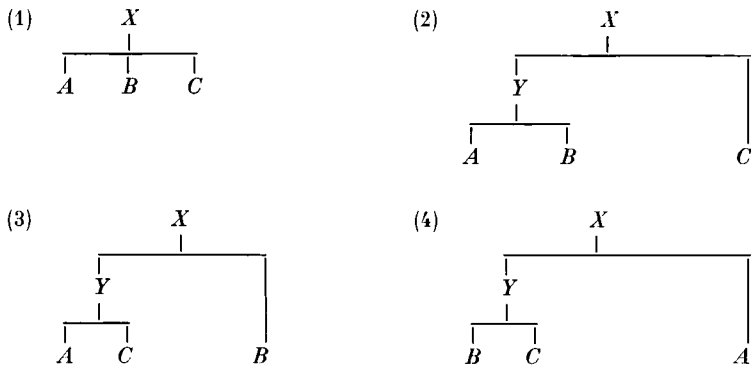
»Close-Relationship» in the Uralian Languages [1]

Suppose three related languages A, B, C. Let it be given that, of these three, no one is directly descended from one of the other two. Then, clearly, the three languages may be related in four ways:

TABLE I



(where X and Y are parent languages). In »Pattern of Descent» No. 2, the relationship between A and B is said to be »closer» than the relationship of C to either A or B. It seems to us that nearly all the evidence for assessing the closeness of relationship between two languages of a family of related languages is comprised in a table made up as follows. Consider the cognate words of the languages — that is, words appearing in two or more languages. Allot a column to each language and a row to each cognate word; if a word is present in a language put a cross in the appropriate cell of the table [2]. Thus:

---

[1] Abbreviation: Ross = A. S. C. Ross, »Philological probability problems», Journal of the Royal Statistical Society B. XII. 19—59.

[2] Cf. Ross pp. 26—27.

TABLE II

Word Number	1	2	3	4	Language Number
1	×	×			
2		×	×		
3	×	×	×		
.					
.					

In the present study we attempt to investigate close-relationship in the Uralian family of languages, using an inventory of words of the type referred to above. The inventory was taken from B. Collinder, Fenno-ugric vocabulary. And, in this context, we should observe, first, that, although it may well be that this source is not perfect, it is nevertheless the only one which is conveniently available. Secondly, if there are omissions in it, as is quite possible, then our conclusion will have been based on only a part of the evidence; nevertheless, this incomplete evidence — if incomplete it be — will suffice for all the conclusions drawn.

Our interest then lies in the question of how close-relationships can be discovered by using quantitative measures of association between languages based on the whole inventory. The evolution of a family of languages is undoubtedly a complex process. As indicated above, we have adopted the junggrammatisch model; that is, we suppose that relationship between a pair of languages implies the existence at some time past of a common ancestor or parent language. The problem is then to construct the family tree of all languages of the family and their lines of descent from their parent languages.

In Glottochronology, a further aim is to attach a time-scale to the lines of descent from the parent language. The construction of a theory to achieve this involves some strict, and generally implausible assumptions as to constant rates of disappearance of words from languages over time. If we consider a language from time $\tau = t$ to $\tau = T$, then it is certainly true to say that the ith of its words existing at $\tau = t$ has a chance $p_{i\tau}$ of dying out in the interval τ, $\tau + \Delta\tau$. The basic attitude of the glottochronologists may be expressed by supposing them to say that all, or, at worst, whole groups of these probabilities are the same. The basic attitude of their opponents is,

essentially, to deny this; some of them would go so far as to say that $p_{i\tau}$ has a different value for every i, τ, or at the least for every i. In other words, that the chance of a word dying out at a particular time has something to do with the word and the time themselves, or at the least with the word itself. It may also be observed that, in the majority of cases in which Glottochronology has been used in practice, there has been a total absence of historical evidence whereby its results can be checked. So we may conclude this section by emphasising that no attempt will here be made to consider time as a factor in relation to family trees of languages.

We now make some general observations on measures of association between related languages and their use in constructing family trees.

Data and notation

It is supposed here that the available data consists of an inventory of N words by means of which it is appropriate to investigate the associations of a group of l languages. Suppose the languages are labelled L_1, L_2 . . . L_l. Most of the work on association measures between languages is based on computing certain quantities. For any pair of languages L_i and L_j, the total numbers of inventory words present in L_i and L_j, n_i and n_j respectively, are calculated. The number of cases in which a word is present in both L_i and L_j, r_{ij}, is also obtained. With the aid of a computer, it would be possible to proceed further and calculate the number of words common to groups of three or more of the languages. Though these quantities would undoubtedly contain information about the association between languages of the family, it will be seen that the difficulty lies in knowing how to exploit or indeed extract this information, particularly when the number of languages is large. Our work will be confined to the use of the r_{ij} only.

Measures of association between language pairs

It has been argued by a number of authors (for example, Ross) that a measure of association should depend only on

the number of words in common, r_{ij}, and not on the number of words absent in both languages; this can be seen to be $N - n_i - n_j + r_{ij}$. Some examples of measures of this type are the following.

