Sir Thomas Young
and
Statistical Evidence of Historical Relationship

William J. Poser
University of Pennsylvania

It has long been recognized that the demonstration of genetic affiliation between languages involves a statistical element, in that it is necessary to exclude the possibility of congruences between languages being due to chance, but it has generally been considered both difficult and unnecessary to engage in explicit statistical calculation. This is largely due to the belief that the comparative method as traditionally applied so reduces the possibility of chance as virtually to exclude it. The controversy that has latterly raged over certain claims of distant genetic relationship, where the evidence adduced is scant and no reconstruction of the history of the putatively related languages follows and confirms the claim of affiliation, has brought about a renewed interest in quantitative methods (e.g. Ringe 1992, Kessler 2001).

Although the mathematical aspects of comparative linguistics are still imperfectly understood, the subject has a long history, much of which appears, from the very spotty references one finds, to be generally unknown. In this note I wish to draw attention to, and comment upon, what I believe to be the first work on this topic, by Sir Thomas Young, M.D. (1773-1829).

Sir Thomas was a scientist and scholar of wide-ranging interests. Although he held a medical degree from Gottingen, his academic appointment was as professor of physics. His contributions include the wave theory of light, the theory of interference of light, the three primary color theory of color vision and the explanation for astigmatism. That Young should have adverted to a linguistic topic is not surprising. Having learned French, German, Latin, Greek, Hebrew and Arabic as a boy, he had a serious interest in linguistics. In 1813 he coined the term ‘Indo-European' by which this family has since been known in English (Koerner 1989:154-157). He made the first and only significant advance in the decipherment of Egyptian hieroglyphic writing prior to Champollion, correctly reading several *cartouches*; indeed, his method was the same as Champollion’s. Had he realized that the phonological component of Egyptian writing was not restricted to royal names, it would probably be Young and not Champollion whose name would be associated with the full decipherment of Egyptian.

Young’s treatment of linguistic relationship appears in a letter

to Captain Henry Kater, read on the 21st of January 1819 to the Royal Society, and subsequently published in its *Transactions* (Young 1819). The paper is devoted primarily to physics, but contains a brief digression on linguistics at pp. 79-82.

Young begins by considering a list of words and the permutations of that list. He computes the number of permutations that leave no words in place, one word in place, two words in place, and so forth. He does not give the formulae on which his computations are based explicitly, but his method is clear. Let us write *A(n,k)*for the number of permutations of an ordered set of n elements that leave *k* elements in place. Young first expressed *A(n,0)*, the number of permutations of an ordered set of *n* elements that leave no elements in place,[1] as the total number of permutations of the set (*n!*) minus the number of permutations that leave one element in place, the number that leave two elements in place, etc., to wit:

$$A(n,0) = n! - \sum_{k=1}^{n} A(n,k)$$

Then he computed the number of permutations of a set of *n* elements that leave *k* elements in place using the equation

$$A(n,k) = A(n-k,0)\binom{n}{k}$$

where $\binom{n}{k}$ is the number of subsets of size *k* of a set containing *n* elements. Since we know that

$$\binom{n}{k} = \frac{n!}{(n-k)!\ k!}$$

the two equations make it possible to build up a table of values of *A(n,k)*.

The probability of each class of permutation is then found by dividing the number of permutations in the class by the total number of permutations, *n!*, which for the list of ten words that he considers is 10! = 3,628,800. He also presents the probability of *k* or more words left in place, implicitly using the fact that

$$P[m \geq k+1] = P[m \geq k] - P[m = k]$$

---

[1] Young's own notation is A_n.

where *m* represents the number of "matches", that is, the number of elements left unchanged by the permutation.

Using the observations that as the number of words in the list becomes large, the probability of there being no coincidence approaches e^{-1} , and that *P[m=n] = P[m=n-1]/n*, Young then derived a similar set of probabilities valid for large *n*.[2] These are as follows:[3]

N	Exactly N Matches	N or More Matches
0	0.3678794	
1	0.3678794	0.6321206
2	0.1839397	0.2642412
3	0.0613132	0.0803015
4	0.0153283	0.0189883
5	0.0030607	0.0036600
6	0.0005109	0.0005943
7	0.0000730	0.0000834
8		0.0000105

Young assumed that the probability that *k* words match in lists from two languages is equivalent to the probability of permuting a list in such a way that *k* elements remain in place, and that the above probabilities could therefore be applied to language comparisons. He applied his technique to a comparison of Basque and Egyptian, adducing the following words as matches:

Basque	Egyptian	Gloss
berria	beri	new
guchi	kudchi	little
oguia	oik	bread
ora	uhor	dog
otsoa	uonsh	wolf
shaspi	shashf	seven

Since by his calculation the probability of six or more coincidences is 0.0005943, less than 1/1000, he takes these similarities to establish a historical relationship between

---

[2]A derivation of the fact that the probability of a permutation that leaves no element in place approaches e^{-1}may be found in Polya et al. (1983:33-34), where it is shown that the finite sums representing the probabilities of no coincidences for lists of length *n* are the first terms of the (infinite) Taylor series expansion of e^{-1}. Although we might fear that Young erred in applying an approximation derived by passing to an infinite limit to lists as small as ten words, in point of fact the approximation is already valid to six decimal places when $n=9$. Young's derivation of this result is not rigorous, but since it is correct there is little point in commenting on his argument here.

