The methods described here are a means of text classification with reference to a certain number of features which can be measured by real numbers. We have applied them to find manuscript families and subfamilies in the different copies of a certain work, but the same methods can be used to classify any finite set of objects with reference to some features, the realizations of which are known for each of the objects; e.g. the objects might be all preserved ancient greek writers and the features 1) relative frequency of the article, 2) of nouns, 3) of *ἅπαξ λεγόμενα* and 4) the quotient of average sentence length and longest sentence.

But now to the stemma problem. Let the objects be N copies X_1, X_2, ..., X_N of the same work and let the features be the similarities S_{X_1}, S_{X_2}, ..., S_{X_N} to each of these copies, where $S_{X_j}(X_i)$ is the similarity of X_i to X_j for $1 \leqslant i, j \leqslant N$. Writing $S_{X_j}(X_i) =: s_{ij}$, we can arrange the mutual similarities of all copies in a quadratic matrix $(s_{ij})_{N \times N}$ such that in the i^{th} row and the j^{th} column we have the similarity of X_i to X_j :

copy \ similarity	S_{X_1}	S_{X_2}	...	S_{X_N}
X_1	s_{11}	s_{12}	...	s_{1N}
X_2	s_{21}	s_{22}	...	s_{2N}
.	.	.		.
.	.	.		.
.	.	.		.
X_N	s_{N1}	s_{N2}	...	s_{NN}

Since all of the later explained methods are based on this matrix of similarities, we have to begin with the definition of similarity measures.

SIMILARITY OF TEXTS

Regarding a text X_i as a sequence of n terms $x_{i1}, x_{i2}, \ldots, x_{in}$ at the second place indexed by the numbers of the corresponding terms of the collation-text (1), we count the *variae lectiones* of X_i and X_j by $|V(\{i, j\})|$, where $V(\{i, j\}) := \{k : x_{ik} \neq x_{jk}, x_{ik} \in X_i, x_{jk} \in X_j\}$, i.e. $V(\{i, j\})$ is the set of places where X_i differs from X_j, and for any subset $\{X_i : i \in I \subset \{1, \ldots, N\}\}$ of the compared texts $X_1, \ldots, X_N$ by $|V(I)|$, where $V(I) = \{k : \exists\, i, j \in I\ x_{ik} \neq x_{jk}\}$, i.e. $V(I)$ is the set of places $k \in \{1, \ldots, n\}$ such that at least two texts of the subset indexed by I have different readings at k. We can define now two similarity measures by

$$s_{ij} := \frac{|V(\{1, \ldots, N\})| - |V(\{i, j\})|}{|V(\{1, \ldots, N\})|}$$

and

$$s'_{ij} := \frac{|V(\{1, \ldots, N\})| - 2\,|V(\{i, j\})|}{|V(\{1, \ldots, N\})|}$$

Of course $|V(\{i, j\})| \leqslant |V(\{1, \ldots, N\})|$ and therefore $0 \leqslant s_{ij} \leqslant 1$ and $-1 \leqslant s'_{ij} \leqslant 1$. These definitions are practicable if we can assume that for any two texts X_i and X_j the average significance of a different reading is almost the same. Otherwise for each place $k \in V(\{i, j\})$ a weight w_k

corresponding to the significance of the difference of x_{ik} and x_{jk} has to be introduced, and for each $k \in V(\{1, ..., N\})$ a weight $w_k^* := \max\limits_{(i,j)} w_k$, which is the weight corresponding to the most significant of all text-differences occurring at k. Then we have the similarities

$$s_{ij} = \frac{\sum\limits_{k \in V(\{1, ..., N\})} w_k^* - \sum\limits_{k \in V(\{i, j\})} w_k}{\sum\limits_{k \in V(\{1, ..., N\})} w_k^*}$$

$$= 1 - \frac{\sum\limits_{k \in V(\{i, j\})} w_k}{\sum\limits_{k \in V(\{1, ..., N\})} w_k^*}$$

and analogously

$$s'_{ij} = 1 - \frac{2\sum\limits_{k \in V(\{i, j\})} w_k}{\sum\limits_{k \in V(\{1, ..., N\})} w_k^*}$$

