*THE VALIDITY OF VARIOUS METHODS OF AUTOMATIC CLUSTERING IN DISCOVERING MANUSCRIPT RELATIONSHIPS**

(Dedicated to Professor Dr. Jürgen Mau on the Occasion of his 60th Birthday)

Since the days of K. Lachmann (1) the main principle of affiliating manuscripts has been to measure scribal errors in comparison with the autograph or, if that has been lost, with an idealized collation or reference manuscript (2). In the meantime, methodology and terminology have been brought to perfection, especially by P. Maas (3), to whom we owe a number of terms (like archetype, hyparchetype, distinctive error, contamination, etc.) and reliable criteria to decide between possible stemmatic constellations. The Benedictine Father J. Froger (4) has summarized the rules of classical textual criticism and its inherent mathematical logic. In doing so he made clear the theoretical presuppositions for an extensive automatisation of textual criticism and affiliation of manuscripts. But in spite of all progress in that field the method has remained the same as it was in the first half of the 19th century : Lachmann's criterion of continuous deterioration of texts from one transmissional step to the next is still considered valid. Two major difficulties arose right at the beginning : the incorporation of contaminated texts (5) into a stemmatic system and the question whether the scribes should be conceded a certain capacity to *ameliorate* texts rather than to *deteriorate* them. But even if these problems could have been solved satisfactorily - and they never have been - the main objection to this method is that it can only be applied to a restricted category of texts with a rather homogeneous tradition (6). Apart from a very few lucky examples (and unlike medieval Latin tradition for instance) intricate textual transmission is the regular case in the earlier stages of Germanic literature. Here the rules of classical manuscript affiliation can scarcely be applied.

This evidence led some scholars like G. Kane (7) to believe that any attempt at all to establish stemmata should be rejected. This opinion holds true as long as conventional methods (Lachmann, Maas, Froger) are applied to text families with a widespread tradition (where most links are missing), with no autograph being at hand, with rather few chances to decide on whether a specific reading is correct or incorrect (in order to reconstruct a reference manuscript), and with a broad spectrum of chronologically and dialectally determined lexical, morphological, phonological and graphical variants. There can be no doubt that a method which relies on the evidence of error types can do little or nothing to discover the genetic relations of texts of this kind.

One of the hardest nuts to crack in the field of complicated textual transmission is the relationship of the Old English interlinear glosses of the Psalter (8). In glossed 'literature' of this kind two additional problems arise. First : The underlying Latin texts can always lead to direct glossing (i.e. that a specific reading is a translation of the Latin lemma rather than an adaptation from the exemplar the copyist had in front of him (9)). This creates polygenetic correspondences which quite often cannot clearly be distinguished from contaminations. Second : There are two main Latin versions of the Psalter, Roman and Gallican (sometimes with Roman readings in the Gallican versions and *vice versa*, and, to complicate things, with readings from the Hebrew version in both of them), and quite a lot of individual variation (often caused by the influence of patristic commentaries).

All fully or partly glossed OE interlinear versions of the Psalter formed the material for the tests described in this article. For an investigation of the genetic relations of these texts (and others with similar problems) *qualitative* methods have to be replaced by *quantitative* ones. Any isolated case - however striking it might be - is of little value because it can be explained either by contamination or by polygenesis. Only a statistically valid set of information increases the likelihood that statements about possible relations will have a satisfying degree of reliability.

Before I start describing the application of statistical methods on the material I think it necessary to give a brief survey of data-storage and the first steps of evaluation. Any investigation of texts with difficulties such as those described above must necessarily be undertaken on a *lexical basis* (10). The manuscripts have to be collated the traditional way (11). An entry is made for each reference where at least one of the versions has a variant reading (this produced ca. 14.000 records). Besides lexicographical information (which I recorded for other purposes) and reference to the entry the record contains a sigla-field which is divided into subfields presenting a true image of the actual combination of the MSS.

(marked by their sigla), for example : =E5G=ABC=DFKGHJI=I. It should be noted that MSS. with double glosses can occur in different subfields of the sigla-field. Numerals following sigla denote any hand in a MS. except the first (i.e. E5 = the fifth scribe in E, A = the first scribe in A). Some peripheral information had to be taken into account : each record is marked according to whether the Latin versions are identical in all MSS. ('R'), whether the versions are clearly separated ('G'), or whether there occurs any form of mixture between the versions ('X'). Some special information turned out to be very useful in the end (12). Each word was classified : nouns, adjectives, adverbs, and verbs by 'N', all form-words by 'F'.

As a basis for all further investigations first of all a frequency-matrix has to be set up. This sounds like a simple task, but some difficulties have to be considered beforehand. Each MS. shows a certain number of isolated readings. The relative frequency of these readings ranges from 2% in one MS. up to as much as 38% in another MS. The relative frequency of isolated readings does not reveal anything about the combination of MSS. in the stemma but, if not neutralized, would be allowed to affect the relative frequency of common readings of any combination of two MSS. In order to come to comparable figures the isolated readings are eliminated before the frequencies of common readings are worked out.