(I) A. Ellegård, »Statistical measurement of linguistic relationship», Language XXXV, 131—56, proposed c_{ij}, the ratio of r_{ij} to the geometric mean of the »richnesses» of the languages, n_i and n_j, as an association coefficient;

$$c_{ij} = r_{ij} / (n_i n_j)^{1/2}.$$

(II) A. Henrici, »Numerical classification of Bantu languages», African language studies XIV, 82—104, used c'_{ij}, the ratio of r_{ij} to the arithmetic mean of the richnesses;

$$c'_{ij} = r_{ij} / \tfrac{1}{2}(n_i + n_j).$$

(III) A similar measure to Henrici's is s — which will be here referred to as the "similarity coefficient" — the proportion of words common to both languages relative to the proportion of words in either language;

$$s_{ij} = r_{ij} / (n_i + n_j - r_{ij}).$$

Neither the inventory size, N, nor the numbers of words absent in pairs of languages enters explicitly into these measures. Further, since all the measures are proportional to r_{ij}, it is to be expected that, whichever of the three measures is used, comparisons between languages should give similar results. Henrici reports that this is true in practice.

Various authors have made attempts to find ways of studying language associations more sophisticated than looking at simple measures of association. We now briefly describe some of these methods.

Statistical measures of association

If a pair of languages had evolved from different parent languages without interborrowing of words, it is nevertheless conceivable that the same word should occur in both languages

purely by chance. A probability model for such a chance mechanism is the following — the argument is conditional upon the richnesses n_i and n_j of languages L_i and L_j being treated as fixed. Suppose that the n_j words of L_j are randomly sampled from the N inventory words, and that the count, r_{ij}, of those words also present in L_i has been made. Then it is well known (cf. for example Ross p. 27) that r_{ij} has a hypergeometric distribution with

$$\text{Mean} = E(r_{ij}) = n_i n_j / N;$$

$$\text{and Variance} = \text{Var}\,(r_{ij}) = \frac{n_i n_j}{N-1}\,(1 - n_i / N)(1 - n_j / N).$$

Further, if n_i and n_j are large, as they usually are for language inventories, the standardised statistic

$$z_{ij} = \frac{r_{ij} - n_i n_j / N}{\frac{n_i n_j}{N-1}\,[(1 - n_i / N)(1 - n_j / N)]^{1/2}}$$

has approximately the Normal distribution with zero mean and unit variance.

We may then apply a significance test to reject the hypothesis of no association. This hypothesis would be rejected at probability level p, where p is the probability of obtaining a value of z_{ij} greater than or equal to the observed value. That is

$$p = \int_{z_{ij}}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-1/2 t^2}\, dt.$$

As there is a strict monotonic decreasing relationship between z_{ij} and p, either the positive size of z_{ij} or the smallness of p, is a measure of association between the languages. Usually it is more convenient to work with z_{ij}, but, for future reference, we give here a short table of the correspondence-values of p and z_{ij}.

At this point it is to be noted that this association measure does depend implicitly on the number of words absent in both languages. As it depends on all four of r_{ij}, n_i, n_j and N, it is

TABLE III

z_{ij}	p
2.00	0.022750100
3.00	0.001349900
4.00	0.000031700
5.00	0.000000287
6.00	0.000000001

possible to write it in terms of the number of words absent in both, $N - n_i - n_j + r_{ij}$. However, from the point of view of statistical significance, it is surely right to consider the size of the inventory N, and so this fact cannot be held a serious disadvantage.

Constant survival rate model for separation of languages

There are some points of connection with another measure of relationship proposed by D. G. Kendall (Ross pp. 41—42). Suppose that, at some point of time, languages L_i and L_j had a common parent consisting of S_{ij} words. Suppose a_i and a_j are the probabilities of survival for any word along the lines of descent to L_i and L_j respectively. It has to be assumed that the probability is the same for all words, which is almost certainly not true (cf. p. 26 above), but, if a_i and a_j are interpreted as average probabilities, the model may be reasonably adequate. It is also assumed that the extinction of any one word is independent of that of any other. Then it can be shown that appropriate estimates of S_{ij}, a_i and a_j are

$$S_{ij} = \frac{n_i n_j}{r_{ij}}, \quad a_i = \frac{r_{ij}}{n_j}, \quad a_j = \frac{r_{ij}}{n_i}.$$

Such estimates would be obtained, for instance, by applying the well-known statistical principle of the method of moments to the problem.

Clearly the quantity S_{ij} will be large when r_{ij} is small, a state of affairs which would perhaps imply an early separation of the languages.

We see, too, that S_{ij} can be related to the numerator of the z_{ij}-statistic, $r_{ij} - n_i n_j / N$, for we can write

$$r_{ij} - n_i n_j / N = n_i n_j \left(\frac{1}{S_{ij}} - \frac{1}{N} \right).$$

As the denominator of the z_{ij}-statistic is reasonably stable for different pairs of languages L_i, L_j with reasonably similar richnesses, comparisons between languages based on the z_{ij} are likely to produce similar conclusions to those based on the S_{ij}.

Anscombe's Method

So far only pairwise comparison of languages has been considered. Anscombe proposed a method of analysis (Ross, pp. 51—53) which not only enables several languages to be considered simultaneously but is also supposed to take account of the interborrowing of words. The following quotation from Anscombe explains the rationale of the method. »— — the number of positive agreements — — between two languages is not an entirely satisfactory measure of relationship by itself, since two languages not of recent common ancestry but which have borrowed extensively from other languages may have a large number of characters in common, while two small closely-related languages may have fewer characters in common. The richness of the languages should somehow be taken into account. The test devised below is for the null hypothesis that all the languages of the set are equally related, in such a way that any one is just as likely to have a particular — — character — — as any other.»

