

A statistical test useful in historical linguistics

Un test statistico utile in linguistica storica

Cinzia Mortarino

Dipartimento di Scienze Statistiche

Università degli Studi di Padova

mortarino@stat.unipd.it

Riassunto: In linguistica storica occorrono metodi per quantificare la similarità fra lingue basandosi sul 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 aver illustrato un approccio già esistente in letteratura, in questo lavoro si propone una variante a tale metodologia illustrandone i vantaggi concettuali e applicativi.

Keywords: phonetic distance, nonparametric test.

1. Introduction

Historical linguistics is one of the many disciplines where statisticians might give their contribution to solve important problems. That discipline investigates ancient relationships among spoken languages, trying to look for similarities, and, most difficult, trying to explain whether similarities are due to *genetic relationship* (a common ancestor language which, for many reasons, split into two or more daughter languages), *borrowing* (different languages spoken by nearby or connected people end to share part of their vocabulary) or simply to *chance*.

At a first glance, if we think at the complexity of a language, chance might seem a weird motivation; this is not so, however, if we consider which aspects of a language are analyzed by historical linguistics to put it in relation with another one. 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. This list is known as the *basic core vocabulary* or the *Swadesh list* (see Swadesh (1950)); it consists of, e.g., body part terms, lower numerals, pronouns, basic actions, certain kinship terms, and naturally occurring non-cultural phenomena. The core vocabulary in many languages consists mostly of very short terms (monosyllabic terms) and since the number of sounds that a human vocal system can produce is limited, the probability of similar words in different languages arising by chance should not be neglected.

The comparison of two wordlists, in order to detect whether the two languages they are extracted from are somehow related, is a method very often used in historical linguistics since Swadesh first proposed it. That method gives however controversial results because the *threshold*, i.e., the minimum number of similarities in a list of length n necessary to conclude that the two languages are related, is set in an arbitrary manner: some linguists require that the threshold cannot be less than 40 (for $n=100$), while other lin-

guists maintain that 15 or 20 would be enough. No real motivation is anyway given for either cutoff.

There is only one reference in the historical linguistics literature, namely Baxter and Ramer (2000), which correctly treats this problem as a decision problem leading it into the context of a statistical test. In that paper the authors suggest to use a controlled list and explicit criteria to identify similarities; in particular, in that paper a *phonetic* criterion is used to check whether there is a match between the words in the two lists associated to each meaning: the number of matches obtained when words are paired by meaning is compared with the result obtained when words are paired at random. The null hypothesis to be tested is that the score obtained when pairing words by meaning is not significantly greater than the scores obtained when pairing words at random. For their test, Baxter and Ramer use a very general algorithm involving similarity of initial consonants only; words match if their initial consonants belong to the same one of the ten classes of consonants defined by Dolgopolsky (1986). For their example, Baxter and Ramer use a comparison between Modern Hindi and Modern English; the wordlist has length $n=33$. Each of the 33 words in English and Hindi is assigned to one of the ten Dolgopolsky classes. In that example, when the words in the two lists are paired by meaning, the algorithm defines $x=9$ pairs as phonetic matches. The next step in the procedure is to estimate the likelihood that a score this high (nine or more matches out of a possible 33) could have occurred by chance.

In section 2 the method used by Baxter and Ramer is presented. In section 3 a new alternative approach is discussed. The comparison between the two approaches is performed in section 4.

2. The method used by Baxter and Ramer (2000)

Let X denote the random variable counting the number of matches in a list of length n when the list items are paired at random. 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. In other words, the distribution of the words among Dolgopolsky classes is fixed for both languages. Let $f_{E,i}$, $i = 1, 2, \dots, 10$ denote the number of words in the i -th class actually included in the English wordlist; in analogous way, let $f_{H,i}$, $i = 1, 2, \dots, 10$ denote the number of words in the i -th class actually included in the Hindi wordlist; in the case analyzed by Baxter and Ramer we observe the following distribution among Dolgopolsky classes:

class symbol	P	T	S	K	M	N	R	W	J	Ø
$f_{E,i}$	5	7	3	1	1	3	1	3	2	7
$f_{H,i}$	4	4	3	8	3	3	0	0	1	7

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

$$r = \frac{1}{n} \sum_{i=1}^{10} f_{E,i} f_{H,i} \quad (1)$$

For the list they use, $r=128/33$ and $P(X \geq 9) = 0.018$.

The test can be performed also without calculating the X distribution, but estimating the p-value by using a simulation approach: *given the two lists*, items can be repeatedly paired at random and at each step the number of matches counted. Baxter and Ramer (2000) perform this analysis, obtaining a p-value 0.011. Since the number of matches when pairing by meaning is significantly higher than expected when pairing at random, the two languages turn out to be related.