But there is still another frequently neglected problem. Scribes may change their exemplar during the course of their work. As the results will show this is not a theoretical presupposition but in fact must have happened fairly often (especially in the larger scriptoria where several versions of the same text existed). To discern these parts of different origin in the MSS. the total material has to be fractioned into portions or sections. After experiments with (arbitrarily chosen) sections of 100, 300, 600, and 1000 records each the most reliable solution (at least in our case) seemed to be to take portions of 500 records each. A section of that size contains enough information to be statistically valid and is small enough that parts originating from different sources normally show up. As a basis for all further investigations we now have as many frequency-matrices as there are sections in the material.

1. GRAPHIC DISPLAY OF FREQUENCY-MATRICES :

A very simple method to give the results far more lucidity than they have in the bulk of frequency-matrices is to take the figures for each MS. (i.e. the relative correspondences with other MSS.) out of the frequency-matrices and have them plotted (in our case sketched automatically on an IBM-plotter) in such a way that the portions are shown in the x-axis, the percentages of correspondence in the y-axis. This produces as many graphs as there are MSS. involved in the investigation. It can be clearly seen from the graphs (cp. diagram 1, relations of the gloss A) that there are obviously two main groups or families of MSS. (ABC and DFKH). The relations of the gloss J (cp. diagram 2) show that the scribe in fact changed his immediate exemplar, some of the sections belonging to the ABC-group some to the DFKH-group. If the material had not been fractioned the total figures for J would have been absolutely uninformative. Even continuous contamination can be detected in the

~~graphs by careful observation of the antitype bunch of curves. If~~ for instance a gloss belonging to type ABC (like most parts of E) shows a significantly higher rate of DFKH-readings than ABC do throughout the text or in certain parts of the text these can be discerned by comparison of the antitype curves of ABC with those of E (13).

2. PEARSONIAN CORRELATION COEFFICIENT :

The correlation coefficient (r) (14) of two columns i and j with n figures (k = 1...n) of the frequency-matrix (s_{ij}) is worked out by means of the following formula :

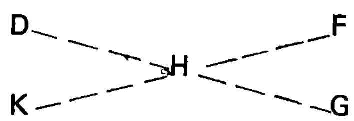
$$r_{ij} = \frac{\Sigma \, (s_{ik} - \bar{s}_i) \, (s_{jk} - \bar{s}_j)}{\sqrt{\Sigma \, (s_{ik} - \bar{s}_i)^2 \; \Sigma \, (s_{jk} - \bar{s}_j)^2}}$$

(where $\bar{s}_i = 1/n \; \Sigma \; s_{ik}$ and $\bar{s}_j$ analogously). With the aid of the correlation coefficient it can be found out for any combination of two columns what degree of linearity they have. The more similar texts are (spoken in terms of behaviour against all others), the higher the correlation coefficients will be. The correlation coefficient ranges from + 1 (in our case : highest possible similarity) to - 1 (highest dissimilarity). Successive comparison of all possible pairings of columns in the frequency-matrix produces a new matrix : the correlation-matrix (cp. diagram 3).

From the ~~distribution of plus- and minus-figures in the matrix it~~ can be seen that (at least in portion 4) there are two distinct families of MSS. : ABCE5J and DFKGH, and that I obviously has an individual position, being somehow connected with DFKGH and more or less indifferent towards ABCE5J. In the ABCE5J-group B is nearer to A than to C and C nearer to A than to B. As J is closest to B and E5 closest to C the enchainment of the MSS. must be :

J – B – A – C – E5

The situation in the DFKGH-group is much more complicated but quite obviously H (or a lost text very near to H) must have a central position in the enchainment as all other MSS. show highest correlation with H :



H itself is nearest to D. The results of more conventional methods applied to the same data show accordance of these constellation-types with the final stemma (cp. diagram 5, stemma of the OE glossed psalters).

The correlation-matrices can be displayed (as has been done with the frequency-matrices) in a graph (cp. diagram 4) in order to quickly recognize how constellation-types change. Apart from the fact that the groups of curves are much better separated than they were in the display of frequency-matrices the main advantage of the correlation-matrices is that fragmentary MSS. recorded in different parts of the same text can be compared as to whether they are derived from the same (or a similar) source or not. In the frequency-matrix their relative correspondence must necessarily be 0.0%, for they never occur in the same sigla-fields. But this indirect comparison of two fragmentary MSS. is only possible if the rest of the tradition is not heterogeneous, i.e. if all other MSS. stick to their groups. If they do not (as for example J when it changes its sources, cp. diagram 2) these outlayers have to be eliminated or neutralized before a comparison of fragmentary MSS. can start.

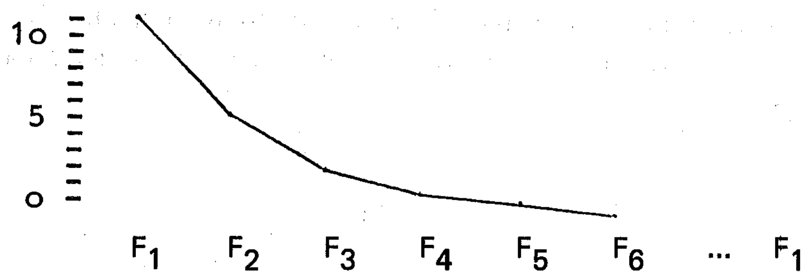
3. FACTOR ANALYSIS :

As has been described excellently by D. Najock "factor analysis aims at reducing the complete set of variables [...] [i.e. in our case the similarities s_{xi} as have been achieved in the correlation-matrix] to a small number of hypothetical variables" (15) F_j (with J = 1...1). This is worked out with the following formula :

$$s_{xi} = \Sigma a_{ij} F_j + U_i$$

"where the factor loading a_{ij} indicates the 'influence' of F_j on s_{xi}" (U_i being a specific factor which can be ignored). For our purpose the common factor F_j can be defined as the similarity to a hypothetical central text. In case that there are two groups of texts "the similarities to the texts of the second family can largely be 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 the texts of the second one" (16). Now a matrix of factor loadings can be set up (cp. diagram 6) where the main factor F_1 is in the first column, the next in importance in the second column, and so on. The main factor clearly divides the texts into two groups ABCE4E5G2J and DFKGH giving I a certain individual position. F_2 shows that G2 and J must have some features in common (which is indeed the case). The individual position of I is confirmed by F_3.