Anscombe's analysis is based on considering the distribution of quantities u_{ij} where

$$u_{ij} = r_{ij} - \frac{T_i + T_j}{l - 2} + \frac{2T}{(l-1)(l-2)},$$

where l = number of languages considered;

T_i = sum of all r_{ij} having one suffix equal to i;

T = sum of all the r_{ij}, $i < j$.

Thus the u_{ij} are the r_{ij} corrected for the richnesses of the languages L_i and L_j. So a large u_{ij} is supposed to imply high association. It is possible to test whether there is association among the family of l languages as a whole. For Anscombe shows that, if there is equal association, a statistic proportional to the sum of the squares of all the u_{ij} has approximately a chi-squared distribution with $^1/_2\ l\,(l-3)$ degrees of freedom. Large values reject the hypothesis of equal association. Individual u_{ij} can be examined to see which languages, if any, are associated.

There are objections to Anscombe's method. First, there is no reason to suppose that richness of the r_{ij} should be corrected linearly by subtraction, as is done in obtaining the u_{ij}. However, it is also true to say that there is no particularly convincing reason for arguing that the r_{ij} are to be corrected multiplicatively as in Ellegård's measure $r_{ij} / (n_i n_j)^{1/2}$. Secondly, there is certainly some danger in correcting for richness on the supposition that it corrects for interborrowing. For suppose that two languages, L_i and L_j, are not subject to interborrowing but have very different rates of decay of words. Then, starting from a common parent, they may well have attained different richnesses at a later period. In these circumstances there seems little to be said for reducing r_{ij} as a measure of association because one language has a very slow rate of decay. The quantity r_{ij} — or r_{ij} / N — would seem to be as good a measure as any in the actual state of affairs, unless indeed the rates of decay themselves can be estimated. Unfortunately information for distinguishing between interborrowing and exceptional rate of decay is hardly ever available. If it can be assumed that there is no interborrowing, then Kendall's method can be used to estimate the rates of decay.

Graphical display methods and the construction of family trees

Suppose the measures of similarity between pairs of the languages L_1, $L_2 \ldots L_l$ are given. It is required to construct a family tree linking similar languages and indicating the

parent languages. If the number of languages is small, this can be done by inspection, though, of course, the judgment as to what size of similarity merits the linking of languages is almost entirely a subjective matter. If we invoke the concept of statistical significance in respect of the similarities, then it could well be suggested that languages not associated by a significant similarity should on no account be linked. In the case that the majority of the languages of a family are found to be mutually associated significantly, the problem of comparing the relative strength of significant similarities becomes dominant. If the number of languages is large, it is convenient to have an algorithm to do the linking automatically.

(I) Single-link cluster analysis[1]

Here the two languages with the largest similarity are linked first, then the two with the next largest similarity and so on, until a point is reached at which the similarities are so low that it is felt no longer desirable to join further languages to groups of languages already linked.

Group-average cluster analysis

A variant of this method consists in only joining a language to a group already linked if its »average» similarity with members of the group is the largest such average similarity out of those outside the group. »Average» can be defined in a number of ways, and a number of different methods of this type exist.

It should be emphasised that there is no such thing as a »best» clustering method, for different methods have their advantages and disadvantages depending on the structure of the data. Judgment of the results of tree construction is again a subjective matter.

[1] Cf. R. M. Cormack, »A review of classification», Journal of the Royal Statistical Society A. CXXXIV, 321—353.

(II) Mapping methods

A similarity between languages may be in part a geographical matter; this would particularly be the case if the presence of words in the languages were due to interborrowing. And, quite apart from interborrowing, the evolution of languages could imply geographical association.

It may therefore be suggested that the similarities between languages could be used to construct coordinates for the languages in a space of a suitable number of dimensions — in the geographical case this number would be two — such that languages with high similarity would have coordinates close together in the space whereas dissimilar languages would be far apart.

In realisation, such methods are known as scaling methods, and two are in common use.

Non-metric multidimensional scaling

Given the l languages L_1, $L_2 \ldots L_l$, with coefficients of similarity S_{ij}, the data is represented in a coordinate space of m dimensions by finding coordinates $\{x_{i\mathrm{K}};\ i = 1, 2 \ldots l;\ \mathrm{K} = 1, 2 \ldots m\}$ for each L_i in a space with metric d. m is to be as small as possible but, as far as possible, to satisfy the monotonicity condition

$$d_{ij} < d_{kq} \text{ when } S_{ij} > S_{kq}.$$

That is, L_i and L_j should be nearer together than are L_k and L_q if L_i and L_j are more similar than L_k and L_q. The choice of metric, d, is at our disposal, but is often taken to be the Euclidean distance. Various rules can be used to deal with "ties" ($S_{ij} = S_{kq}$). Coefficients of stress can be defined to measure how well the monotonicity condition is satisfied in a given number of dimensions. The method is said to be non-metric, because it depends directly only on the rank-ordering of the S_{ij} (by reason of the condition $S_{ij} > S_{kq}$) and only indirectly on their absolute magnitudes. Further details of

the method will be found in J. B. Kruskal, »Multidimensional scaling by optimising goodness of fit to a non-metric hypothesis», Psychometrika XXIX, 1—27.

Principal coordinates analysis

A similar method which uses the absolute magnitudes of the S_{ij} more directly is that known by the above title. The theory underlying the method is essentially the following.

