WILHELM OTT

COMPUTERS AND CRITICAL EDITIONS

In the text-based disciplines, reliable editions of literary texts are the basis for almost any kind of scholarly research. It is no wonder, therefore, that many editing projects are being undertaken and that those projects range among the most expensive undertakings[1] in the humanities.

The possibility of mechanizing single steps necessary in the course of preparing a critical edition was for the same reason being discussed even before computers capable of handling textual data were available. The Hinman Collator, an opto-mechanical device developped in the 40s by Charlton Hinman for the internal collation of more than 80 volumes of the First Folio Edition of Shakespeare's works, is perhaps the best known example of such mechanical aids. Since then, computers have revolutionized not only the sciences, but with some delay also research in the humanities.

It can not be the purpose of this paper to give a detailed history of computer applications to textual editing[2]. I prefer to concentrate on the methodological and technical aspects of computer applications to the preparation of critical editions and to do this against the background of

[1] For German texts dating from the 8th to the 16th century alone, *Germanistik*, 31 (1990) 234-242 lists 335 current edition projects.

[2] For an overview, see the chapter *Textual Criticism*, in S. HOCKEY, *A Guide to Computer Applications in the Humanities*, London, 1980, p. 144-167, and *Textual Editing with a Computer*, in R.L. OAKMAN, *Computer Methods for Literary Research*, Columbia, 1980, p. 113-138. A more systematic approach is offered by P.L. SHILLINGSBURG in his book *Scholarly Editing in the Computer Age*, Athens and London, 1986. Additional information may be found in the reports on various conferences, above all in : *La pratique des ordinateurs dans la critique des textes*. Colloques Internationaux du Centre National de la Recherche Scientifique, No. 579, Paris, 1979. For the first two decades of computer-aided textual criticism, see also my *Bibliographie : EDV im Editionswesen*, in *Sprache und Datenvarbeitung*, 4, 1980, p. 179-184 which covers the period until March 1983 (sic).

the experiences we made at the University of Tübingen and of the tools we have developed.

It is a truism that every edition has its own special problems, depending on the genre of the text and on the time in which it was written. When editing an ancient text, where the oldest surviving manuscripts are younger, perhaps by several centuries, than the lost autograph of the author, then other methods of editing are necessary than for the edition of a text of the nineteenth or twentieth century, where the first printed edition(s) and the author's manuscripts are both available. Despite these differences, there are many technical problems which are common to every editorial project.

Since 1965, when Helmut Praschek published *Die Technifizierung der Edition – Möglichkeiten und Grenzen*[3], it has become common in discussions like this one to break up the editing process itself into several successive and logically distinct steps and to consider the ways in which the computer may be able to help at each stage[4] :

1. Collection and preliminary listing of the witnesses (manuscripts, early quotations ; for more modern texts : drafts, autographs, early editions)
2. Collation of the texts to discover variants
3. Evaluation of the results of collation (genealogy ; final listing of witnesses to be considered for the edition ; selection of variants to be included in the apparatus)
4. Constitution of the copy text
5. Compilation of the critical and historical apparatuses
6. Preparation of indexes
7. Preparation of the printer's copy from the results of 4-6
8. Publication of text, apparatuses, indexes.

[3] In H. KREUZER and G. GUNZENHÄUSER (eds.), *Mathematik und Dichtung. Versuche zur Frage einer exakten Literaturwissenschaft*, München, 1965, p. 123-142.

[4] The following distinction of 8 steps is taken from W. OTT, H.W. GABLER, P. SAPPLER, *EDV-Fibel für Editoren*, Stuttgart–Tübingen, 1982.

For Praschek, these steps have to be preceded by the definition and elaboration of the editorial principles valid for the current project. Laying down, first of all, the principles and requirements of the work is, of course, necessary not only for editions, but for all kinds of scholarly work. The fact that it may be necessary to refine or revise them during the progress of the work is no objection to this basic rule — which sometimes seems to be forgotten when a project relies essentially on computer applications. The starting point should not be what the computer can do easily, but what is required by the discipline. The computer can supply neither the relevant questions nor the methods ; nor can it take over the responsibility for the results. Professional scholarship in the humanities profits most from the computer when it takes it seriously in its role as a tool and does not misuse it as a problem-supplying and problem-solving device.