As the importance of factors decreases according to the following graph :



it seems to be legitimate to concentrate on the first two or three factors when the texts are classified, which can be done automatically by the machine. Classification of this kind produces two-dimensional graphs (cp. diagram 7) where the clustering of MSS. is clearly visible. Experiments with three-dimensional models yielded even better results but require a great deal of time and energy as the models have to be constructed by hand.

4. CLUSTERING IN N-DIMENSIONAL SPACE :

The position of a manuscript x_i in n-dimensional space is fixed by the relations it has to all other MSS., i.e. the similarities with others which can be taken. either from the frequency-matrix or from the correlation-matrix. The position of x_i (with n figures (k = 1...n) in its column) is defined by means of the Euclidian norm :

$$/\vec{x}_i/ \quad := \quad \sqrt{\Sigma\ x_{ik}^{\ 2}}$$

This produces a cloud of points in n-dimensional space, each point representing and characterizing the position of a MS. The next step is to work out the distances d_{ij} between any combination of two MSS. x_i and x_j :

$$d_{ij} \quad := \quad \sqrt{\Sigma\ (x_{ik} - x_{jk})^{\ 2}}$$

Going through all possible combinations of columns a distance-matrix is produced which should be normalized in such a way that all distançes range from - 1 to + 1. The distance-matrix can be evaluated with two different methods :

a. Divisive method :

Two points with the greatest distance recorded are looked up in the distance-matrix and for each of the other points in the cloud it is worked out whether it is nearer to one or the other of the two most distant points, thereby dividing the whole cloud into two sections or subsets. Then dividing starts again with one of the two subsets, progresses with the second subset, and so on down to the smallest possible units of two points. This successive division can be displayed in a dendrogram (cp. diagram 8). It should be noted that a dendrogram is not a stemma. The dendrogram (read from top to bottom) shows the increasing similarities from each divisive step to the next : the smaller the subsets, the higher the similarity.

b. Agglomerative method :

In a way agglomeration means reversing the above method by looking up the smallest distance in the cloud of points. When that has been found the centroid for these two points is worked out (17) and agglomeration goes on until no single subset of one is left in the cloud. Again the results can be displayed in a dendrogram (cp. diagram 9) which in this case should be read from bottom to top.

The results offered by the divisive and the agglomerative method vary in details (as can be seen in section 2, cp. diagrams 8 and 9). In all cases the results of agglomeration were much nearer to what

~~could be expected from the final stemma (cp. diagram 5). The~~ reason for this may be illustrated with a simple example. If the divisive method is applied on the following string of eight points :

1 2   3 4 5   6 7 8

no. 5 would go with 6, 7, 8 after the first divisive step whereas with the agglomerative method it would never be in contact with these but go with 3 and 4 where it actually belongs.

Furthermore users should be warned that the divisive method takes about 5 times more computer time than the agglomerative method.

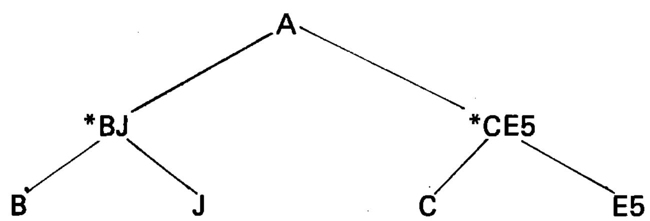
RESULTS

None of the methods that have briefly been described above leads directly to stemmata. Factor analysis and the agglomeration of clusters in n-dimensional space produce a very good survey of how MSS. are grouped, but genetic relations in the material cannot easily be detected, except in a very few cases : The position of G2J in section 4 for example (cp. diagram 7) leads to the assumption that the two, obviously belonging to the family ABCE4E5, form a certain subgroup with common features, and that they may (perhaps) originate from the same source. Likewise the arrangement of E3E14 in the dendrogram of section 2 (cp. diagram 9) seems to force us to accept that the two hands copied from the same exemplar. So the information that can be achieved by applying these methods is restricted to a very rough first classification of the MSS. Consequently factor analysis and clustering in n-dimensional space should be used as a first step to group large traditions with several dozens or even hundreds of MSS. As far as I can see, however, the main range of application lies in the field of stylo-statistics. To solve problems of authorship, scriptoria, etc. on the basis of stylistical (or other) features automatic classification of this kind will be extremely helpful as no answers to genetic questions are required.