It is first necessary to define the similarity of a language with itself. This is not a problem for any of our similarity measures. For instance, the maximum value of the z_{ij}-statistic is $\sqrt{N-1}$, and this is attained when the statistic is computed for the similarity of a language with itself. Next, a metric between the languages is defined by

$$d^2_{ij} = S_{ii} + S_{jj} - 2S_{ij},$$

where d_{ij} is the distance between L_i and L_j. Now the S_{ij} are corrected for mean similarities by forming the matrix α of elements

$$\alpha_{ij} = S_{ij} - \bar{S}_i + \bar{S}_j + \bar{S},$$

where $\bar{S}_i$ is the row mean of the $l \times l$ S_{ij} matrix,
$\bar{S}_j$ is the column mean of the S_{ij} matrix,
$\bar{S}$ is the average of all elements S_{ij}.

The latent roots $\lambda_1, \lambda_2 \ldots \lambda_l$ and vectors of the matrix α are computed. The latent vectors are normalised so that the sums of the squares of the elements are equal to the corresponding latent roots. Finally suppose that the latent roots $\lambda_1, \lambda_2 \ldots \lambda_l$ are ordered so that $\lambda_1 > \lambda_2 > \ldots > \lambda_l$.

Then, in l-dimensional space, languages L_i, L_j have coordinates given by the i, respectively j, elements of the l latent vectors. The Euclidean distance between points with these coordinates is d_{ij}, so the representation of the distances between the languages by these points is a perfect one in l dimensions. If we wish, we can seek to represent the languages in fewer than l dimensions (say, u) by using just the first u

coordinates. Suppose the distance between L_i and L_j using just the first u coordinates is $d_{ij}^{(u)}$. Then we can assess the adequacy of using just u dimensions to represent the data by the stress coefficient

$$\sum_{i<j} \left\{ [d_{ij}]^2 - [d_{ij}^{(u)}]^2 \right\} \Big/ \sum_{i<j} [d_{ij}]^2$$

which can be computed as

$$\sum_{i=1}^{u} \lambda_i \Big/ \sum_{i=1}^{l} \lambda_i.$$

Further details of this method can be found in J. C. Gower, »Some distance properties of latent root and vector methods in multivariate analysis», Biometrika LIII, 325—338.

Again it is not possible to say which of the mapping methods will give the most meaningful results for a particular set of data. On the whole, the two methods gave similar results on the linguistic data on which they have been tried.

THE URALIAN LANGUAGES

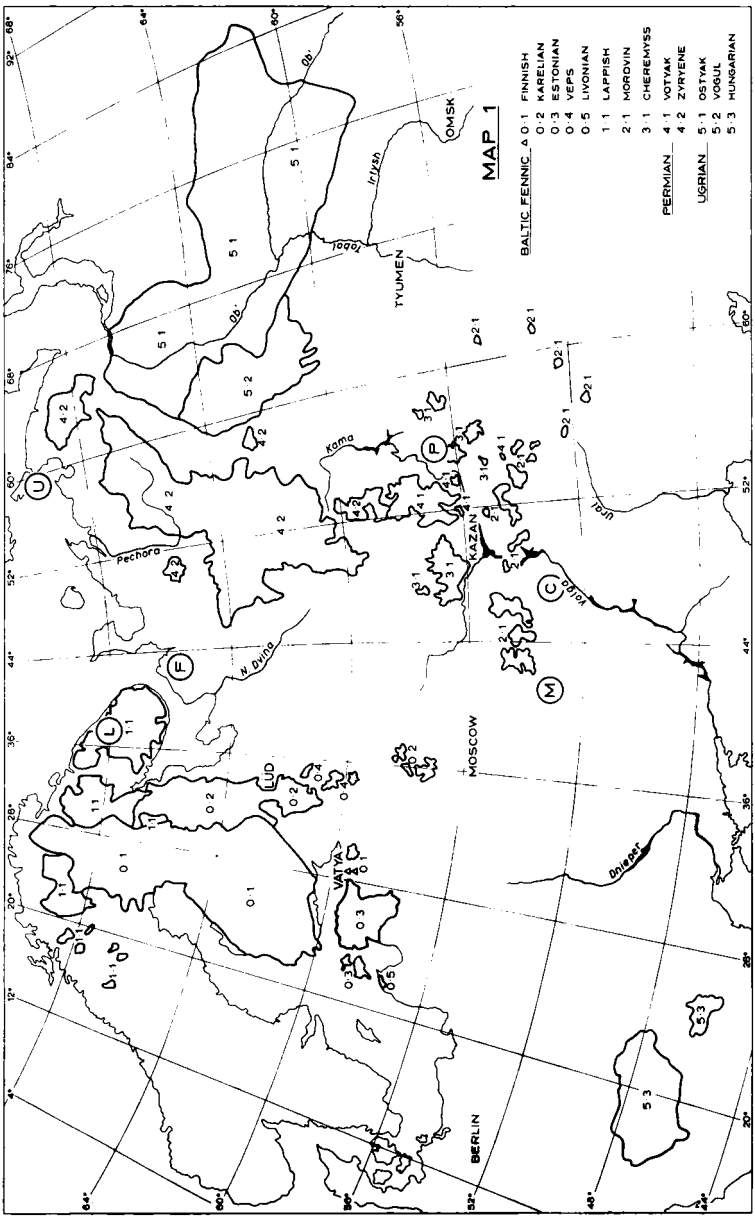
In this section many of the techniques just discussed are applied to the Uralian languages. Of some techniques, the results are not reported here because they have given very similar results to those of others.