A further modification of s_{ij} and s'_{ij} is convenient, if great parts of the text, e.g. more than half a page, have been omitted either in X_i or in X_j, because according to the above definitions, every term occurring in only one of X_i and X_j would be counted as a different reading and therefore X_i and X_j

would seem to be very different, even if the parts preserved in both X_i and X_j do not differ at all. It is convenient, therefore, to separate these large omissions from simple *lacunae* and to use a restriction of $V(\{i, j\})$, $V(\{1, ..., N\})$ and the corresponding sums of weights to those parts of X_i and X_j which are not covered by one of the large omissions, i.e. if O_{ij} is the set of places k covered by large omissions in X_i or X_j and if $P_{ij} := \{1, ..., n\} \setminus O_{ij}$, i.e. if P_{ij} is the set of places preserved in both X_i and X_j (including simple *lacunae*), we use $V(\{i, j\}) \cap P_{ij}$ and $V(\{1, ..., N\}) \cap P_{ij}$ for the definition of s_{ij} and s'_{ij}. Now the larger the omissions, the less informative are the resulting similarities. This fact can be taken account of by multiplying the former s'_{ij} by $\frac{|P_{ij}|}{n}$, i.e. we have now

$$s'_{ij} = \left(1 - \frac{2 \sum\limits_{k \in V(\{i,j\}) \cap P_{ij}} w_k}{\sum\limits_{k \in V(\{1, ..., N\}) \cap P_{ij}} w_k^*} \right) \frac{|P_{ij}|}{n}$$

and similarly

$$s_{ij} = \left(1 - \frac{\sum\limits_{k \in V(\{i,j\}) \cap P_{ij}} w_k}{\sum\limits_{k \in V(\{1, ..., N\}) \cap P_{ij}} w_k^*} - \frac{1}{2} \right) \frac{|P_{ij}|}{n} + \frac{1}{2}$$

with $-1 \leqslant s'_{ij} \leqslant 1$ and $s'_{ij} \to 0$ for $|P_{ij}| \to 0$
and $0 \leqslant s_{ij} \leqslant 1$ and $s_{ij} \to \frac{1}{2}$ for $|P_{ij}| \to 0$.

For our tests with the methods of automatic classification we have used the similarity

$$s'_{ij} = 1 - \frac{2\,|V(\{i, j\}) \cap P_{ij}|}{|V(\{1, \ldots, N\}) \cap P_{ij}|},$$

assuming that for any two texts X_i and X_j the means of the difference weights are almost the same (2). The reader may define other similarities appropriate to the individual conditions of text tradition. In the above definitions clearly $V(\{i, j\}) = V(\{j, i\})$ and $P_{ij} = P_{ji}$, hence $s_{ij} = s_{ji}$, i.e. the similarity matrix $(s_{ij})_{N \times N}$ is symmetrical.

EXPERIMENTS WITH CORRELATION MATRICES AND FACTOR ANALYSIS

Both methods use the correlation matrix $(r_{ij})_{N \times N}$ where r_{ij} is the Pearsonian correlation coefficient of $(s_{i1}, s_{i2}, \ldots, s_{iN})$ and $(s_{j1}, s_{j2}, \ldots, s_{jN})$, i.e.

$$r_{ij} = \frac{\sum_{k=1}^{N} (s_{ik} - \bar{s}_i)\,(s_{jk} - \bar{s}_j)}{\sqrt{\sum_{k=1}^{N} (s_{ik} - \bar{s}_i)^2 \cdot \sum_{k=1}^{N} (s_{jk} - \bar{s}_j)^2}},$$

where $\bar{s}_i = \frac{1}{N} \sum_{k=1}^{N} s_{ik}$ and $\bar{s}_j$ analogously.

a) Correlation

In a text-enchainment consisting of fairly long chains, the main branches parting from the centre of the enchainment can be separated, assuming that

1) $r_{ij} < O$ indicates that X_i and X_j are not in the same branch and

2) within a branch, texts more distant to the centre have lower correlation coefficients to any certain text of another branch than texts less distant to the centre.

If there are texts X_k and X_l such that $r_{ik} > O$ and $r_{jk} < O$ but $r_{il} < O$ and $r_{jl} > O$, then X_i and X_j are of different branches and hence for all texts of a branch, let us say $\{X_{j_1}, ..., X_{j_b}\}$, there is a chain of implications $r_{ij_1} > O \Rightarrow r_{ij_2} > O \Rightarrow \ldots \Rightarrow r_{ij_b} > O$ which is the same for all i. Now if a branch is long enough, the outermost text X_{j_1} has negative correlation coefficients to all texts of another branch. Then a chain of implications of the above form shows that all texts $X_{j1}, \ldots, X_{jb}$ are of the same branch (3).