Fortunately enough, in the preparation of a critical edition, there are a lot of technical tasks and problems to be carried out which can be substantially supported by approved computer methods and existing software.

I. — THE LAST STEP : PUBLISHING THE EDITION

The publication of text, apparatuses and indexes is the last of the above mentioned steps in the preparation of a critical edition. In the past, it used to be at the same time one of the most laborious ones.

Typesetting errors, unavoidable with traditional typesetting, would be destructive to a critical edition, whose main purpose is to present the exact wording of the text as intended by the author, or as existing at a given time, and to provide reliable documentation of the variants of its transmission or of the different forms the text showed during its genesis. The worth and usefulness of a critical edition depend entirely on the reliability and correctness of the text and apparatus.

A. Computer typesetting

Technological development in the past 25 years has made it possible for this last step in the preparation of an edition to lose much of its drudgery. Since the middle of the 60s, typesetting can be done auto-

matically from a « manuscript » recorded by computer. In 1970, at the first ALLC Conference in Cambridge, Harold Love spoke of a « revolutionary element », of an area where the computer was not only « streamlining old methods », but in which the editing process itself was beginning to « show fundamental modifications », when he discussed the possibility of « the text being assembled on magnetic tape and set by the computer itself »[5]. In 1986, Shillingsburg[6] repeats this view : « at its best, it has the potential to revolutionize scholarly editing and publishing ».

But also possible disadvantages of computer typesetting were seen by both authors. According to Love, « the only drawback is that one would not in the present state of printing technology end up with quite so handsome a piece of book-production » (p. 56) — a fear which has become reality to an extent which he perhaps did not foresee. The reason today is, however, not the current state of printing technology, but the « do-it-yourself »-attitude of hobby homeworkers among humanities scholars (supported by the relevant advertising of the desk-top-publishing industry) who believe that they can do not only without professional typesetting equipment but also without any typographic training and experience.

Beyond the absolute requirement that a critical edition must not contain misprints, typesetting critical editions poses technical problems which other scholarly publications do not have. And since the setting of critical editions is normally not the daily business of composing firms, this may still lead to problems.

B. The pagination problem

When typesetting critical editions, one of these problems is the relatively complex pagination and the setting of the apparatuses, of which often more than one is found at the bottom of the page — sometimes in addition to normal footnotes. Unlike normal footnotes, appa-

[5] H.H. LOVE, *The Computer and Literary Editing : Achievements and Prospects*, in R.A. WISBEY (ed.), *The Computer in Literary and Linguistic Research*, Cambridge, 1971, p. 47-56 ; quotation from p. 55.

[6] In his book quoted in footnote 2, p. 138.

ratuses cannot be broken into lines before pagination ; they are traditionally printed as blocks of text with justified right margins, leaving a bit more blank space between the single notes within the line.

A further peculiarity of critical apparatuses for prose texts is that the single annotations are keyed to the relevant text words not by reference numbers printed in the text and repeated at the bottom of the page, but by counting the lines on every page and prefixing the critical annotation with the number of the line of text in which the annotated text word is found. This means that the content of the single lines itself depends on the pagination (which itself depends on the number of critical notes necessary for the single pages).

For critical editions, the advantage of computer composition depends to a large extent on the question of whether this relatively complex pagination can be included in the automatic procedure. Therefore, the typesetting program which is part of TUSTEP, the « TUebingen System of TExt processing Programs », has as a standard feature automatic pagination, without human intervention at a terminal, of texts containing not only footnotes, but in addition up to 9 different apparatuses at the bottom of the page[7]. Of course, the text must show a more extensive markup for a program which provides automatic pagination than for galley production or interactive pagination.

The advantage of automatic pagination lies not only or even primarily in saving some time and money. Consistently applied, it may really revolutionize the method of organizing an editing project, as the following example shows.

When vol. III of series VI of the complete works of Gottfried Wilhelm Leibniz[8] had been finished, which contains in chronological order more than 100 documents of very different length, the editors had found that with this procedure they could really, without catastrophic

[7] Cf. W. OTT, *A Text Processing System for the Preparation of Critical Editions*, in *Computers and the Humanities*, 13, 1979, p. 29-35 ; a more complete description is contained in the following papers : *Automatisierung von Seitenumbruch und Register-Erstellung beim Satz wissenschaftlicher Werke*, in H. HEILMANN (ed.), *4. Jahrbuch der EDV* Stuttgart–Wiesbaden, 1975, p. 123-143 ; *Integrierte Satzherstellung für wissenschaftliche Werke*, in *Der Druckspiegel*, 31 (1976), p. 35-44.

