Leeds Studies in English

Article:

Ruby Roberts and Alan S. C. Ross, 'A Note on the Interpretation of the Statistics of Variant Forms in Philology', *Leeds Studies in English*, 2 (1933), 7-13

Permanent URL:
https://ludos.leeds.ac.uk:443/R/-?func=dbin-jump-full&object_id=134472&silo_library=GEN01

A NOTE ON THE INTERPRETATION OF THE STATISTICS OF VARIANT FORMS IN PHILOLOGY.[1]

There is a problem, arising frequently in many branches of philology, which may be typified by the following example.

In the Lindisfarne Gospels there is a variation in the nom. acc. pl. of masc. *-o*-stems between the older *-as* and a later *-es* due to weakening of *a* to *e*. The statistics[2] for the words recorded a 'fair number' of times (see below) are as follows:

	No. of forms in *-as*	No. of forms in *-es*
dæg	28	—
diowl	20	14
engel	11	4
fisc	5	4
heofon	9	—
hlaf	15	—
ðegn	153	—
ðreat	11	—
waras	18	—

The word *fisc* shows a comparatively high proportion of forms in *-es* (5 *-as*: 4 *-es* as against in all 270 *-as*: 22 *-es*). Are we entitled to use this fact as evidence to show that *a* tended to be especially weakened after *sc*?

It does not appear to have been realized hitherto that the solution of all such problems is primarily a mathematical, not a philological, problem: when different forms occur in varying proportions it is evident that no conclusion as to any particular case which appears to differ from the others can be drawn unless either most of the statistics agree in establishing some sort of norm, or some special theory is available to explain

[1] Our thanks are due to Mr. Glenny Smeal, Lecturer in Statistical Method and Computation in this University, for the assistance which he has so kindly given us on the mathematical side of the question.

[2] From H. C. A. Carpenter, *Die Deklination in der nordhumbrischen Evangelienübersetzung der Lindisfarner Handschrift* § 234.

their variation. In the absence of a theory the primary tests of agreement will necessarily be those based on chance occurrence; hence the preliminary qualifying condition is that the statistics (excluding the particular case) should be 'homogeneous.' The following may be recommended as a practical procedure for discovering whether a set of statistics of two variant forms is homogeneous or not:

(i) Exclude all cases in which the total number of forms of both types is less than 10.[3]

(ii) Arrange the remaining statistics in a table of the following type:

	No. of forms of Type A	No. of forms of Type A'	No. of forms of both Types
1st Case	a_1	a'_1	$a_1 + a'_1$
2nd Case	a_2	a'_2	$a_2 + a'_2$
...	..	..	...
...	..	..	...
...	..	..	...
...	..	..	...
...	..	..	...
...	..	..	...
...	..	..	...
...	..	..	...
...	..	..	...
...	..	..	...
...	..	..	...
mth Case	a_m	a'_m	$a_m + a'_m$
Total No.	n	n'	$n + n'$

[3] No conclusion of the type under discussion can be drawn from such statistics. Note that if the total number of cases is small this limit can be lowered to 9 or even 8.

(In this table a_1, a'_1 denote the number of forms of Types A and A' in the first of the cases considered—which are, in all, m in number; a_2, a'_2, . . . denote those in the second, . . . case etc.; n and n' denote the total number of forms of Types A and A'.)

(iii) In each case take the value of

$$\frac{(an'-a'n)^2}{a + a'}$$

(iv) Add all these values together and divide the sum by the product $n \times n'$. This is the function χ^2.

(v) In a table[4] of values of χ^2 look up the value given in the horizontal row against $(m-1)$ (i.e. one less than the number of pairs of statistics) and proceed along this row to the figure given in the vertical column under .05. If the value of χ^2 obtained in (iv) is greater than this figure in the table then the statistics are not homogeneous.

Note. If $(m-1)$ is greater than the largest value given at the beginning of a horizontal row take

$$\sqrt{2\chi^2} - \sqrt{2m-3} \qquad ;$$

if this is greater than 2 the statistics are not homogeneous.

If the statistics are not homogeneous no conclusion as to any particular case can be drawn from them. In actual practice it seems probable—at least from the small amount of material so far available for study—that philological statistics rarely approach anywhere near[5] satisfying this preliminary test for homogeneity and that consequently no special conclusion as to a particular form in a group can possibly be based on them.

Thus, applying the procedure to the statistics given above for the nom. acc. pl. of the masc. *-o-* stems in *-as*: *-es* in Lindisfarne:

(i) All the statistics except those for the words given above must be excluded because the total number of forms is less than 9.