The graphic display of frequency-matrices and operations with the correlation coefficient lead us much nearer to an interpretation in a stemmatic sense (even telling us about contaminative influences). One first step into that direction was the production of enchainments as could be seen above. But the enchainment

J -- B -- A -- C -- E5

for example still has to be interpreted and it is a long way from here to the final stemmatic constellation (cp. diagram 5) :



All sorts of other data have to be taken into account like the numbers of exclusive coocurrences of words in any combination of two or more mss. (an item of information easily supplied by the computer once the material has been stored in machine-readable form) and more "usual" philological arguments like date of MSS., dialectal features, and so on. But even if all this has been considered carefully, very often a decision for possible stemmatic constellations depends on logic, probability, or even guesswork.

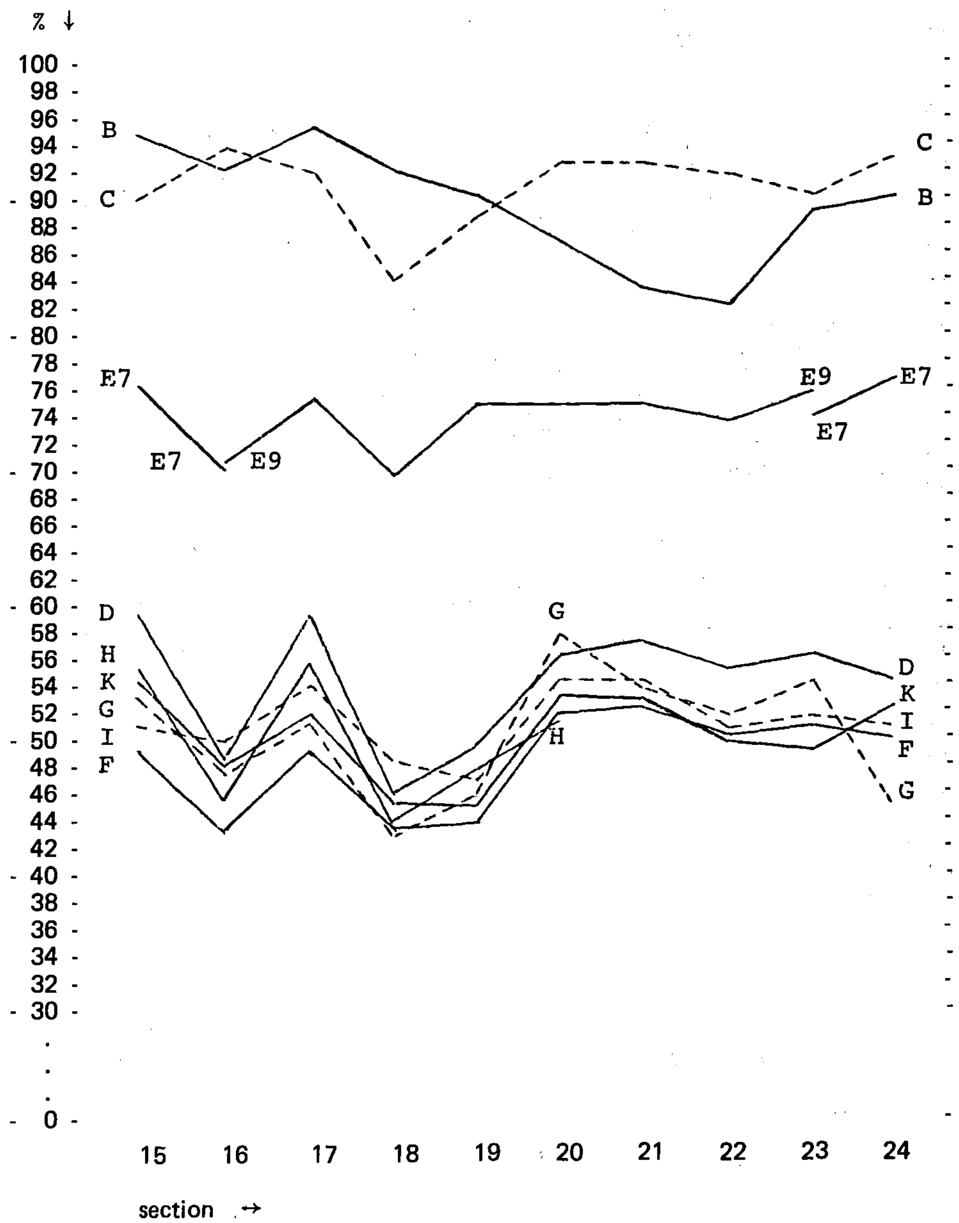
Thus stemmata that have been constructed for loosely linked traditions (e.g. only 15 of a much larger number of OE glossed psalters are still extant) should never be declared as perfect arrangement of the original constellation, but should be regarded as an imperfect visual aid towards a better knowledge of the genetic relations of texts.