Example

Similarity-matrix of 8 texts A, ..., H :

	A	B	C	D	E	F	G	H
A	1.000	.333	.556	.556	.333	.111	-.333	-.111
B	.333	1.000	.778	.333	.111	-.111	-.556	-.333
C	.556	.778	1.000	.556	.333	.111	-.333	-.111
D	.556	.333	.556	1.000	.778	.556	.111	.333
E	.333	.111	.333	.778	1.000	.778	.333	.556
F	.111	-.111	.111	.556	.778	1.000	.556	.778
G	-.333	-.556	-.333	.111	.333	.556	1.000	.778
H	-.111	-.333	-.111	.333	.556	.778	.778	1.000

Correlation-matrix derived from the similarities :

	A	B	C	D	E	F	G	H
A	1.000	.679	.788	.600	-.063	-.531	-.818	-.743
B	.679	1.000	.961	.306	-.381	-.775	-.962	-.922
C	.788	.961	1.000	.459	-.253	-.701	-.944	-.886
D	.600	.306	.459	1.000	.666	.170	-.297	-.148
E	-.063	-.381	-.253	.666	1.000	.812	.427	.572
F	-.531	-.775	-.701	.170	.812	1.000	.825	.919
G	-.818	-.962	-.944	-.297	.427	.825	1.000	.966
H	-.743	-.922	-.886	-.148	.572	.919	.966	1.000

For j = A, ..., H we have the implications

$$r_{Aj} > O \Rightarrow r_{Bj} > o \Rightarrow r_{Cj} > o \Rightarrow r_{Dj} > O,$$

i.e. ABCD are of the same branch, and

$$r_{Hj} > O \Rightarrow r_{Gj} > O \Rightarrow r_{Fj} > O \Rightarrow r_{Ej} > O,$$

i.e. EFGH are of the same branch,

furthermore $r_{DH} < O$ and $r_{EH} > O$,

but $r_{DA} > O$ and $r_{EA} < O$,

i.e. D and E are not in the same branch; hence ABCD and EFGH are of different branches and since there are no more texts than these, there is no further branch.

The same figured data, by the affiliation programmes of MAU and AHNERT (4), yield exactly the same results :

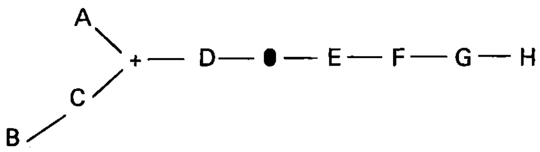
table of constellation-types :

1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	0	0	0	0	1	0
0	1	1	0	0	0	0	0
0	0	0	0	0	0	1	1
1	1	1	0	0	0	0	0
1	1	1	1	0	0	0	0
1	1	1	1	1	0	0	0
1	1	1	1	1	1	1	1

enchainment-output :

B...C...+...D...E........F
A....... +...D...E........F
G...H.........................F

i.e. the enchainment is



where + denotes a lost text and ● the centre of the enchainment. Obviously ABCD and EFGH are two branches parting from the centre.

b) Factor analysis

Factor analysis aims at reducing the complete set of variables, i.e. in our case the similarities $S_{X_1}, \ldots, S_{X_N}$, to a small number of hypothetical variables $F_1, \ldots, F_l$ (called common factors) such that any similarity S_{X_i} is for the greatest possible part composed of some of the common factors F_j and for the rest by a specific factor U_i which does not contribute to any other similarity, i.e. the similarities S_{X_i} are linear combinations of factors,

$S_{X_i} = \sum_{j=1}^{l} a_{ij} F_j + U_i$, where the factor loading a_{ij} indicates the "influence" of F_j on S_{X_i}. In our case a factor is the similarity to a hypothetical text. Now if there are distinct main branches or families of texts, the similarities to all texts of a certain family can be reduced for a great part to the similarity to a hypothetical central text of this family. Hence, if there are more than two families, for any family there will be one factor F_j such that a_{ij} is great for all X_i of this family but for no other or that a_{ij} is small for all X_i of this family and great for a contrasting family. If there are two families only, the similarities to the texts of the second family can be largely reduced to the dissimilarity to the hypothetical central text of the first family, i.e. the two families are distinguished by a single factor F_j with high a_{ij} for the texts of the first family and low a_{ij} for those of the second one (in this case $j = 1$).