(ii) Let Type A represent the forms in *-as*, Type A' those in

[4] Such as that found in R. A. Fisher, *Statistical Methods for Research Workers* (1933 edition), pp. 104-5.

[5] An extremely convenient consequence of the fact that the statistics usually fall so far short of satisfying the test is that, even if a comparatively large error has been made in compiling them, the conclusions reached will be unaltered.

-es. Let *dæg*=1, *diowl*=2, *engel*=3, *heofon*=4, *hlaf*=5, *ðegn*=6, *ðreat*=7, *waras*=8. Then we have the following table:

	No. of forms of Type A	No. of forms of Type A'	No. of forms of both Types	Value of $\frac{(an'-a'n)^2}{a+a'}$
1	28		28	9072
2	20	14	34	330074
3	11	4	15	49536
4	9		9	2916
5	15		15	4860
6	153		153	49572
7	11		11	3564
8	18		18	5832
Total	265	18	283	455426

(iv) From this table we have $m=8$, $n=265$, $n'=18$.

Sum of values of $\frac{(an'-a'n)^2}{a+a'} = 455426$.

Hence $\chi^2 = \frac{455426}{18 \times 265} = 95{\cdot}5$.

(v) From the table of values of χ^2 we find that the figure given in the 7th (i.e. $\overline{m-1}$th) row under the column .05 is 14.067, which is very much less than 95.5.

Consequently the statistics are not homogeneous and therefore no special conclusion as to the word *fisc* can be based on the fact that the forms with *-es* occur comparatively frequently. This fact is thus actually irrelevant to any consideration of the question whether *a* tended to be especially weakened after *sc*.

A simple example from the Ayenbite of Inwyt—a text which, by reason of the fact that it faithfully represents the dialect of a particular individual writing at a known time and place, is peculiarly well-suited to the study of variant forms—will serve to illustrate further the unhomogeneity of philological statistics. The statistics given below fall far short of satisfying the homogeneity test.

In the Ayenbite of Inwyt there is a variation between the spellings *c* and *k*. In the case where the consonant occurs initially before a front vowel or *n*, the Mn.E. conditions[6] obtain. Thus in the words *kende* (72 x), *keste* (20 x), *kyng* (64 x), *knawe* (91 x), *knygt* (38 x)[7] only forms with *k* are recorded. But in the cases where Mn.E. has *c* initially (i.e. before a back vowel, *l* or *r*) there is a variation between *c* and *k* in the Ayenbite. Applying the procedure described above:

(i) All the statistics except those given below must be excluded because the total number of forms is less than 10.

(ii) Let Type *A* represent the spelling with *c*, Type *A′* that with *k*. Let Case:

No. 1 = *clene*	No. 4 = *cloð*	No. 7 = *corn*
No. 2 = *clepie*	No. 5 = *come*	No. 8 = *creft*
No. 3 = *cliue*	No. 6 = *conne*	No. 9 = *crist*

[6] See O. Jespersen, *A Modern English Grammar* i, § 2.326.

[7] In these and the following examples derivatives are included.

Then we have the following table:

	No. of forms of Type A	No. of forms of Type A'	No. of forms of both Types	Value of $\frac{(an'-a'n)^2}{a+a'}$
1	106	17	123	818546
2	104		104	113256
3	27		27	29403
4	25		25	27225
5	282		282	307098
6	77	16	93	1117613
7	19		19	20691
8	14		14	15246
9	142		142	154638
Total	796	33	829	2603716

(iv) From this table we have $m = 9$, $n = 796$, $n' = 33$.

Sum of values of $\dfrac{(an' - a'n)^2}{a + a'} = 2603716$.

Hence $\chi^2 = \dfrac{2603716}{796 \times 33} = 99.1$.

(v) From the table of values of χ^2 we find that the figure given in the 8th (i.e. $\overline{m-1}$th) row under the column .05 is 15.507, which is very much less than 99.1. Consequently the statistics are not homogeneous.

Summarising, we may say that a preference for a particular variant in the case of a given word in a variation of this type cannot, in general, be considered on statistical grounds to have any philological significance unless the statistics are homogeneous and consequently satisfy the above mathematical test.[8]

RUBY ROBERTS.
ALAN S. C. ROSS.

[8] Some examples of arbitrary preference in the Lindisfarne Gospels are dealt with in an article by A. S. C. Ross, about to appear in *Modern Language Notes*.