

Some further Improvements to a Statistical Test useful in Historical Linguistics

Una nuova versione per un test statistico utile in linguistica storica

Cinzia Mortarino

Dipartimento di Scienze Statistiche, Università degli Studi di Padova
e-mail: mortarino@stat.unipd.it

Riassunto: In linguistica storica occorrono metodi per quantificare la similarità fra lingue attraverso il confronto di particolari liste di parole estratte dal vocabolario completo. In particolare, si pone il problema di valutare se il numero di elementi simili nelle due liste sia sufficientemente alto per rifiutare l'ipotesi di somiglianza casuale. Dopo un breve richiamo dei lavori presenti letteratura, in questo lavoro si propone una estensione in grado di utilizzare in modo più efficiente i dati disponibili. In particolare, ci si sofferma sul calcolo della distribuzione esatta della nuova funzione test sotto l'ipotesi nulla e si procede ad un confronto della potenza della nuova metodologia rispetto a quella già esistente.

Keywords: phonetic distance, nonparametric test, power comparison.

1. Introduction

Historical linguistics investigates ancient relationships among spoken languages, trying to look for similarities. In many situations the comparison between two languages is done with reference to a wordlist; in particular, it is possible to construct a special list of meanings which are found in all cultures, and which are sensibly assumed to be less subject to change and borrowing than the general vocabulary of the language.

Baxter and Ramer (2000) correctly treat this problem as a decision problem leading it into the context of a statistical test. In that paper a phonetic criterion is used to check whether there is a match between the words associated to each meaning in the two lists. Baxter and Ramer use a very general algorithm involving similarity of the *first consonant* only; words match if their initial consonants belong to the same one of the ten classes of consonants defined by Dolgopolsky (1986). The null hypothesis to be tested is that similarity is due to chance. Let X denote the random variable counting the number of matches in a list of length n . Although the authors do not explicitly state it, they calculate the distribution of X (the distribution of the test statistics under the null hypothesis) *conditioning on the actual lists* of both languages. Baxter and Ramer (2000) report that the probability distribution of X is quite complicated, but it is closely approximated by a Poisson distribution with parameter r , which depends on the wordlists to be compared.

In Mortarino (2004) a different approach is used to construct the distribution of the test statistics under the null hypothesis. Since linguists really would like to test whether phonetic class correspondences in words paired by meaning can arise by chance, it would be better to proceed without conditioning on the actual translations of both languages, but only conditioning on the theoretical list of meanings. In that case, under the null hypothesis, initial phonemes are, for any language, associated to meanings at random;

thus

$$X \sim \text{Bin}(n, p), \quad (1)$$

where p depends upon the distribution of phonemes among the classes. That value could be estimated either as 0.1565 or, using a sample of 1958 meanings translated in 10 languages, as 0.1293.

In section 2 a different distance function is described and the correct null hypothesis distribution function is calculated. In section 3 some results about approximated distributions and power comparisons are briefly discussed.

2. A different test statistics

The crucial problem of the procedure above described relies on the waste of information due to the use only of the first consonant of each translation. The approach here suggested allows to exploit much more information extending the comparison from one consonant to three. From the linguistic point of view, this is not just a matter of information quantity, but essentially of information quality, since the comparison of the first letter might be affected by prefixation rules which are typical of the grammar of each language and which could mask a true genetic relationship.

The new distance between two homologous translations is calculated as follows. The words are processed through a linguistic function, i.e. SOUND⁽¹⁾, retaining the first three phonetic consonants and classifying them into Dolgopolsky classes. Let $a_1a_2a_3$ and $b_1b_2b_3$ represent the strings that should be compared. The distance, D_3 , is computed as the sum of character differences; for a single comparison, D_3 ranges from 0 to 3. In order to evaluate the distribution function of D_3 , let's define ${}_1\pi_i$, $i = 1, 2, \dots, 10$, the probability that a phoneme of the i -th class could arise in position 1, i.e. at the beginning of the word. Similarly, let's define ${}_2\pi_i$ and ${}_3\pi_i$, $i = 1, 2, \dots, 10$, the probability that a phoneme of the i -th class could arise, respectively, in position 2 and 3. Since many words are very short, we have to consider also ${}_2\pi_V$ and ${}_3\pi_V$, the probability that positions 2 and 3, respectively, could be empty. If we denote