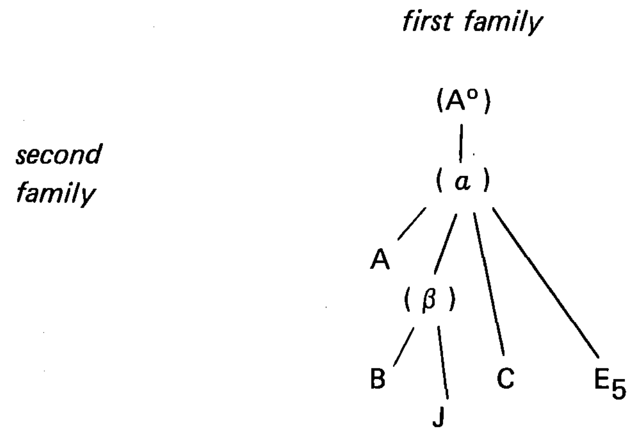
Example (8)

Matrix (a_{ij}) of factor loadings :

	F_1	F_2	F_3	F_4	F_5	.	.	.
E_5	.465	.613	.518	-.209	.152	.	.	.
E_{15}	-.796	.148	.428	.240	-.144	.	.	.
E_{16}	-.733	.050	.605	-.050	.040	.	.	.
A	.708	.502	-.029	-.170	-.411	.	.	.
B	.697	.462	-.069	.164	.425	.	.	.
C	.666	.534	-.171	.245	-.063	.	.	.
D	-.801	.182	-.431	-.262	.055	.	.	.
F	-.787	.366	-.062	.260	-.154	.	.	.
G	-.817	.343	-.084	-.202	.037	.	.	.
H	-.807	.199	-.161	.234	.060	.	.	.
I	-.679	.571	-.042	-.108	.047	.	.	.
J	.466	.776	-.056	.025	-.069	.	.	.
K	-.879	.298	-.150	-.030	.060	.	.	.

Obviously the loadings a_{i1} of the main factor F_1 separate the texts into two main families $\{A, B, C; E_5, J\}$ and $\{D, E_{15}, E_{16}, F, G, H, I, K\}$ with E_5 and J having a certain individual position, which is shown again by F_2. This is in accordance with the results of conventional methods applied to the same data (one specimen sector of 500 records of the old English glossed

Psalter versions) : according to F. BERGHAUS (2) the most probable stem is as follows :



where A almost equals *a*. Therefore, since a_{A1}, a_{B1}, a_{C1} and $a_{E_5 1}$, a_{J1} are very similar, the hypothetical text represented by F_1 is different from the lost text *a*.

Searching for subfamilies one could now try to apply the same methods of correlation and factor analysis to the single main families, then to the single subfamilies etc. But this is much more complicated than the following methods of automatic classification (5).

METHODS OF AUTOMATIC CLASSIFICATION

a) Divisive method

The position of a text X_i among the other texts is given by the N-tupel

$(s_{i1}, s_{i2}, ..., s_{iN})$ of the similarities to them. $(s_{i1}, ..., s_{iN})$ is a point of the N-dimensional space R^N, and the distance d_{ij} between two texts X_i and X_j can be defined by the Euclidean norm :

$$d_{ij} := \| (s_{i1}, ..., s_{iN}) - (s_{j1}, ..., s_{jN}) \|$$

$$:= \sqrt{\sum_{k=1}^{N} (s_{ik} - s_{jk})^2}$$

Of course $d_{ij} = d_{ji}$, i.e. the distance-matrix $(d_{ij})_{N \times N}$ is symmetrical. Furthermore, for a subset A_r of n_r texts, i.e. for $A_r = \{X_{i_1}, ..., X_{i_{n_r}}\}$, the centroid is defined by

$$\bar{s}_{A_r} := \frac{1}{n_r} \sum_{X_i \in A_r} (s_{i1}, ..., s_{iN})$$

$$= \left(\frac{1}{n_r} \sum_{X_i \in A_r} s_{i1}, ..., \frac{1}{n_r} \sum_{X_i \in A_r} s_{iN} \right)$$

For X_i let $s_i := (s_{i1}, ..., s_{iN})$.

Now the complete set of texts $A = \{X_1, ..., X_N\}$ is divided into two disjoint subsets A_1, A_2 by the following process (6) :

1) Let $A_1 = \{X_i\}$, $A_2 = \{X_j\}$, where X_i and X_j are texts with maximal

distance d_{ij} in $(d_{ij})_{N \times N}$.

2) In the remaining set $A \setminus A_1 \setminus A_2$ find
X'_i with minimal distance from $\bar{s}_{A_1}$ and
X'_j with minimal distance from $\bar{s}_{A_2}$.

3) If $\| s'_i - \bar{s}_{A_1} \| < \| s'_j - \bar{s}_{A_2} \|$, add X'_i to A_1 (i.e. form a new A_1 by uniting the former A_1 and $\{X'_i\}$).
If $\| s'_i - \bar{s}_{A_1} \| \geqslant \| s_j - \bar{s}_{A_2} \|$, add X'_j to A_2.

4) Repeat 2) and 3) until all N texts are distributed among A_1 and A_2.

