THE AUTOMATISATION OF QUOTATION-FINDING

Quotation in the sense used here means not only verbal or literal citation but any kind of similarity between at least two passages of different texts or of different parts of the same text. Similarities may be due to simple word-for-word copying, to plagiarism in the general sense that certain formulae or patterns or ideas are borrowed, to the variation of common material, or to mere coincidence.

Especially in earlier periods of literature originality did not necessarily imply producing new ideas but also consisted of the variation of known sources.

Thus it is of some interest to the philologist to know where similar phrases occur in a fixed corpus of texts. Conventional methods of quotation-finding always mean time-consuming mechanical work : texts have to be collated line for line or, if fortunately they happen to be at hand, concordances must be examined. As in most cases where excerpting is involved here again the machine can be a useful help for the philologist and supply him with the necessary information for his interpretations.

But how can the computer be told to find passages where for example, patristic literature quotes the Old or New Testament, where the Old English poem *Beowulf* and *Andreas* use the same formula, or which particular line of Shakespeare's *Macbeth* coincides with any other line of the rest of the author's work ?

The following suggestions and solutions have been discussed in the computer branch of the Classics Dept. of Göttingen University (1). The test material for the investigation were the 32244 lines of the A-text (British Museum Ms. Cotton Caligula A.IX) of *Layamon's Brut* (2).

As hypothesis, quotation or similarity is defined as the correspondence of at least three words in a collated passage with three words from a ten word sequence of the passage in question. This provides redundant information, but redundancy can be reduced by neglecting commonplace words with a total frequency of more than fifty. In similar passages commonplace words (as well as of course ἅπαξ λεγόμενα) are generally not very informative. The hypothetical limits (3 out of 10, maximum total frequency of 50) are subject to experiment. They depend on type and length of the texts in question.

$T_{1...n}$ is to represent a certain quantity of texts to be compared with T_x, the investigated text. $T_{1...n}$ and T_x are united in a single mass storage file where the entries are numbered monotonously from 1 to m (3).

If a concordance to $T_{1...n}$ is available only T_x is sorted and merged into the concordance. If not, a combined concordance is made out of $T_{1...n}$ and T_x with a) cited word, b) entry number (EN) and c) context. This concordance is lemmatized and the lemmata are numbered in such a way that every cited word belonging to the same lemma has the same lemma number (LN). Synonyms can be given the same lemma number in order to include passages where the sense is similar but where non-identical words have been used.

The LN- and EN-columns of the concordance give all necessary information for the next steps of the investigation. The rest of the concordance can be

ignored. LN and EN form an index with LN being the first sorting key and EN the second, i.e. within each block of the same LN (LB) the EN are in ascending order. The EN belonging to T_x should be at the beginning of each LB.

The procedure of quotation-finding can be shown in a diagram or flow-chart (See Appendix).

The index of LN and EN is read to the end of the first LB. If there is no occurrence of T_x in LB the next block is taken as long as there are blocks left. If an EN belonging to T_x is found the whole block is stored in such a way that this block receives a block number (BN) which is identical with the first EN.

For example :

LN1	EN1
:	EN2
:	EN3
:	:
:	:
LN1	ENk

is transformed to a block (BN1 identical EN1) :

BN1	EN1
:	EN2
:	EN3
:	:
:	:
BN1	ENk

If EN2 also belongs to T_x a second block is formed and stored :

BN2	EN2
:	EN3
:	:
:	:
:	:
BN2	ENk

and so forth until there is no T_x left in LB. When the index has been read to the end ('STOP' in flowchart) the stored data are sorted with BN being the first sorting key and EN the second.

Within each block of identical BN the EN of T_x now occur as often as there are significant words in the line which is represented by EN. Line numbers not belonging to T_x occur as often as they have identical (or synonym) words with the line of T_x.