$${}_1\pi_{\bullet}^2 = \sum_{i=1}^{10} {}_1\pi_i^2, \quad {}_2\pi_{\bullet}^2 = \sum_{i=1}^{10} {}_2\pi_i^2 + {}_2\pi_V^2, \quad {}_3\pi_{\bullet}^2 = \sum_{i=1}^{10} {}_3\pi_i^2 + {}_3\pi_V^2, \quad (2)$$

after some calculations, it could be shown that

$$P(D_3 = 0) = {}_1\pi_{\bullet}^2 {}_2\pi_{\bullet}^2 [{}_2\pi_V + {}_3\pi_{\bullet}^2 (1 - {}_2\pi_V)] \quad (3)$$

$$\begin{aligned} P(D_3 = 1) = & {}_1\pi_{\bullet}^2 {}_2\pi_{\bullet}^2 (1 - {}_3\pi_{\bullet}^2) (1 - {}_2\pi_V) \\ & + {}_1\pi_{\bullet}^2 (1 - {}_2\pi_{\bullet}^2) [{}_3\pi_V {}_2\pi_V + {}_3\pi_{\bullet}^2 (1 - {}_2\pi_V)] \\ & + (1 - {}_1\pi_{\bullet}^2) {}_2\pi_{\bullet}^2 [{}_2\pi_V + {}_3\pi_{\bullet}^2 (1 - {}_2\pi_V)] \end{aligned} \quad (4)$$

$$\begin{aligned} P(D_3 = 2) = & {}_1\pi_{\bullet}^2 (1 - {}_2\pi_{\bullet}^2) [(1 - {}_3\pi_V) {}_2\pi_V + (1 - {}_3\pi_{\bullet}^2) (1 - {}_2\pi_V)] \\ & + (1 - {}_1\pi_{\bullet}^2) {}_2\pi_{\bullet}^2 (1 - {}_3\pi_{\bullet}^2) (1 - {}_2\pi_V) \\ & + (1 - {}_1\pi_{\bullet}^2) (1 - {}_2\pi_{\bullet}^2) [{}_3\pi_V {}_2\pi_V + {}_3\pi_{\bullet}^2 (1 - {}_2\pi_V)] \end{aligned} \quad (5)$$

$$P(D_3 = 3) = (1 - {}_1\pi_{\bullet}^2) (1 - {}_2\pi_{\bullet}^2) [(1 - {}_3\pi_V) {}_2\pi_V + (1 - {}_3\pi_{\bullet}^2) (1 - {}_2\pi_V)]. \quad (6)$$

⁽¹⁾ This function is implemented in the linguistic software STARLING, created by Prof. S. Starostin, inside the project "The Tower of Babel". The data here used were extracted from the multilingual database of that project (<http://starling.rinet.ru/babel.htm>).

Using a sample of 1300 meanings translated in 40 languages, the following estimate for D_3 distribution function was obtained:

${}_3d_i$	0	1	2	3
$\hat{P}(D_3 = {}_3d_i)$	0.003513	0.055728	0.311211	0.629549

(7)

D_3 represents the distance between the translations of a single meaning. If N meanings are considered, a single distance, S_N , is computed as the sum of N independent and equally distributed components. Its distribution could be evaluated through the probability generating function. The probability generating function (p.g.f.) of a random variable, X , taking values in the non-negative integers, is defined to be

$$G(u) = E(u^X). \quad (8)$$

For a discrete random variable with probability distribution

x_i	0	1	2	3
$P(X = x_i)$	a	b	c	$(1 - a - b - c)$

(9)

the p.g.f. is $G_X(u) = a + bu + cu^2 + (1 - a - b - c)u^3$. For a known result (see, for example, Grimmett and Stirzaker (1982, p. 79)), the p.g.f. of the sum of N independent and identically distributed random variables, S_N , is

$$G_{S_N}(u) = [G_X(u)]^N = [a + bu + cu^2 + (1 - a - b - c)u^3]^N. \quad (10)$$

By iterated applications of Newton's binomial theorem, we can obtain

$$G_{S_N}(u) = \sum_{n=0}^N \binom{N}{n} \left[\sum_{\ell=0}^{N-n} \xi(n, \ell) u^\ell \right] \cdot \left[\sum_{m=0}^n \psi(n, m) u^{2n+m} \right], \quad (11)$$

where

$$\xi(n, \ell) = \binom{N-n}{\ell} a^{N-n-\ell} b^\ell, \quad \psi(n, m) = \binom{n}{m} c^{n-m} (1 - a - b - c)^m. \quad (12)$$