Then by the same process 1)-4)both A_1 and A_2 are divided into two subsets, then these subsets are divided etc. until no subset contains more than two texts.

b) Agglomerative method

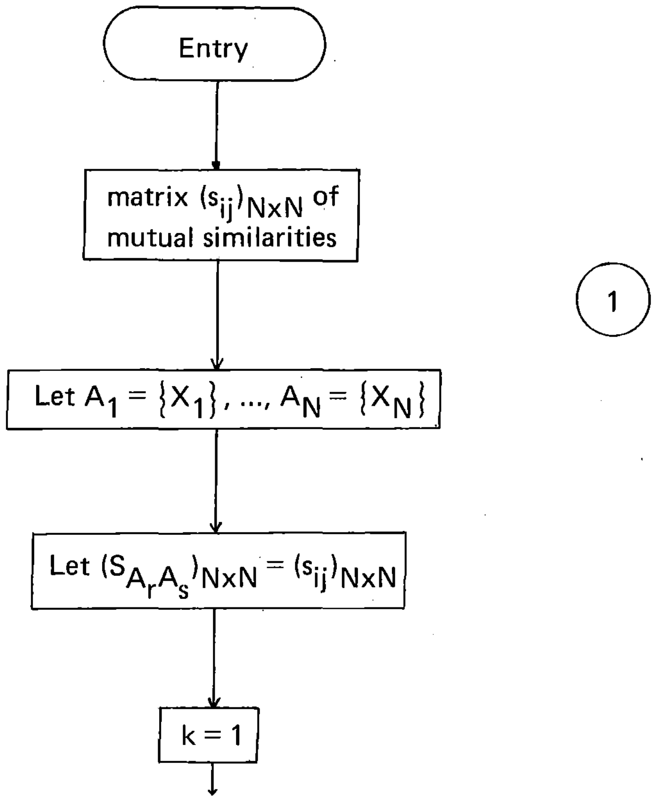
Instead of successive division of the complete set of texts, we now start with the single texts as the smallest sub-families or subsets and proceed by successive union of the two most similar subsets (7). The similarity $S_{A_r A_s}$ of two sets A_r and A_s of n_r and n_s texts respectively is given by

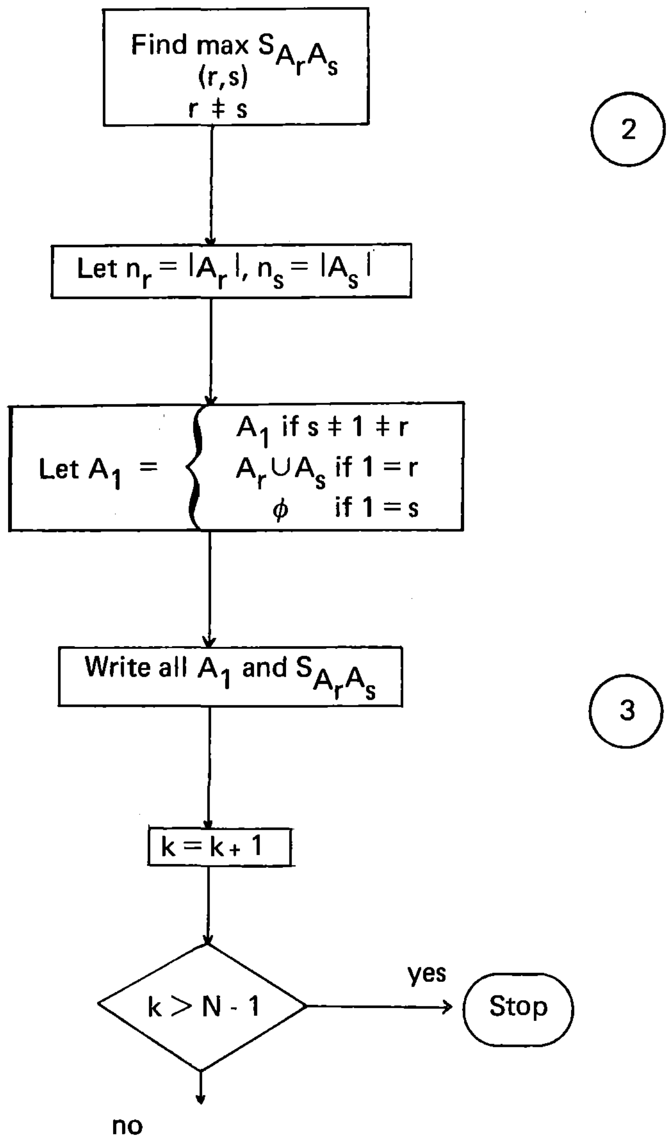
$$S_{A_r A_s} := \frac{1}{n_r n_s} \sum_{X_i \in A_r} \sum_{X_j \in A_s} s_{ij}.$$