[8] G.W. LEIBNIZ, *Philosophische Schriften*. Hrsg. v. d. Leibniz-Forschungsstelle der Universität Münster, 3. Band : 1672-1676, Berlin, 1980.

financial consequences, change not only single pages, but even the whole arrangement of the single pieces contained in this volume. They had done this only a few weeks before the final printing of the volume, after having checked everything in page proofs on the basis of some preliminary arrangement. As a consequence of this experience, the editor decided, for future volumes, not to wait 5 years until all the material needed for a full volume was prepared, dated and commented, but to print a small volume of about 200 pages every year « ad usum collegialem ». This « Vorausedition » shows text and apparatuses in the same typographical form as provided for the final edition, only with a different page width and depth and in smaller type ; it not only serves as a report on the progress of the work, but also makes available the hitherto unpublished materials to colleagues all over the world. In this way, the critical remarks and suggestions of other scholars interested in this work are stimulated and can be taken into account for the selection of the pieces to be contained in the next volume, for their chronological arrangement and for final corrections, before the volume in question is published. To date, 8 volumes of the « Vorausedition » have been printed. As Schepers says in the preface to the 8th volume (Oct. 1989), a further essential advantage of this procedure lies in the strict time schedule for every step to be carried out periodically every year[9].

C. Special characters and fonts

In the past, a further problem for composing critical editions may have been the availability of characters beyond those used in the everyday business of composing and printing firms. In the age of electronic composition the availability of special characters should theoretically no longer present any problems. The fast spread of hardware-independent page-description languages like POSTSCRIPT, for which a great variety of font libraries are available on the market, will further remove the obstacles which one is still likely to meet in today's practical work. At present, however, in order to get professional quality, it remains good advice to rely on firms or institutions who are accustomed to respective tasks.

[9] Cf. also H. SCHEPERS, *EDV-Erfahrungen einer Edition*, in *Philosophisches Jahrbuch*, 88, 1981, p. 159-164.

II. — TOOLS FOR PREPARING THE « MANUSCRIPT » OF THE EDITION

If word processing and automated typesetting were all that computers can contribute to scholarly editing, this would not deserve an extra paper in this conference. These functions are not at all special to critical editing or to scholarly work in general. In editing projects, electronic data processing can — and should — not only support typing, correcting and typesetting, but also the scholarly part of editorial work itself which corresponds to steps 2-6 in the above schema of editorial tasks.

Automatically collecting the variant readings or, for modern texts, the differences in subsequent stages during the development of a text ; safely administrating and evaluating them — in editions of antique texts, including questions of stemmatic relationship ; gaining a complete and exact overview of the usage of words and spellings, of prepositions and other syntactic features, of orthographic peculiarities of single manuscripts ; getting persistent control over the editor's own decisions and the alterations made in the course of the editing process to the text and apparatus ; exploiting the text with indexes, comments, references to sources and cross-references to similar passages : those functions which are beyond mere word processing may add new qualities to editorial work by replacing « impressionism and the random instance »[10] by a systematic approach based on the complete material.

Not all of these steps are of equal weight for every edition ; their importance depends on many factors. When the autograph of the text to be edited is available, then methods for determining the genealogy of the witnesses, for reconstructing an archetype and for constituting an « original » text, which are essential for editions of classical Greek and Latin texts, are without meaning. However, the tools which are used to determine the variants present in the different versions of a text as a result of its tradition by manual transcription through centuries, may be used also for detecting and recording the differences in a work which has developed to its final stage through different phases of corrections, additions and revisions.

[10] LOVE 1971 (cf. above footnote 5), p. 54.

It cannot be the task of this short paper to make up for a comprehensive introduction to the usage of the single tools, or to examine in detail how the methodology to be applied in the different steps can be supported by the computer. I will therefore, after some hints to other, more « standard » procedures and programs, concentrate on one of these tools only, namely automatic collation, which is especially relevant for the preparation of critical editions.