After a rearrangement of terms in previous expression, it could be shown that the coefficient of u^j , i.e. the probability that S_N equals j , $j = 0, 1, 2, \dots, 3N$, is given by

$$P(S_N = j) = \sum_{t=0}^{t_{max}} q_{tj}, \quad t_{max} = \frac{(N+1)(N+2)}{2} - 1, \quad j = 0, 1, \dots, 3N, \quad (13)$$

where

$$q_{tj} = \binom{N}{k} \xi(k, y) \psi(k, r) \quad \begin{matrix} t = 0, 1, \dots, \frac{(N+1)(N+2)}{2} - 1 \\ j = 0, 1, \dots, 3N \end{matrix} \quad (14)$$

and

$$k = \begin{cases} 0 & \text{if } t = 0 \\ 1 + \sum_{i=2}^{t_{max}} I_{(\sum_{h=1}^i h, \infty)}(t) & \text{if } t > 0 \end{cases}, \quad r = t - \sum_{h=0}^k h, \quad y = j - 2k - r, \quad (15)$$

with the further restriction $\xi(k, y) = 0$ if $k + y > N$ or $y < 0$.

3. Some additional topics

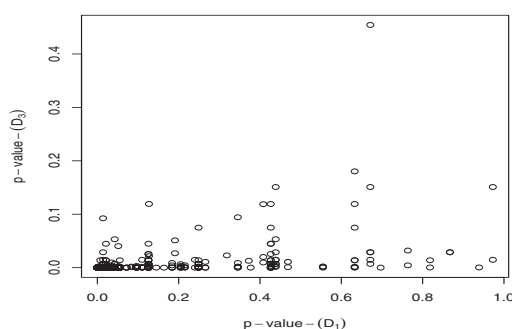
In this kind of application, N is usually around 30. A normal approximation of the exact probability distribution might, in general, be quite adequate. Attention must however be paid to the strong asymmetry of D_3 distribution. For $N=30$, the distribution of S_N still shows a non-negligible asymmetry.

In order to compare the efficiency of the test here proposed (based on D_3) with its simpler version (based upon the distance of the first consonant only, say D_1) both tests were applied to each pair of languages in a group of 40 IndoEuropean languages (780 comparisons). Figure 1 shows the scatterplot of p-values of the two tests. Since genetic relationships are well established into that group, it is evident that D_3 leads to a much more powerful test.

In order to assess whether estimates (7) for D_3 distribution are quite stable with reference to the sample used, a bootstrap procedure was used: estimates were calculated from 200 subsamples of length 50 from the whole list of 1300 meanings. The range of each estimate is quite narrow.

A final question was then addressed. The 33 item list of meaning is selected, as described above, according to linguistic criteria (i.e. stability in time and space). In order to see if the choice of the list could influence the power of the test, we proceeded as follows. Given a pair of languages, both tests were applied to 200 lists of length 33 randomly selected from the biggest 1300 item list. The p-values of both procedures have great fluctuations, but the new procedure leads to more stable values.

Figure 1: *P-values of the two procedures applied to pairwise comparisons among 40 IndoEuropean languages.*



References

- Grimmett G., Stirzaker D. (1982) *Probability and Random Processes*, Clarendon Press, Oxford.
- Baxter W. H., Ramer A. M. (2000) Beyond lumping and splitting: probabilistic issues in historical linguistics, in: *Time Depth in Historical Linguistics*, Renfrew C., McMahon A., Trask L. (Eds.), vol. 1, 167–188.
- Dolgopolsky A. B. (1986) A probabilistic hypothesis concerning the oldest relationships among the language families of northern Eurasia, in: *Typology, Relationship, and Time: a Collection of Papers on Language Change and Relationship by Soviet Linguists*, Shevoroshkin V. V., Markey T. (Eds.), Ann Arbor(MI), Karoma, 27–50.
- Mortarino C. (2004) A statistical test useful in historical linguistics, in: *Atti della XLII Riunione Scientifica SIS, Bari, 9-11 giugno 2004*, vol. 2, 107–110.