For example :

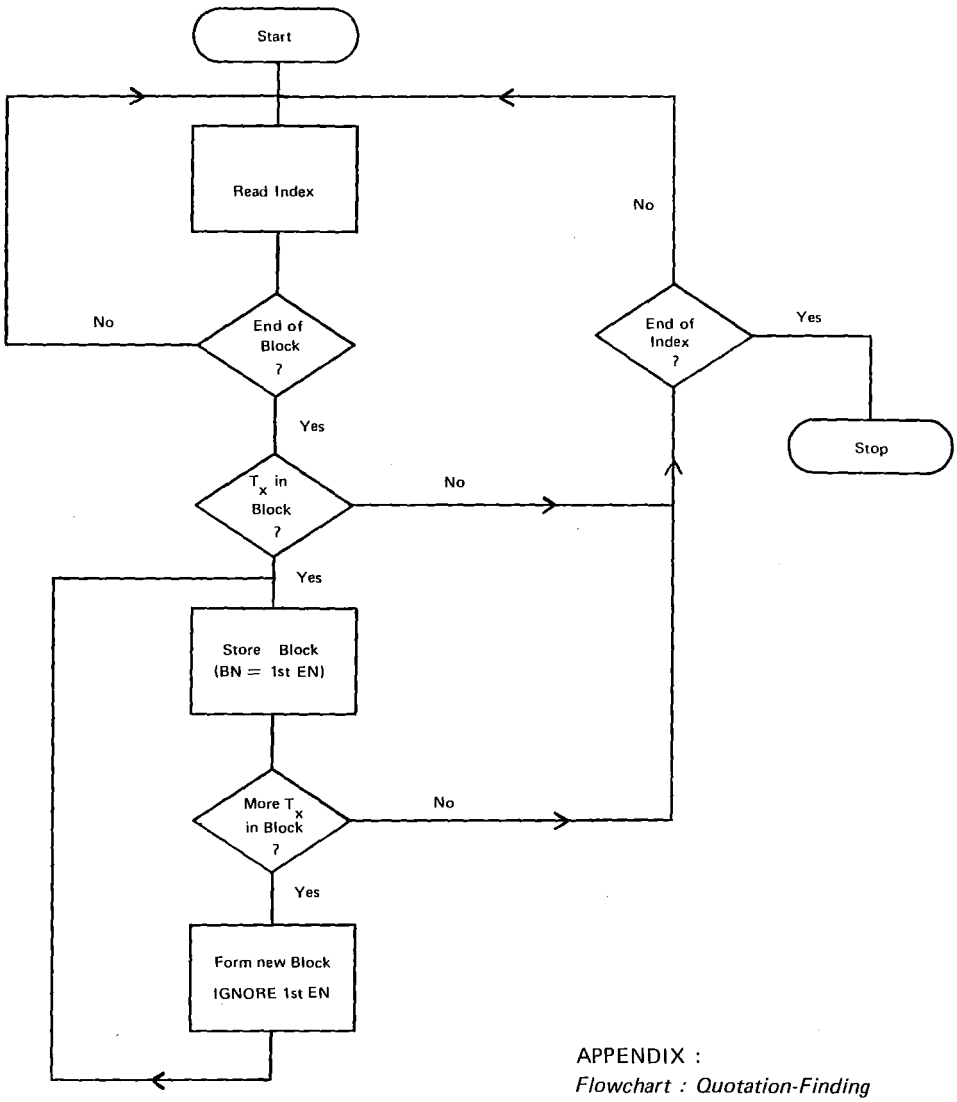
BN10	EN10
:	EN10
:	EN10
:	EN10
:	EN152
:	EN161
:	EN161
:	EN161
BN10	EN487

Three of the four significant words of entry 10 coincide with three words of entry 161.

By means of direct or random access the corresponding lines in T_x and $T_{1...n}$ are looked up and printed. Now the philologist can decide which type of similarity has been found and whether the similar passages are of any interest to him (4).

Sem. f. Engl. Phil.
34 Göttingen

Frank-G. BERGHAUS



APPENDIX :
Flowchart : Quotation-Finding

NOTES

(1) I am indebted especially to Prof. Dr. J. MAU, the chairman of the branch, for his many useful suggestions for improving the efficiency of the computer programmes.

(2) The text is based on the edition of Frederic Madden, *Layamon's Brut or Chronicle of Britain,* 3 vols, London 1847, which has been brought into machine-readable form by Herrn L. SCHOLL. An exhaustive concordance of similar passages in Old English poetry is in progress and will be published, I hope, in the near future.

(3) A single entry in the file may be a verse of poetry, a sentence of prose, or even any mechanically divided sequence of text. In the test material it was convenient to take verses. T_x was simulated by taking a section of the *Brut* containing the first 2000 lines, $T_{1...n}$ was the rest of the text.

(4) All programmes for the investigation are available in Göttingen and can be acquired on request.