We have — naturally — taken the major division of the family, into Finno-Ugrian and Samoyede, as given. As far as the inventory of words is concerned, there is very little overlap between the two subdivisions.

FINNO-UGRIAN

For this subdivision six languages or language-groups[1] were considered: Finnish (F), as representing Baltic Fennic; Lappish (L); Mordvin (M); Cheremyss (C); Permian (P), that is Zyryene and Votyak; and Ugrian (U), that is, Vogul, Ostyak

[1] In the context of these six — and only in this context — *Language* means 'language' *or* 'language-group'.


MAP 1
BALTIC FENNIC
0:1 FINNISH
0:2 KARELIAN
0:3 ESTONIAN
0:4 VEPS
0:5 LIVONIAN
1:1 LAPPISH
2:1 MOROVIN
3:1 CHEREMYSS
PERMIAN
4:1 VOTYAK
4:2 ZYRYENE
UGRIAN
5:1 OSTYAK
5:2 VOGUL
5:3 HUNGARIAN
OMSK
TYUMEN
KAZAN
MOSCOW
BERLIN
Pechora
N Dwina
Kama
Volga
Ural
Irtysh
Ob
Dniepr

and Hungarian. It will be observed that we have not taken the obvious opportunity of joining up Mordvin and Cheremyss into one language-group -- they are undoubtedly closely-related, forming, as they do, the Volga Fennic group. We did this of set purpose, so as to leave a control: would our data and methods in fact reveal that these two languages are closely-related?

The inventory for this subdivision consisted of $N = 965$ words. Table IV gives (1) the richnesses, n_i (last column), (2) the number of words in common, r_{ij} (above the diagonal), (3) the similarity coefficients $S_{ij} = r_{ij} / (n_i + n_j - r_{ij})$ (below the diagonal). From the Table it is apparent that the similarity coefficients are all much the same, with most language-pairs having about 35 % of the words in either in common. This would seem to suggest that the Finno-Ugrian languages constitute a family without very much close-relationship.

Map 1 shows the regions in which the various language-groups are located. A Principal Coordinates analysis of the r_{ij} and n_i was carried out to show the relative locations of the languages implied by these numbers. The first two coordinates were then plotted on the geographical map after suitable adjustment of the scale, the origin and the orientation of the principal coordinate axes. The stress-coefficient for a two-dimensional representation is high (48 %) and a three-dimensional system with stress 29 % is better. However, the two-dimensional principal coordinates are very similar to those produced by the non-metric scaling method, which has a stress of only 10 %. Thus the two-dimensional coordinates do provide a reasonable representation of the similarities implied by the r_{ij}. The principal coordinate points are shown on the map by the ringed language-group letters. If we leave Ugrian out of account, the closeness of these points to the corresponding geographical regions is striking. As Ugrian is composed of two regions far apart, the geographical centre has no meaning, and so a correspondence cannot be expected. We conclude that there is quite a strong relationship between language similarity and geographical proximity.

In Table V, the values of the z_{ij}-statistic are given above the diagonal and the u_{ij} obtained from Anscombe's analysis below it. The result of Anscombe's test was that the hypothesis

TABLE IV

	F	L	M	C	P	U	
F		263	196	201	243	412	479
L	0.40		165	176	213	352	451
M	0.33	0.28		170	161	249	298
C	0.32	0.28	0.36		197	295	341
P	0.35	0.31	0.27	0.33		387	446
U	0.48	0.40	0.30	0.35	0.36		786

that the languages are equally related, in the sense that any one is just as likely to have a particular word as any other one, is rejected at a very low probability level of 1 in 10 000. Of more interest are the u_{ij} in their rôle as measures of association corrected for the richnesses of the languages. The languages most strongly associated are Mordvin and Cheremyss — our control is thereby vindicated — with Permian and Ugrian also strongly related. It was argued that a value of z_{ij} much greater than 3 would provide strong statistically significant evidence of association on the grounds that, for the given richnesses of the two languages, there would be more words in common than would be expected by chance. Many of the language-pairs are covered by this rubric, with Mordvin and Cheremyss again strongly linked. But Lappish is not significantly related to either Permian or Ugrian by this argument. Principal coordinate maps of the z_{ij} and u_{ij} can be constructed in the same way as was the r_{ij} map of Fig. 1. As they are similar to this, and confirm the impression that strength of association is broadly related to geographical proximity, they are not reproduced here.

TABLE V

	F	L	M	C	P	U
F		5.1	6.7	4.3	2.8	3.6
L	10.0		3.6	2.2	0.6	−2.5
M	0.0	5.5		9.4	3.2	1.1
C	−19.5	−8.0	43.0		5.3	3.0
P	−18.0	−11.5	−6.5	5.0		3.9
U	27.5	4.0	−42.0	−20.5	31.0	

The Finno-Ugrian family tree and its parameters

According to Kendall's model for the separation of two languages, L_i and L_j, the parameters of the model are, first, a_i, the probability of survival of any root in L_i, along the line of descent to L_i and, secondly, S_{ij}, the number of words in the common parent of L_i and L_j.