From this definition there follows the recursion formula

$$S_{A_r \cup A_s,\, A_1} = \frac{n_r}{n_r + n_s} S_{A_r A_1} + \frac{n_s}{n_r + n_s} S_{A_s A_1},$$

i.e. after the union of the two most similar subsets we need not use the previous definition of $S_{A_r A_s}$ to find the new similarities. The complete classification can be carried out according to the following flowchart :





Find max $S_{A_r A_s}$ (r,s) $r \neq s$
2
Let $n_r = |A_r|$, $n_s = |A_s|$
Let $A_1 = \begin{cases} A_1 & \text{if } s \neq 1 \neq r \\ A_r \cup A_s & \text{if } 1 = r \\ \phi & \text{if } 1 = s \end{cases}$
Write all A_1 and $S_{A_r A_s}$
3
$k = k + 1$
$k > N - 1$
yes
Stop
no

Let

$$s_{A_rA_1} = \frac{n_r}{n_r + n_s} s_{A_rA_1} + \frac{n_s}{n_r + n_s} s_{A_sA_1}$$

and $s_{A_1A_r} = s_{A_rA_1}$

↓

Go to 2

Instead of the mutual similarities, we can use similarities of position in the complete set of texts. This is done by inserting at ① the following steps :

matrix $(d_{ij})_{N\times N}$
of distances

↓

$$\text{Let } E = \frac{2}{N\,(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}$$

↓

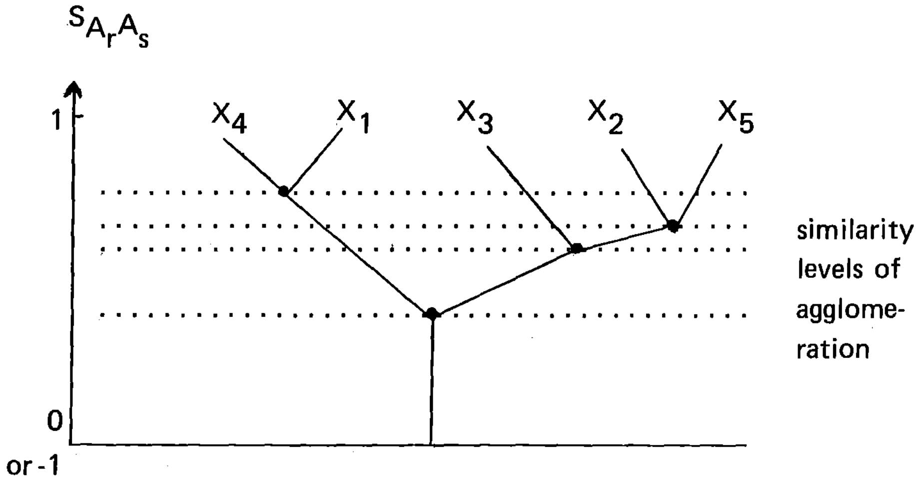
Let $s_{ij} = \dfrac{E - d_{ij}^2}{E + d_{ij}^2}$

↓

matrix $(s_{ij})_{N\times N}$ of similarities of position

↓

For each agglomeration-step the resulting partition into text-families (one less) and the similarity level of agglomeration is given by ③ . The complete classification can be represented by a dendrogram of the following type :



(A similar diagram can be given for the divisive method)

One must notice, however, that such dendrograms are generally no stems but a classification into text-families such that within each family A_r the average similarity is lower than within each subfamily of A_r. Though a direct dependence of texts cannot be shown by these methods of automatic classification, they will be useful especially in the cases of widespread contamination and great numbers of texts to be compared.

Dietmar NAJOCK

NOTES

(1) Cf. my contribution in this periodical n° 2, 1972.

(2) J. MAU intends to describe the computer programmes of our tests and F. BERGHAUS will compare the results of automatic classification with those of the conventional methods in the tradition of old English Psalter versions, both in this periodical.

(3) Since $r_{ij} < 0 \Leftrightarrow \sum_{k=1}^{N} (s_{ik} - \bar{s}_i)(s_{jk} - \bar{s}_j) < 0$, the same method can be applied to the covariance-matrix instead of the correlation-matrix.

(4) Cf. this periodical, n^os^ 3 and 4, 1972.

(5) A survey of statistical methods of automatic classification and further literature is given by H.H. BOCK in *Statistische Methoden II,* ed. by E. WALTER, *Lecture Notes in Operations Research and Mathematical Systems* vol. 39, Berlin 1970.