A. Indexes, word lists, concordances

Apart from word processing software, which allows the entering, correcting, editing, formatting and printing of texts, the most commonly available software for handling textual data are indexing programs, which decompose a text into its components (e.g. single word forms, or single paragraphs), record the exact location of these constituents (e.g. adding codes for the source text and references like page- and line-numbers) and sort them according to rules to be defined by the user. This offers the possibility of having available — in any stage of the editorial work and for every state of elaboration of the text — concordances and word lists arranged alphabetically, or according to word endings, or to word frequency, and to consult them when working on the text and the corresponding annotations. Updated versions of these lists are easily available when required by the progress of the work.

Indexing programs are therefore not only useful when preparing the indexes to be printed with the edition (step 6). The invaluable contribution which the availability of these lists can make to the editorial work especially for step 4, the constitution of the text of the edition, has been stressed ever since the role of the computer for the preparation of editions has been discussed[11]. In fact, we have found that editions which have been prepared in the conventional way, without computers, have revised their text just before final printing, when, due to the fact that the text had been transcribed on a computer for photocomposition, a KWIC (keyword in context) concordance was available at almost no cost, except for computer time and printer output.

[11] Cf. e.g. PRASCHEK 1965 (see above footnote 3), p. 134f.

B. Numerical computation, statistical evaluation

The functions which gave the computer its name, involving calculation with numerical values, take part also in the generation of word indexes, when absolute and relative frequencies are to be given for each entry. For the preparation of an edition, especially of ancient texts, numerical calculation plays a further role. If more than two manuscripts are available which offer different versions of the text, it may be useful to start from a list of variants and to calculate the relative proximity or distance of the different witnesses on the basis of the number of variant readings they have in common.

The application of methods of mathematical statistics, including cluster analysis and other specially designed methods for determining stemmatological relations, has been for more than two decades the subject of a large number of publications, for which Jacques Froger's book *La critique des textes et son automatisation*[12] is one of the early and prominent examples. The pertinent discussions seem however to have ceased more or less at the beginning of the 80s.

Indeed, no great progress seems to have been achieved in the last 20 years, at least as far as the application of numerical methods to questions of manuscript affiliation and stemmatology is concerned. The initial enthusiasm has given way to a certain scepticism as expressed by R.H. Pierce in 1987 : « The use of computers and spreadsheets makes it possible to include a very great number of ms. and readings in a data set, e.g. those representing an entire Gospel. If this is done, any hope of assessing the material directly will be gone ; and investigators will be obliged to place their trust in the methods they employ. This they ought to be unwilling to do unless they have already satisfied themselves about their reliability by testing them on material which they are satisfied they can evaluate without their aid »[13]. Also P. M. W. Robinson, who at a first glance seems to be more optimistic regarding statistical methods,

[12] J. FROGER, *La critique des textes et son automatisation*, Paris, 1968.

[13] R. HOLTON PIERCE, *Multivariate Numerical Techniques Applied to the Study of Manuscript Tradition*, in B. FIDJESTØL et al. (eds.), *Tekstkritisk Teori og Praksis. Nordisk symposium i tekstkritikk. Godøysund 19.-22. mai 1987*, Oslo, 1988, p. 24-45.

prefers to rely on putting his material into a data base, using it for « speedy development and testing of hypotheses concerning manuscript relations »[14].

The development of the last 20-30 years shows that serious attempts to use the computer to help with the problems of textual editing are very far from automatic textual criticism or textual editing in its philological phase. The computer can however supply a firm basis for text-critical work.

C. Automatic collation

Apart from the attempts to solve stemmatological problems with the aid of the computer, collecting and recording variant readings was the second field for which computational help was sought early on, especially when preparing a variorum edition of ancient texts.

Collating is a basic requirement for all editions for which more than one « witness » containing the text or one of its preliminary stages is to be evaluated. On the result of this first step all further editorial steps are founded ; on its exactness depends the reliability of the whole edition. Therefore, the increase in exactness, completeness, and reliability which can be achieved by computer application is far more important for this step than the possible savings in time and money.

But, beyond collecting and recording variant readings, automatic collation has other fields of application, two of which are especially relevant for textual editing :

a) it can be an economic replacement for proofreading and manually correcting texts in any case where absolute exactness of transcription is essential (as, e.g., when entering the text to which all the other witnesses are to be collated manually or automatically) ;

b) it can serve as an accurate, fast and reliable tool for checking (and for keeping track of) the results of any manual or automatic operation performed on the text.