The estimates of S_{ij} for each pair of Finno-Ugrian languages are given in Table VI. Only one of the values of S_{ij} is greater than the inventory size $N = 965$, but many of the values are close to it. If Mordvin and Cheremyss are excluded it would not be unreasonable to postulate a parent language containing the 965 words from which the other languages are directly descended. (Of course it is profitless to speculate as to the number of words of the parent language which have disappeared in a l l its descendants). The smaller figures for Mordvin and Cheremyss suggest that they had an intermediate parent from which several hundred of the original words had already disappeared.

Each pair of languages L_i and L_j provides estimates of a_i and a_j along their lines of descent, and these are given in Table VII. Inspection of the rows of this table shows a high degree of stability in the estimates of the a, with probabilities of survival for Finnish, Lappish and Permian of between 0.5 and 0.6. But Mordvin and Cheremyss retain about 55 % of the words along their common line of descent but only 30 % to 40 % along their lines of descent with other languages. Thus one family tree suggested by these figures is that given in Fig. 1, which involves two parent languages I and II and

TABLE VI

	F	L	M	C	P	U
F		821	728	813	879	914
L			814	874	944	1 007
M				598	825	941
C					772	909
P						906
U						

TABLE VII

Language used with L_i to derive a_i

Language L_i	F	L	M	C	P	U
F		0.58	0.66	0.59	0.54	0.52
L	0.55		0.55	0.52	0.48	0.45
M	0.41	0.37		0.50	0.36	0.32
C	0.42	0.39	0.57		0.44	0.37
P	0.51	0.47	0.54	0.58		
U	0.86	0.78	0.83	0.86	0.87	

seven probabilities or rates of survival labelled in an obvious notation a_{L}, a_{F}, a_{P}, a_{U}, a_{II}, a_{M}, a_{C}.

The statistical problem of how best to estimate the parameters from the data of the r_{ij} and the n_i is difficult. Some of the standard statistical techniques, such as maximum likelihood estimation, are intractable, because of the difficulty of writing down the marginal likelihood function for the data. The approximative technique employed can be briefly described as follows. It was assumed that Parent I contained all 965 of the inventory words. Then, for instance, the expected value of r for Lappish and Finnish is $965a_{\mathrm{L}}a_{\mathrm{F}}$. By taking logarithms, the expected value of $\log(r/965)$ is approximately $\Theta_{\mathrm{L}} + \Theta_{\mathrm{F}}$, where Θ_{L} and Θ_{F} are $\log a_{\mathrm{L}}$ and $\log a_{\mathrm{F}}$ respectively. Thus the problem was changed to one involving a model linear in the Θ parameters; least squares techniques were used to estimate the Θ, and hence the a. Further details of this approximative method of analysis will not be considered here, as the only purpose has been to show that there are values of the parameters which make the model fit the data quite well. In Table VIII the estimated parameter values and their approximate standard errors are given. These can be used to obtain the expected or predicted values of the model (e.g. $965a_{\mathrm{L}}a_{\mathrm{F}}$ for r_{LF}) and approximate standard errors for these. The estimates of the a are much as would be expected from Table VII, and the predicted values give quite a good approximation to the observed data. In fact the model fits the data better than could reasonably be expected when it is remembered that it assumes that all words have the same probability of survival and decay.

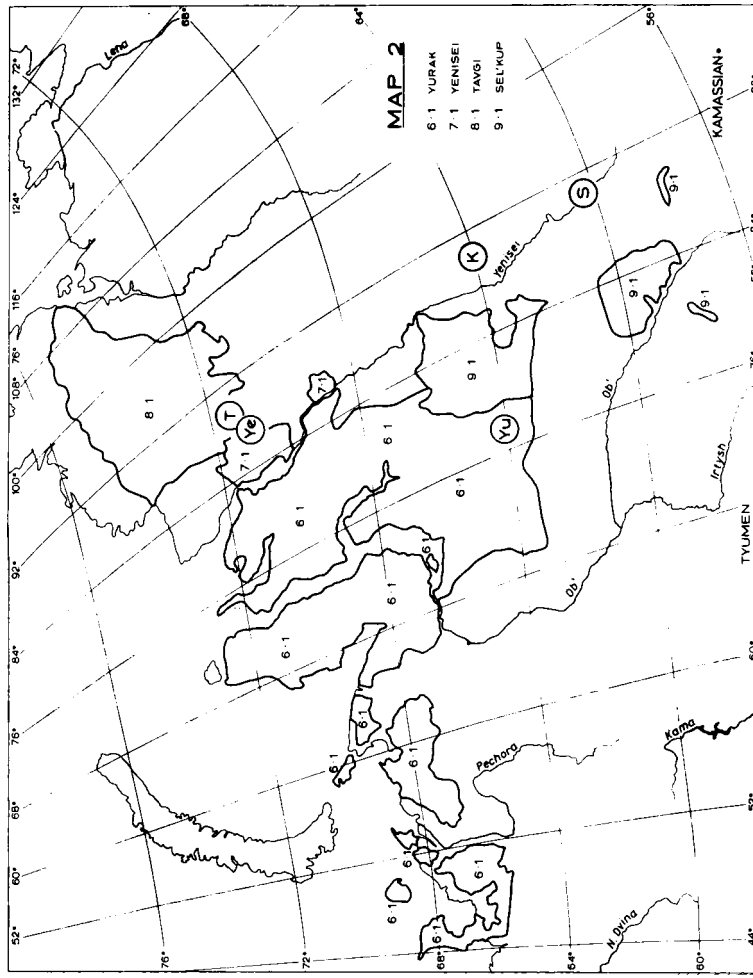

MAP 2
6·1 YUKAK
7·1 YENISEI
8·1 TAVGI
9·1 SEL'KUP
Lena
Yenisei
Pechora
Kama
N. Dvina
Irtysh
TYUMEN
KAMASSIAN
K
S
T
Ye
Yu

Figure 1.