(6) According to GOWER (see (5)).

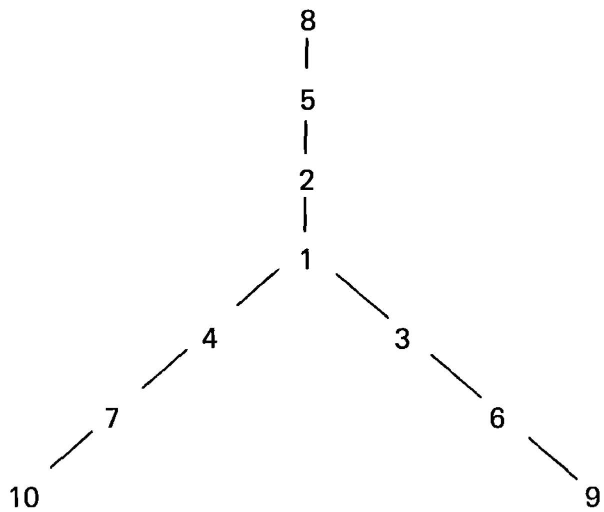
(7) This principle of agglomeration is the "average-link pair-group method" of SOKAL/SNEATH (see (5)). Instead of the successive union of the most similar subsets, i.e. of subsets A_r and A_s with maximal $S_{A_r A_s}$, we have also used the union of those

subsets A_r and A_s which contain the most similar texts not yet combined (first searching for maximal s_{ij} such that $X_i \in A_r$, $X_j \in A_s$, $r \neq s$ and then uniting the respective subsets A_r and A_s). The results are almost the same, seldom better or worse than those of the average-link pair-group method (see (2)). The second principle of agglomeration, however, which might be called single-link pair-group method, does not guarantee that the average similarity within a family is generally lower than within a sub-family of this family.

(8) The following three examples of (fictive) symmetrical text-enchainments show a certain difficulty in the application of factor analysis in some cases.

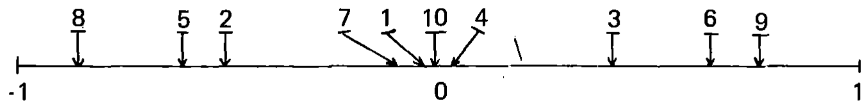
Example 1

The enchainment

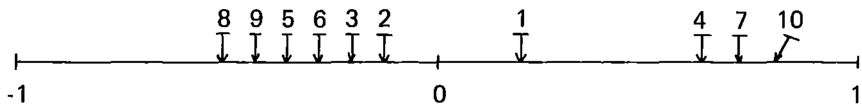


with equal similarities for all pairs of texts which are separated by the same number of intermediate texts, yields the following maps from the texts $\{1, ..., 10\}$ to the factor loadings in $[-1, 1]$:

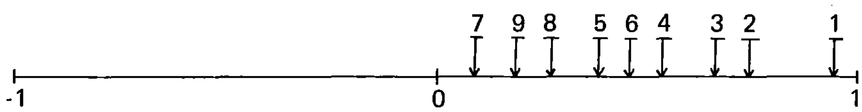
F_1 :



F_2 :



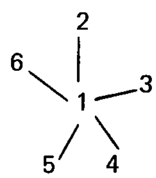
F_3 :



F_1, F_2 separate the families and F_3 represents the centre, but in spite of the symmetry of the enchainment, the representation of the families differs by increasing factor loadings for increasing indices of the factors. This effect, which can be seen more clearly in example 2, leads in example 3 to the destruction of families.

Example 2

The enchainment



with the similarity-matrix

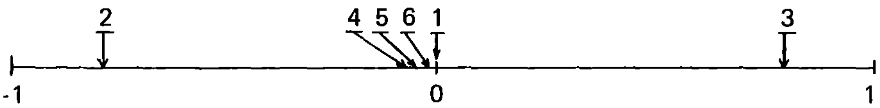
	1	2	3	4	5	6
1	1.0	.8	.8	.8	.8	.8
2	.8	1.0	.6	.6	.6	.6
3	.8	.6	1.0	.6	.6	.6
4	.8	.6	.6	1.0	.6	.6
5	.8	.6	.6	.6	1.0	.6
6	.8	.6	.6	.6	.6	1.0

yields the following matrix of factor loadings :