Sem. f. Engl. Phil.
34 Göttingen
Nikolausberger Weg 15

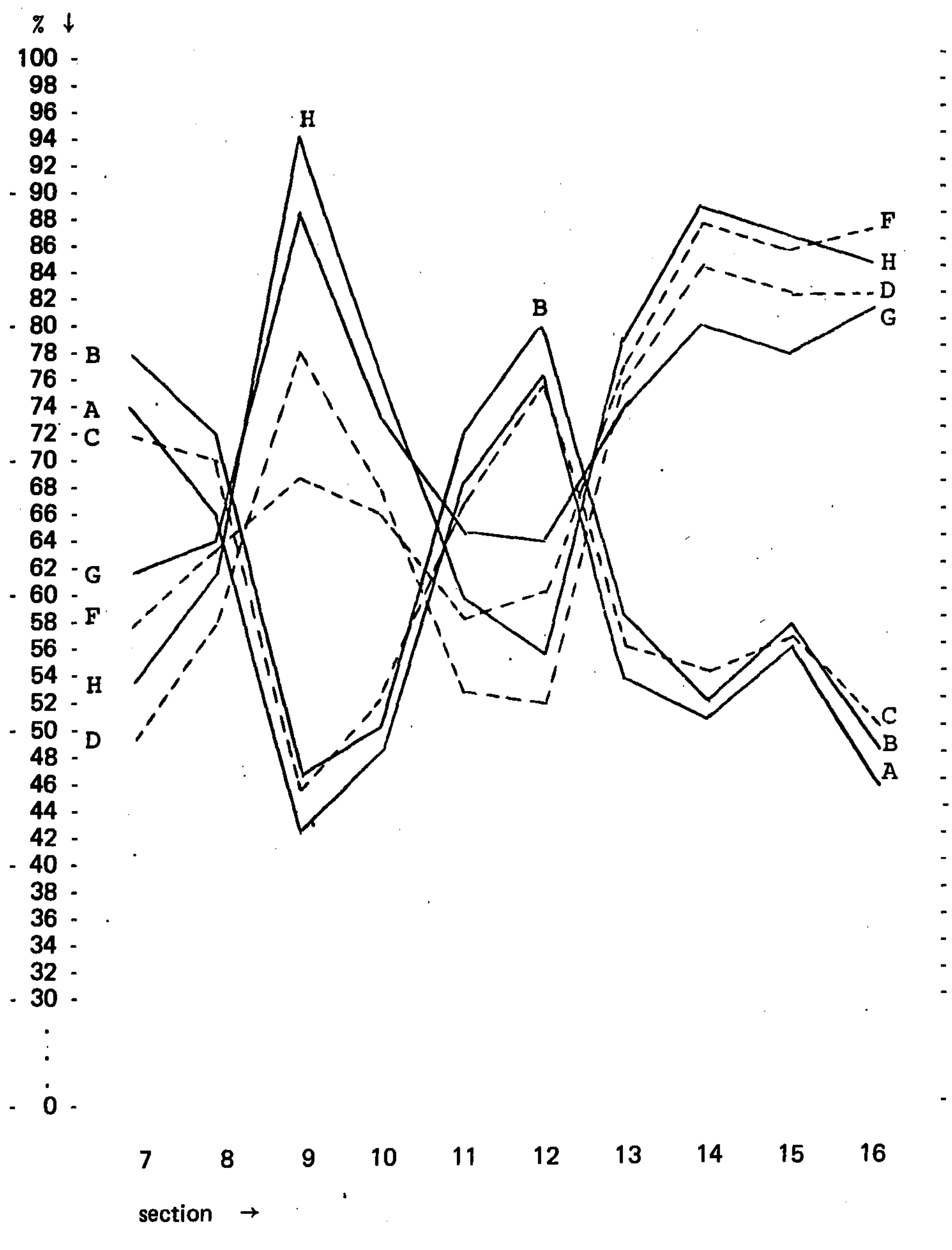
Frank-G. BERGHAUS

DIAGRAMS

1. Relations of the gloss A (sections 15 - 24) :



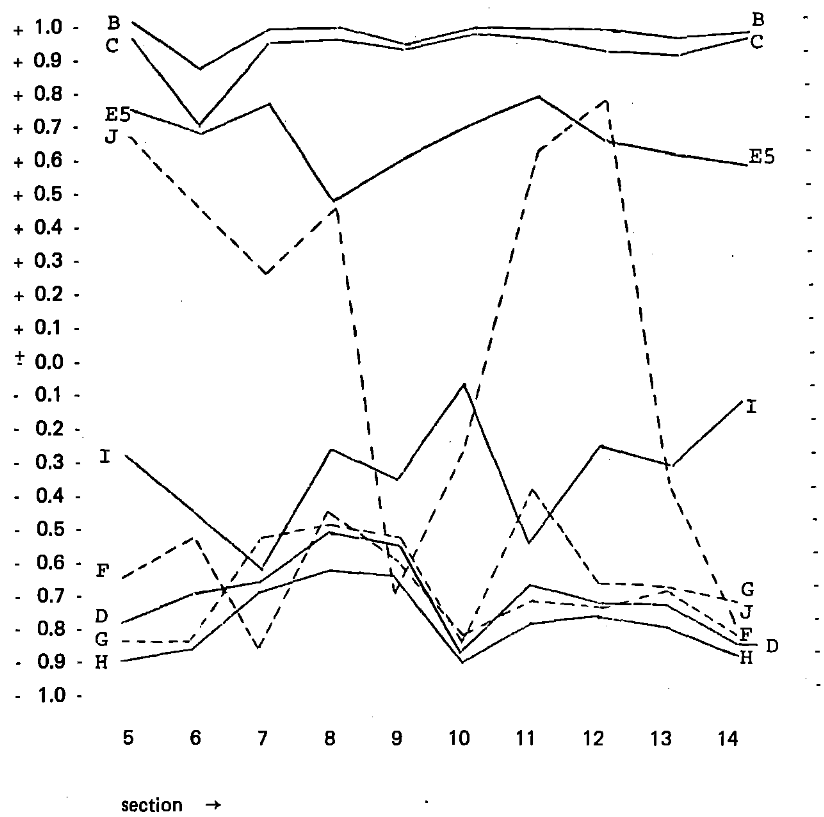
2. Relations of the gloss J (sections 7 - 16)