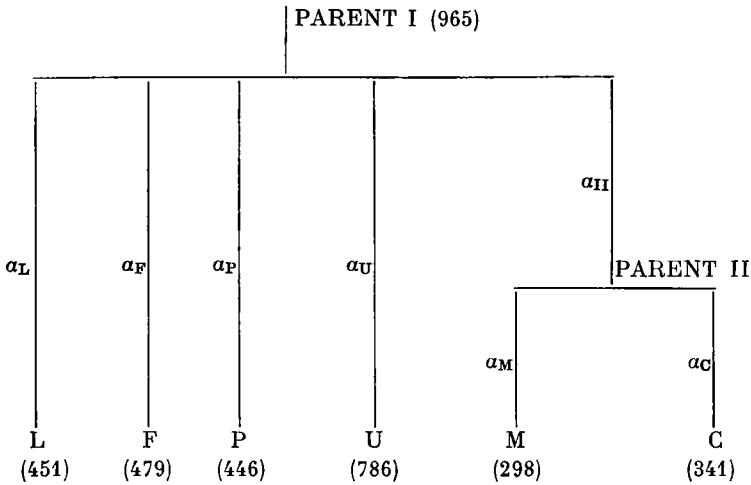


TABLE VIII

Estimates of the Finno-Ugrian a and predicted values for the n_i and r_{ij}

a_L	a_F	a_P	a_U	a_{II}	a_M	a_C
0.48	0.54	0.49	0.79	0.75	0.46	0.51
(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.04)	(0.06)

	Observed	Predicted			Observed	Predicted	
n_L	451	460	(13)	r_{FP}	243	254	(10)
n_F	479	524	(15)	r_{FU}	412	414	(16)
n_P	446	468	(13)	r_{FM}	196	179	(7)
n_U	786	762	(21)	r_{FC}	201	202	(8)
n_M	298	330	(10)	r_{PU}	387	370	(14)
n_C	341	371	(11)	r_{PM}	161	160	(6)
r_{LF}	263	250	(9)	r_{PC}	197	180	(7)
r_{LP}	213	223	(8)	r_{UM}	249	260	(10)
r_{LU}	352	362	(14)	r_{UC}	295	293	(11)
r_{LM}	165	157	(6)	r_{MC}	170	170	(11)
r_{LC}	176	177	(7)				

The figures in brackets are approximate standard errors for estimates of parameters or predicted values.

SAMOYEDE

For this subdivision five languages were considered: — Yurak (Yu); Tavgi (T); Yenisei (Ye); Sel'kup (S); Kamassian (K), as representing Sayan-Samoyede. The inventory for this subdivision consisted of $N = 468$ words. Table IX is the counterpart of Table IV. Apart from implying a strong connection between Yenisei and Tavgi, the similarity coefficients are not very helpful, as they show a uniform picture of relationship with about 40 % of the words in either of a language pair in common.

Map 2 [1] is the counterpart of Map 1. The principal coordinates analysis of the r_{ij} in two dimensions gave a stress coefficient of about 30 % compared with the non-metric scaling stress of about 2 % but the language coordinates are similar in the two cases. As before, the points were plotted on the map and the resemblance of the map of language similarity to the geographical one is again quite striking. Table X is the counterpart of Table V. The hypothesis that the languages are equally related is rejected at a probability level of less than 1 in 10 000. Inspection of the table shows that the closest associations are between Yenisei and Tavgi, and then, between Sel'kup and Kamassian. The only z that are not statistically significant are those between Yurak and Sel'kup or Kamassian. As in the case of Finno-Ugrian, a Principal Coordinate map of the z shows a strong similarity to the actual geography.

TABLE IX

	Yu	T	Ye	S	K	
Yu		154	171	178	165	342
T	0.42		149	126	118	180
Ye	0.47	0.65		135	125	196
S	0.42	0.40	0.42		159	264
K	0.42	0.43	0.44	0.49		215

[1] Having in mind what A. Joki, Kai Donners Kamassisches Wörterbuch (Lexica Societatis Fenno-ugricae VIII), p. XVIII, and, again, Die Lehnwörter des Sajansamojedischen (MSFOu 103), p. 27, has to say, we have entered Kamassian at Abalakova (55°04′ N. × 94°50′ E.).

The Samoyede family tree and its parameters

Tables XI and XII are the counterparts of Tables VI and VII respectively. Only one of the values of S_{ij} is greater than $N = 468$.