[14] P.M.W. ROBINSON, *The Collation and Textual Criticism of Icelandic Manuscripts. (1) : Collation ; (2) : Textual Criticism*, in *Literary and Linguistic Computing*, Vol. 4 (1989) no. 2, 99-105 ; Vol. 4 (1989) no. 3, 174-181.

The potentiality of automatic collation, its usefulness and importance for scholarly work in the humanities in general and in textual editing in particular is seriously underestimated not only by many users, but also by the designers of respective programs. It therefore seems to be worthwhile to make some more words about it and its application.

« Rationalizing » and « increased reliability of the results » are often only two sides of the same coin ; in many cases, the rationalization of certain steps by computer application is the pre-condition for their being carried out at all. In order to be suitable as a tool consistently applied in everyday work, a computer program must therefore meet some basic requirements. For automatic collation, these requirements can be summarized as follows[15] :

1. Collation must be really automatic ; it must not rely on manual intervention.
2. For the task of collecting variant readings, collation must be possible for an indefinite number of versions.
3. The results of collation must not only be made visible for inspection by the scholar's eye, but must also be available in a form which allows their further processing by other programs.

1. What does « automatic collation » mean ?

1.1. Re-alignment of the versions after encountering variant readings

While in the first publications on automatic collation a major role was assigned to the problem of automatic realignment of the texts after

[15] A more explicit examination of these requirements I gave in the following papers : *The Output of Collation Programs*, in D.E. AGER et al. (eds.), *Advances in Computer-aided Literary and Linguistic Research*, Birmingham, 1979, p. 41-51 ; *Transcription errors, variant readings, scholarly emendations : software tools to master them*, in : *Association Internationale Bible et Informatique, Actes du Second Colloque International « Bible et Informatique : méthodes, outils, résultats », Jerusalem, 9-13 juin 1988*, Paris–Genève, 1989, p. 419-434.

variants have been encountered[16], many PC-based collation programs advertised nowadays for collecting variants rely on manual intervention[17].

An approach like this prevents many possible applications of computer-aided collation : if it is not really automatic[18], it will never be used for consistently checking the results of the single steps of interactive or rule-based work on a text, or for repeating the collation if for some reason the collation basis changes during the editing process.

[16] Cf. the detailed report in R.L. OAKMAN, *Textual Editing and the Computer. Review Essay*, in *Costerus : Essays in English and American Language and Literature*, New Series, Vol. IV, Amsterdam, 1975, p. 79-106. This problem has been taken up again by ROBINSON 1989 (see above footnote 14) who describes in detail his solution to it.

[17] So the program *COLLATIO*, which was first described by H.W. GABLER and W. KREITMAIR (in B. GREGOR and M. KRIFKA (eds), *Computerfibel für die Geisteswissenschaften*, München, 1986, p. 203-211), or the program *URICA !* (*User Response Interactive Collation Assistant*) described by R.L. CANNON and R.L. OAKMAN (*Interactive Collation on a Microcomputer : the URICA ! Approach*, in *Computers and the Humanities*, 23,6, p. 469-472) : « The system performs a word-for-word comparison, locking the keyboard as soon as a word in the master text no longer matches a typed word in the second text. The user then enters the variant text at the keyboard until a point of match is reestablished » (p. 470 ; or, when working on two files stored on disk, « when a mismatch is found, the machine beeps and pauses for the user to determine why the mismatch occurred », p. 471). This procedure responds to what Praschek proposed in 1965 (p. 131, see above footnote 3), and is a modern reanimation of the punched card verifier used in the 60s. Both programs add the possibility of recording the variants which have been classified by the user as « insertions », « deletions » or « changes » in a file of variants, together with their reference.

[18] Of course, there are practical limits to really fully automatic re-alignment ; there will always be a compromise between this requirement and the expense in computing time necessary to achieve it. When at a single location an insertion or a deletion extends over more than about two typewritten pages, even the program VERGLEICHE contained in TUSTEP requires help by the user. The reason is that the longer a mismatch is, the more words (which may show « minor » differences) must be recognized as « basically identical » or corresponding to each other before the texts can be realigned. The help required by the program can, however, be given in advance, in the form of pairs of references indicating for each of the two text versions the passages which the program is required to collate. Of course, the program records which passages have been included at all in the collation.