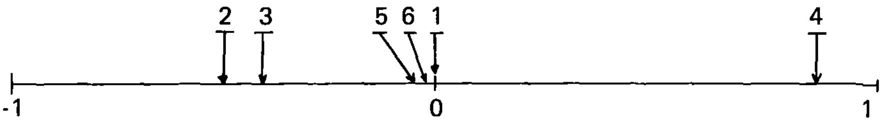
	F_1	F_2	F_3	F_4	F_5
1	.999	-.000	.000	-.000	-.000
2	.293	-.795	-.446	-.367	-.260
3	.293	.776	-.408	-.299	-.238
4	.293	-.014	.881	-.290	-.231
5	.293	-.013	-.014	.929	-.227
6	.293	-.013	-.014	-.013	.956

i.e. we have the maps

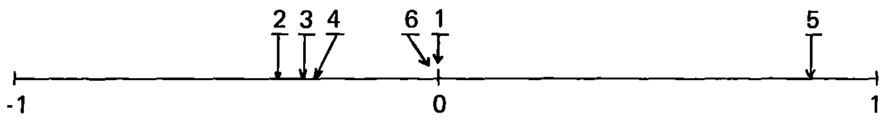
F_2 :



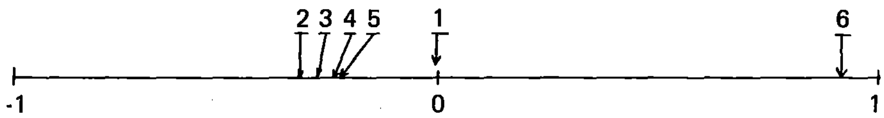
F_3 :



F_4 :

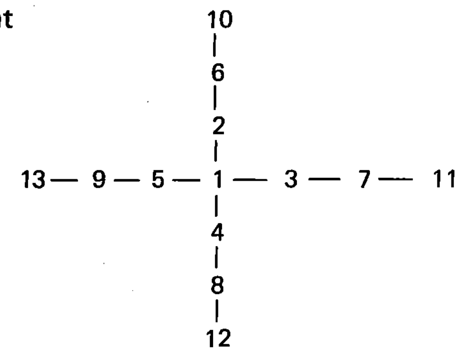


F_5 :



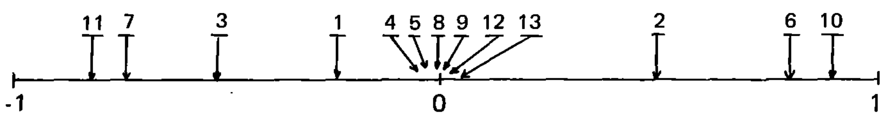
Example 3

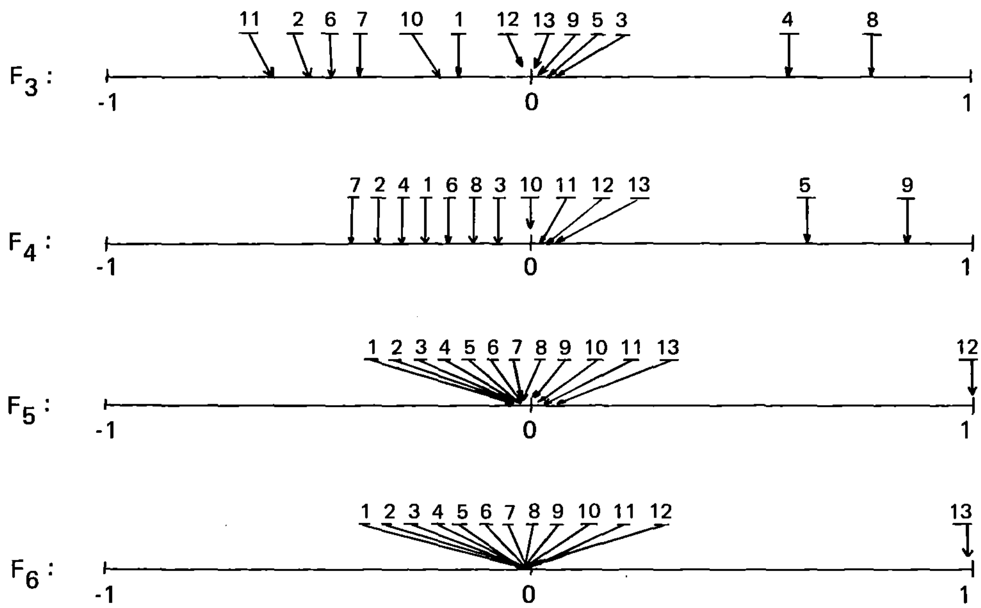
The enchainment



with symmetrical similarities as in examples 1 and 2 yields the following maps :

F_2 :





Here the correlation method is clearly superior.