The following points are of help in the devising of a family tree that might represent the data adequately. The Tavgi and Yenisei figure for S_{ij} of 237 is the smallest, and, in view of their strong association, it is reasonable to postulate a parent from which about half the original inventory has been lost. This is supported by the a_i figures which are very similar for the Tavgi and Yenisei rows and columns which show their similarity to other languages. The same kind of argument will justify a common parent for Sel'kup and Kamassian containing about 350 of the inventory words. Yurak, with its high proportion of words surviving, seems quite distinct from the other four languages. There arises the question whether the parent of Sel'kup and Kamassian could be the same as that of Tavgi and Yenisei. The difference in the estimated numbers of words in the parents suggests that this is not the case, and, so, trying to fit a family tree with the same immediate parent proves a failure. On the other hand, the entries in Tables XI and XII suggest that the parent of Sel'kup and Kamassian and that of Yurak and Tavgi must have been fairly closely related. After a certain amount of trial and error with different family trees having the properties described above, we concluded that the tree in Fig 2 was the one that fitted the data best. It involves three parent languages I, II and III, and seven rates of survival labelled in an obvious notation a_{Yu}, a_{II}, a_S, a_K, a_{III}, a_T, a_{Ye}. These parameters were estimated

TABLE X

	Yu	T	Ye	S	K
Yu		4.8	5.9	−3.1	1.6
T	−4.3		14.2	4.7	6.7
Ye	1.7	20.0		4.6	6.6
S	2.7	−9.0	−11.0		7.0
K	0.0	−6.6	−10.6	17.4	

from the data in the same approximative way that was used for Finno-Ugrian, and the results are given in Table XIII, the counterpart of Table VIII. It is evident that the model produces predicted values which fit the observed data well, and all observed values are within one or two standard errors of the predicted values.

CONCLUSIONS

A number of different measures of association and methods of analysis have been tried with the Uralian data. It is not possible to say which method is — in any sense — the best, since they have different objectives, contribute different information, and rely on different assumptions. The construction of the family trees is undoubtedly the most ambitious part of the analysis. It is also the part which relies most heavily on simplifying assumptions and evolutionary models. Though the results apparently fit the data quite well, it would be foolish not to recognise that it is in this part of the analysis that we may be most seriously wrong. Mapping methods in respect of association measures seem particularly useful for this kind of data, for they show the extent to which language similarities are related to the geographical proximity of the languages. For the Uralian languages, this extent was substantial.

TABLE XI

	Yu	T	Ye	S	K
Yu		400	392	507	446
T			237	377	328
Ye				383	337
S					357

TABLE XII

	Yu	T	Ye	S	K
Yu		0.85	0.87	0.67	0.77
T	0.45		0.76	0.48	0.55
Ye	0.50	0.83		0.51	0.58
S	0.52	0.70	0.69		0.60
K	0.48	0.65	0.64	0.73	

Figure 2.

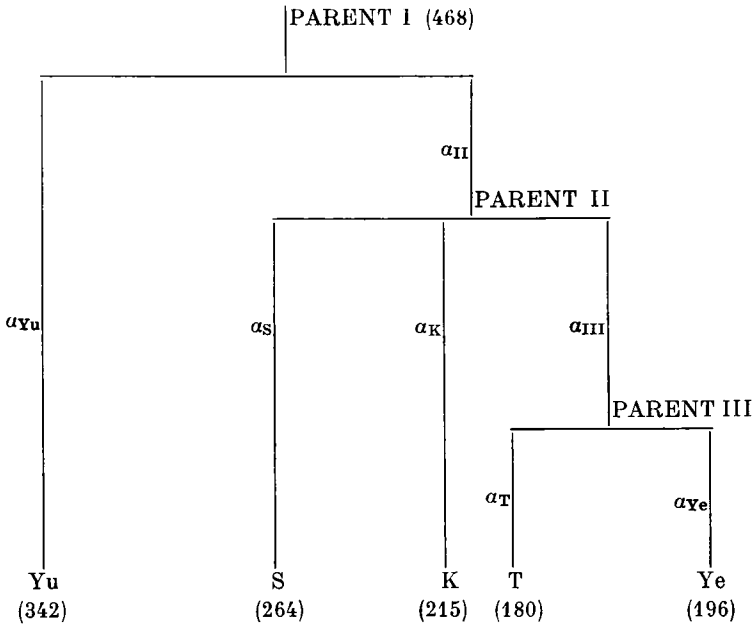


TABLE XIII

Estimates of the Samoyede α and predicted values for the n_i and r_{ij}

α_{Yu}	α_{II}	α_S	α_K	α_{III}	α_T	α_{Ye}
0.77	0.77	0.68	0.61	0.72	0.73	0.79
(0.03)	(0.05)	(0.03)	(0.03)	(0.06)	(0.05)	(0.06)

	Observed	Predicted			Observed	Predicted	
n_{Yu}	342	360	(13)	r_{YuYe}	171	157	(7)
n_S	264	246	(10)	r_{SK}	159	151	(6)
n_K	215	221	(9)	r_{ST}	126	129	(6)
n_T	180	189	(8)	r_{SYe}	135	140	(6)
n_{Ye}	196	204	(8)	r_{KT}	118	116	(5)
r_{YuS}	178	189	(8)	r_{KYe}	125	126	(5)
r_{YuK}	165	170	(7)	r_{TYe}	149	149	(9)
r_{YuT}	154	145	(6)				

The figures in brackets are approximate standard errors for estimates of parameters or predicted values.

The question of how the information from these two methods of mapping and family tree construction should be combined is a very difficult one. The family tree model is one which gives an implied temporal ordering to the concept of one language changing to another even if no time scale is attached to the branches. Mapping methods have been used here with some success to give information about how language changes have taken place spatially but we have not been able to infer any temporal ordering from these results. To combine the methods, a model for language change incorporating both geographical and temporal ideas is needed but we lack both such a model and the historical evidence by which it could be tested.

PAUL DAVIES and ALAN S. C. ROSS