[3]Young did not give the probability of 0 or more matches, which is obviously 1.0, nor the probability of exactly 8 matches, which is 0.0000091. This latter is not actually an omission, though it may seem to be one from my arrangement of the table, which is not the same as Young's, which reflected the manner in which he comuted the values. Since the values of interest are the probabilities of *n* or more matches, and Young computed the value of this probability by subtracting the probability of exactly $n-1$ matches from the probability of $n-1$ or more matches, he did not need to know the probability of exactly 8 matches in order to compute the probability of 8 or more matches and so did not display

Egyptian and Basque (p.82, emphasis mine).

> ... *if we consider these words as sufficiently identical to admit of our calculating upon them*, the chances will be more than a thousand to one, that, at some very remote period, an Egyptian colony established itself in Spain: for none of the languages of the neighboring nations retain any traces of having been the medium through which these words have been conveyed.

Of course, a relationship between Egyptian and Basque is not generally accepted today, and those few who do accept such a relationship would consider it to result from the membership of both in a much larger family (e.g., 'Boreal' or 'Proto-World') within which Egyptian is much more closely related to the Afro-Asiatic languages, as generally held,
and Basque is much more closely related to certain languages of the Caucasus, Athabaskan, and Chinese, as maintained by some. From the point of view of virtually all scholars at present, Young's conclusion is erroneous.

Young's error was to apply an inapplicable mathematical model. His model would apply if all languages had the same phonological lexicon, that is to say, the same phonological inventory and the same phonotactics, differing only in the pairings of sound and meaning, and if only words **identical** in sound and meaning were counted as matching. Since, as his example shows clearly, he did not require anything near identity to count a pair of words as a match, and since it is not the case that languages have the same phonological lexicon, his mathematical model is inapplicable, and the probabilities he computed are useless for estimating the probability of chance resemblance. As I hope to show in a sequel to this note, Young was by no means the last author to fall victim to this error.

Young's mathematical model is inapplicable for another reason as well. The model assumes that the list of words is chosen randomly, but in fact Young has selected these six words precisely because they resemble each other. In order to apply his model correctly, it would be necessary to select the words to be considered a priori, then determine the number of matches. Since the number of small subsets of a lexicon is very large, the probability of finding a small set of matches is high. For example, the number of combinations of 1000 items taken six at a time is:

$$\binom{1000}{6} = \frac{1000!}{(1000-6)!6!} = \frac{1000*999*998*997*996*995}{720} = 1{,}368{,}173{,}298{,}991{,}500$$

well over a quadrillion. With a larger lexicon to choose from, the number of combinations is even larger.

Although Young failed to realize how the lack of identity between the Basque and Egyptian words he put forward invalidated his calculations, he was nonetheless dimly aware that the latitude allowed in determining a match affected the validity of the

calculation. This we can see in the contingency he expressed in the emphasized words in the passage above, as well as in his criticism of a comparison involving words he considered to be too different to count as matches.

> ... if we adopted the opinions of a late learned antiquary, the probability would be still incomparably greater that Ireland was originally peopled from the same mother country: since he has collected more than 100 words which are certainly Egyptian, and which he considers as bearing the same sense in Irish; but the relation, which he has magnified into identity, appears in general to be that of a very faint resemblance: and this is precisely an instance of a case, in which it would be deceiving ourselves to attempt to reduce the matter to a calculation. (p.82)

Young was mistaken both in the application of his mathematical results to practice and in the results he achieved, but he was fully aware that the demonstration that similarities between languages are unlikely to be due to chance is not in itself sufficient to establish genetic affiliation. As the following two passages show, he realized that historical connection between languages may be due either to common descent or to diffusion.

> Thus, if we were investigating the relations of two languages to each other, with a view of determining how far they indicated a common origin from an older language, *or an occasional intercourse between the two nations speaking them*...(p.79; emphasis mine:WJP)

> ...it would be more than 10 to 1 that they must be derived in both cases from some parent language, or *introduced in some other manner*...(p.82; emphasis mine:WJP)

Sir Thomas' greatest insight may well have been his recognition of the danger that incorrect mathematical arguments could be used to mislead the unwary, a prophecy that has, unfortunately, come true in our time.

> ...it would be extremely easy to pervert this application in such a manner, as to make it subservient to the purpose of clothing fallacious reasoning in the garb of demonstrative evidence. (p.79)

We would do well to heed these words.

Author's address:

William J. Poser
Department of Linguistics
619 Williams Hall
University of Pennsylvania
PHILADELPHIA, PA 19104-6305
U.S.A.

Email:wjposer@unagi.cis.upenn.edu

References

Kessler, Brett. 2001.
The Significance of Word Lists.
Stanford, Calif.: Center for the Study of Language and Information.

Koerner, E. F. Konrad. 1989.
Practicing Linguistic Historiography.
Amsterdam & Philadelphia: Benjamins.

PÓlya, George, Robert E. Tarjan & Donald R. Woods. 1983.
Notes on Introductory Combinatorics.
Boston: Birkauser.

Ringe, Donald. 1992.
On Calculating the Factor of Chance in Language Comparison.
Philadelphia: American Philosophical Society.

Young, Thomas 1819. "Remarks on the Probabilities of Errors in Physical Observations, and on the density of the earth, considered, especially, with regard to the reduction of experiments on the pendulum".
Philosophical Transactions of the Royal Society of London 109.70-95.