtein V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, vol. 10 no. 8; pp.: 707 – 710
- Matisoff J. A. 1990. On megalocomparison. *Language*, vol. 66 no. 1; pp.: 106 – 120
- Miller G. A., Newman E. B., Friedman E. A. 1958. Length-frequency statistics for written English. *Information and Control*, vol. 1; pp.: 370 – 389
- 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
- Serva M., Petroni F. 2008. Indo-European languages tree by Levenshtein distance. *European Physics Letters*, vol. 81; pp.: 68005

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

Townsend, W. C. 1975. *A handbook of homophones of general American English*. International Friendship, Waxhaw

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