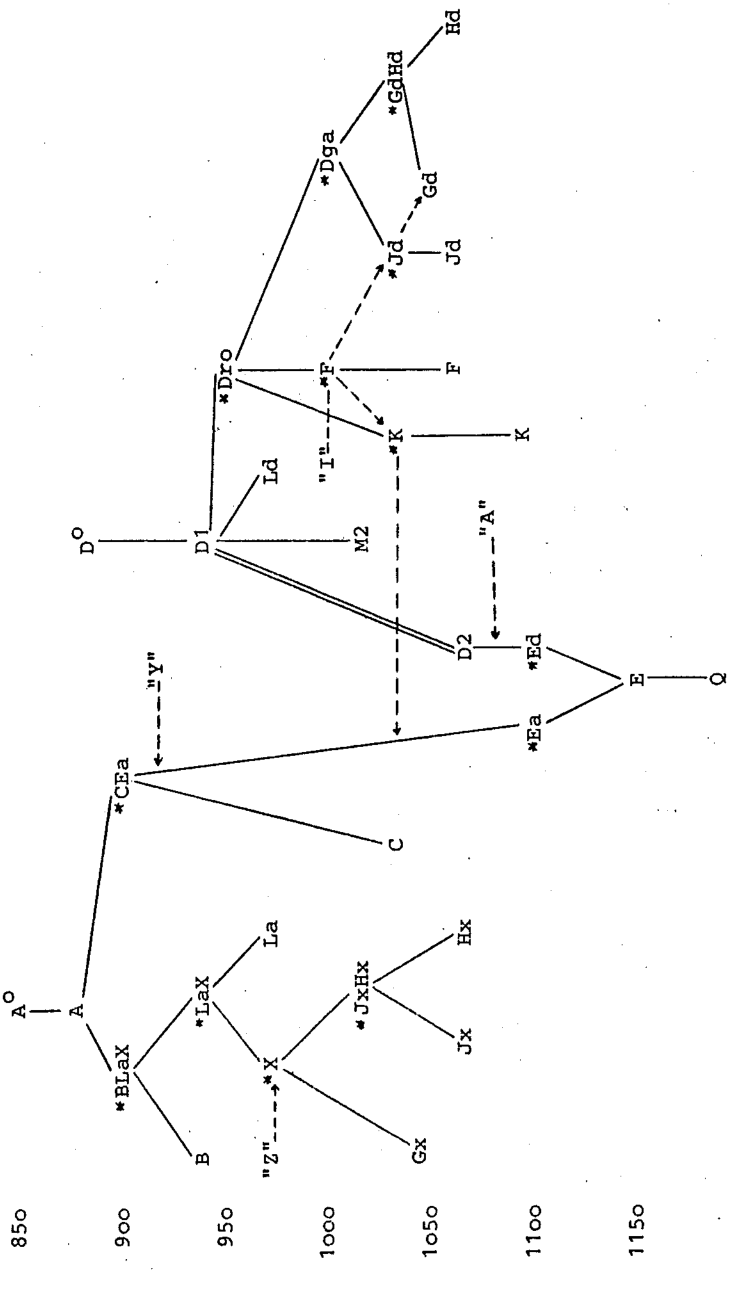
3. Correlation coefficients (section 4, reduced to 11 mss.)

	A	B	C	E5	J	D	F	K	G	H	I
A	1.00	.99	.98	.67	.43	-.54	-.52	-.49	-.34	-.49	-.04
B	.99	1.00	.97	.66	.47	-.56	-.54	-.52	-.37	-.51	-.07
C	.98	.97	1.00	.70	.40	-.50	-.46	-.47	-.29	-.46	-.00
E5	.67	.66	.70	1.00	.42	-.40	-.33	-.32	-.21	-.27	.12
J	.43	.47	.40	.42	1.00	-.74	-.56	-.72	-.46	-.69	-.17
D	-.54	-.56	-.50	-.40	-.74	1.00	.83	.89	.82	.98	.23
F	-.52	-.54	-.46	-.33	-.56	.83	1.00	.80	.83	.86	.56
K	-.49	-.52	-.47	-.32	-.72	.89	.80	1.00	.78	.90	.33
G	-.34	-.37	-.29	-.21	-.46	.82	.83	.78	1.00	.87	.45
H	-.49	-.51	-.46	-.27	-.69	.98	.86	.90	.87	1.00	.34
I	-.04	-.07	-.00	.12	-.17	.23	.56	.33	.45	.34	1.00