3. An alternative approach

A different approach could be used to construct the distribution of the test statistics under the null hypothesis. As previously mentioned, the null hypothesis in Baxter and Ramer (2000) claims that the score obtained when pairing words by meaning is not significantly greater than the scores obtained when pairing words at random. Really, what we would like to test is whether similarities in words paired by meaning are due to chance; in other words, since the algorithm here used compares only the classes of the initial consonants of the two words, we would like to test whether phonetic class correspondences in words paired by meaning can arise by chance. In order to test this, it would be better to proceed without conditioning on the actual lists of both languages, but only conditioning on the theoretical list of meanings selected according to Swadesh's criterion. In that case, under the null hypothesis, for any language initial phonemes are associated to meanings at random; thus for each list item

$$P(\text{match}) = \sum_{i=1}^{10} \pi_{E,i} \pi_{H,i}, \quad (2)$$

where $\pi_{E,i}(\pi_{H,i})$, $i = 1, 2, \dots, 10$ denote the probability that a phoneme of the i -th class could theoretically arise in the English (Hindi) language.

If we assume independence among correspondences of the items in the list¹, then

$$Y \sim \text{Bin}(n, \sum_{i=1}^{10} \pi_{E,i} \pi_{H,i}). \quad (3)$$

Very often no real motivations exist for different π_i in different languages; moreover, a language specific approach would be of little use. For these reasons, we will assume that $\pi_{E,i} \equiv \pi_{H,i}$, $i = 1, 2, \dots, 10$. These quantities can be easily estimated with p_i , $i = 1, 2, \dots, 10$, the proportions of phonemes in the different classes:

class symbol	P	T	S	K	M	N	R	W	J	Ø
p_i	0.067	0.089	0.089	0.178	0.022	0.044	0.111	0.044	0.067	0.289

Hence the parameter in (3) can be estimated with $\sum_{i=1}^{10} p_i^2 = 0.1565$.

A more reliable (from the linguistic point of view) estimate for π_i , $i = 1, 2, \dots, 10$, can be based upon a big sample of words from different languages. Using a sample of 1958 meanings translated in 10 languages², the following distribution into phonetic classes was obtained:

¹This is not a restrictive assumption, since the theoretical list of meanings has been constructed avoiding "duplicated" meanings, i.e., couples of meanings closely related (like "what" and "who").

²The data here used were extracted from the multilingual database of the project "The Tower of Babel" designed by Prof. Starostin (<http://starling.rinet.ru/index2.htm>).

class symbol	P	T	S	K	M	N	R	W	J	Ø
q_i	0.131	0.144	0.174	0.106	0.051	0.029	0.098	0.078	0.014	0.174

leading to

$$\sum_{i=1}^{10} q_i^2 = 0.1293. \quad (4)$$

Similar values for expression (4) were obtained sampling from different groups of languages. For the example discussed by Baxter and Ramer, where 9 matches were observed, we would obtain

$$\begin{aligned} P(X \geq 9 | X \sim \text{Bin}(33, 0.1293)) &= 0.02115 \\ [P(X \geq 9 | X \sim \text{Bin}(33, 0.1565)) &= 0.06215] \end{aligned}$$

For the English/Hindi example the results of both approaches substantially agree: *for this particular case*, the Poisson distribution and the Binomial distributions are very similar, as a consequence of known results about the link between Poisson and Binomial random variables. This is not, however, a general feature of the two approaches, since the value of r in (1) depends on the actual wordlist used.

4. Conclusions

As outlined at the end of the previous section the two methods might in some cases lead to similar distributions for the test statistics. The main difference lays however in the argumentation used to define the problem as a statistical test. The approach here proposed seems to be more suitable for the specific problem, since the null hypothesis which is really tested is closer to the linguistic aim. Moreover the test is much more flexible since it is easier to apply it to different wordlists (with the same length, n , but with different items or with different lengths) because the X distribution does not depend on the actual lists; in addition it is possible to compare p-values obtained with respect to different couples of languages (with a conditional approach it would be meaningless). Since that test has to be used by linguists, the approach here proposed is much more simple: we have an exact distribution without the necessity to run simulations.

References

- Baxter W.H. and Ramer A.M. (2000) Beyond lumping and splitting: probabilistic issues in historical linguistics, in: *Time Depth in Historical Linguistics*, Renfrew C. McMahon A. and Trask L., eds., volume 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. and Markey T.L., eds., Ann Arbor(MI): Karoma, 27–50.
- Swadesh M. (1950) Salish internal relationship, *International Journal of American Linguistics*, 16, 157–167.