4. Correlation coefficients of the gloss A (sections 5 - 14)



5. Stemma of the OE. glossed psalters (except I and M)

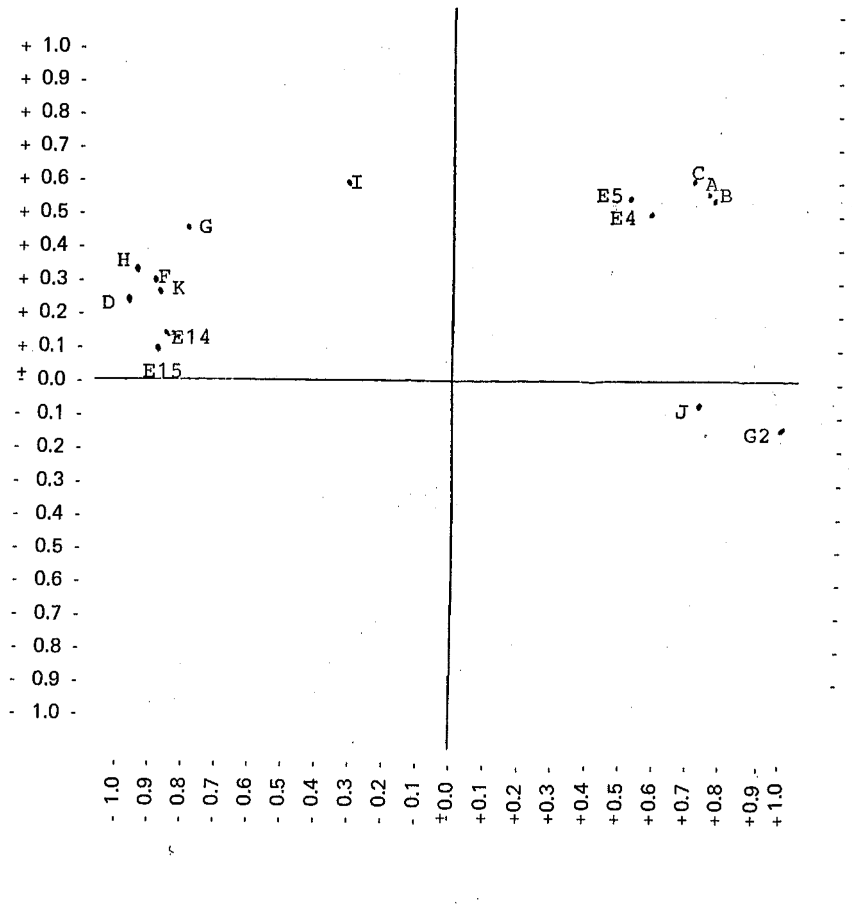


Extrait de la Revue (R.E.L.O.)
XIV, 1 à 4, 1978. C.I.P.L. - Université de Liège - Tous droits réservés.

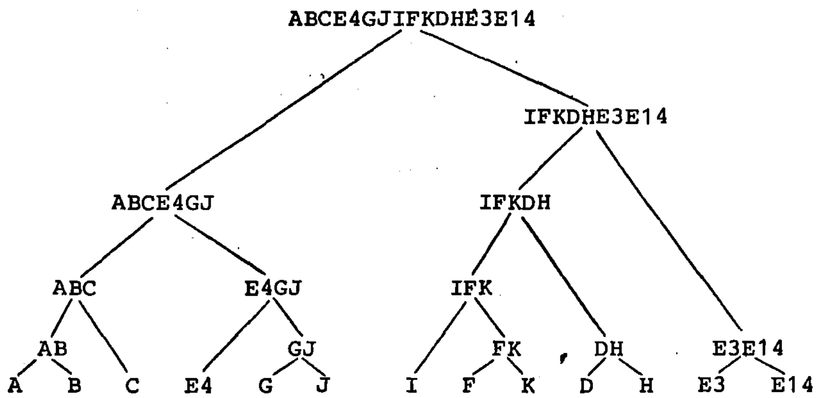
6. Matrix of factor loadings (section 4) :

	F_1	F_2	F_3	F_4	F_5	F_6
E4	.609721	.498957	.255560	.151141	-.059342	-.522751
E5	.527118	.578071	-.074489	-.601827	-.071418	-.074105
E14	-.827251	.155863	-.207885	-.453850	-.001308	.030916
E15	-.835425	.099514	.039115	.100916	.348253	-.006575
A	.758522	.579346	-.240987	.120761	.022068	.100390
B	.777124	.547265	-.235090	.114909	.070676	.105619
C	.715401	.594520	-.216328	.198894	.061801	.162670
D	-.930962	.203891	-.200485	.044969	.054437	-.033105
F	-.867904	.299693	.230087	.035112	.078870	.125566
G	-.779674	.448707	.079915	.081011	.257324	-.010616
G2	.951146	-.142518	.103636	-.083572	.100455	.057385
H	-.911919	.326332	-.112245	-.031602	.055247	-.094616
I	-.294396	.561042	.673008	-.009786	-.271064	.235743
J	.746007	-.069324	.340241	-.208744	.515564	-.007357
K	-.877558	.270213	-.094734	.057243	-.066080	-.122139

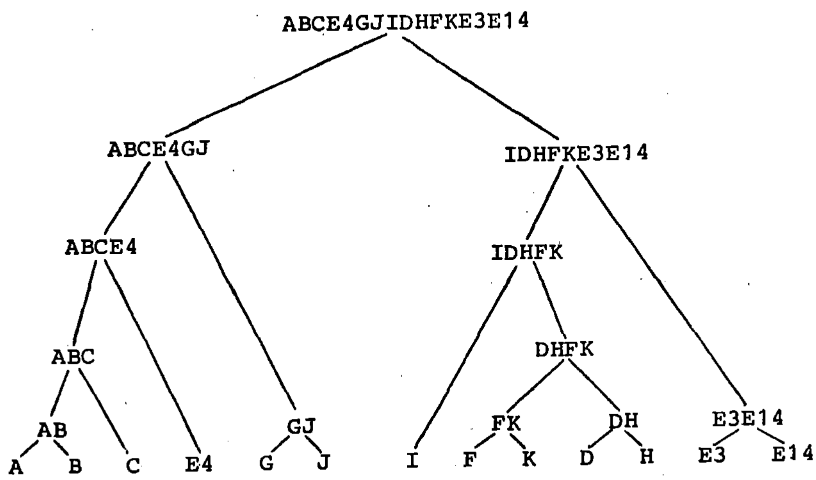
7. Two-dimensional graph of factor loadings (section 4)



8. Dendrogram (divisive method, section 2) :



9. Dendrogram (agglomerative method, section 2) :



NOTES

* The backbone of this article is a lecture held at McGill Univ., Montreal, and at the Pontifical Institute for Medieval Studies, Toronto, in November 1976.

(1) Cp. his introductions to *T. Lucreti Cari de Rerum Natura Libri Sex* (Berlin 1871, 4th ed.) and *Zu den Nibelungen und zur Klage* (Berlin 1836).

(2) There has been quite a long discussion (especially in France) in the past years whether "faute" should better be replaced by "inncvation" in order to achieve more objectivity in textual criticism, cp. for example : Quentin, *Essais de la critique textuelle* (Paris 1926); Bédier, *Romania* 54 (1928), p. 181; Roques, *Romania* 69 (1946/47), p. 116ff.; Fourquet, *Romania* 70 (1948/49), p. 85ff. But even operating with "innovation" instead of "faute" still implies thinking in categories of comparison of single readings in the MSS. with the subjective conception of the author's language and way of writing. A first step towards getting rid of subjectivity in textual criticism and making relative correspondences the basis of stemmatic conclusions has been done by Bévenot, *The Tradition of Manuscripts. A Study in the Transmission of St. Cyprian's Treatises* (Oxford 1961).

(3) *Textkritik* (Leipzig 1960, 4th ed.).

(4) *La critique des textes et son automatisation* (Paris 1968).

(5) Cp. Maas, *Textkritik*, p. 30 : "Gegen die Kontamination ist kein Kraut gewachsen"; cp. also Vidmanovà, "Les textes contaminés et l'ordinateur", *Revue* [de l'Organisation Internationale pour l'Etude des Langues Anciennes par Ordinateur] (1972), I, p. 5ff.

(6) The characteristics of a textual tradition to which Lachmann's criteria can be applied have excellently been described by Stackmann, "Mittelalterliche Texte als Aufgabe", *Festschrift für Jost Trier*, ed. by W. Foerste and K.H. Borck (Köln/Graz 1964), p. 240ff., see especially p. 246f.

(7) Cp. his introduction to *Piers Plowman I : The A Version*, (London 1960).

(8) The relationships of the OE glossed psalters (cp. also diagram 5) have been investigated by myself with more or less conventional methods in my doctoral thesis *Die Verwandtschaftsverhältnisse der altenglischen Interlinearversionen des Psalters und der Cantica* (Göttingen 1977, to appear in 1978).

(9) Very often direct glossing cannot easily be detected but some obvious examples can be found in the Eadwine Psalter [see my edition *Eadwine's Canterbury Psalter. An Edition with Notes and Collations of All Interlinear Versions of the Psalter* (to appear : Toronto 1978)] : Ps. 102.7 *felle hundes : pellicano*, Ps. 103.11 *on ðæm londum : onagri*, and others.

(10) Even very small lexical differences like for instance OE *hyhtan* versus *gehyhtan* must count as full lexical divergences. Experiments prove that variations like the one quoted (as well as variations in form-words) have the same quality as have "normal" variants like OE *hyhtan* versus *hopian;* cp. the chapter "Mögliche Beziehungen zwischen D und A" in my thesis, see above, fn. 8. This is why I cannot agree at all with Kane's opinion, *Piers Plowman I,* p. 59 : "the 'herd of dull commonplace readings' is an unreliable source of evidence for genetic relation".

(11) Automatic procedures have been proposed by Froger, "La collation des manuscrits à la machine électronique", *Bulletin de l'Institut de Recherches et d'Histoire des Textes* 13 (1964/65), p. 135ff., but applied to texts with extreme phonological, morphological, and graphical variation are far too complicated.

(12) The distribution of 'N' - to 'F' -records (in the OE psalters) is ca. 70% to 30%. Typical correctors show a significant deviation by having ca. 85% to 15% 'N' - to 'F' -records. This observation sometimes allows a definite decision between stemmatic possibilities and the historical classification of compilational layers in the material, cp. for instance the chapter "Die Glosse I" in my thesis, see above, fn. 8.

(13) Contaminative influences of K on E could be detected in sections 6-12 and 15-18, cp. the chapter "Mit ABC verwandte Teile von E" in my thesis, see above, fn. 8.

(14) For a full description see for instance Kreyszig, *Statistische Methoden und ihre Anwendungen* (Göttingen 1975, 5th ed.), p. 300ff.

(15) "Automatic Classifications of Texts by Methods of Multivariate Statistics", *Revue* (1973), II, p. 39.

(16) *Ibid.*, p. 40.

(17) *Ibid.*, p. 42.