1.2. Delimiting variant readings

Apart from the problem of automatic realignment (which the interactive programs avoid), a further problem arises when recording the variants : the subdivision of the mismatching pairs of strings or word sequences into meaningful pairs of variant readings.

No program can provide subdivisions which are meaningful in every instance. Therefore, it would have been a chance for programs relying on interactive work to provide means for the user to group adjacent differing words into meaningful units of variant readings. This could have added a quality which programs working automatically cannot provide. But even this chance has been missed : in URICA !, which relies on human intervention in every case of mismatch, « in the case of replacements, it is necessary to move through each text until the position of rematch is found »[19].

This is less than can be achieved by a fully automatic approach which tries, as TUSTEP does, to group the pairs of variant readings as « atomistically » as possible, on a word-by-word basis[20]. As long as the program recognizes two varying words as being « principally » the same, they are treated as pairs of variant readings, no matter whether the following pair of words also shows differences or not.

Of course, the result of this procedure often does not make sense. So, the grouping of the pair of variant readings « Der Knabe] Des Knaben » would be far more adequate than the « atomistic » approach of subdividing them into two pairs of variant readings : « Der] Des » and « Knabe] Knaben ». It does not, however, seem feasible to provide language-independent algorithms which could identify meaningful larger units of sense. Compared to possible unsatisfying algorithms, the « atomistic » approach at least has the advantage of regularity. Further scholarly use of the automatically collected material,

[19] CANNON and OAKMAN, 1989 (see footnote 17), p. 471.

[20] Cf. ROBINSON, 1989 (see footnote 14), p. 102 : « The fewer cases of variants involving more than one word found by COLLATE the more useful its results ». — Robinson apparently did not know the TUSTEP program VERGLEICHE when he wrote that « it was early clear that it is just in this that the methods of earlier machine collations fail ».

e.g. for creating entries for a critical apparatus, will in any case require intelligent philological revision. For this revision, the breaking down of variants, as far as possible, to the level of single words seems to be a good starting point.

2. Collating many versions of a text

When more than two versions of a text must be collated, this can be achieved in different ways : either, the program itself allows a certain number of input files to be read in parallel, or provisions are made for the output of successive pair-wise collations of a text to the same « basic » text to be cumulated according to the needs of a project. It is this solution which we have chosen in TUSTEP, thus avoiding any limits which may be imposed by the availability of central computer memory or the number of files which an operating system allows to be open at the same time. What seems, at first glance, to be a limitation, namely that only two files are collated at a time, is in reality the condition that allows there to be no technical limits regarding the number of collations to be performed : by subsequent sort and merge runs, one can cumulate as many « variant reading » files as required by a project. And for the case that, in a later stage of the work, it becomes necessary to include a further text in the collation, only this one collation must be made additionally, and its results merged with the previous ones.

3. Recording the variant readings for further processing

When recording the variants for further processing, each entry must show at least the following elements :

a. complete indication of the location where the variant reading occurs (e.g. page-, line- and word-number or pair of such for multi-word variants)

b. a sign for the kind of variant (omission, addition, replacement)

c. for additions or replacements, the wording of the variant reading.

These three elements are sufficient for a complete description of the differences contained in two texts. For the purpose of collecting variant readings for a variorum edition, a fourth element must be added which

enables the identification of the source of a variant after merging the different variant files resulting from the single pair-wise collations :

d. code identifying the text bearing the variant reading (manuscript code)

The four elements mentioned hitherto allow the sorting and merging of variant readings in a way similar to that required by a critical apparatus : location of variant reading, sign for kind of variant (omission, addition, replacement), wording of variant, and code for the witnesses bearing this variant. Also all the technical information necessary for preparing a genetic apparatus, which shows in parallel lines the different stages of the development of the text, is contained in these four elements.

In a critical apparatus, normally also the wording of the edited text is repeated, followed by a square bracket. Though it can be retrieved easily by program through the location element contained in each entry in the variant file, it is useful to add it as a fifth element to each entry :

e. wording of the text of the collation basis.

This fifth element is required if the recorded variants are the basis for further research on their character and on the character of the texts in which they are contained. Variant spellings which are typical for the single witnesses may easily be found after a sort run in which the first sorting criterion is not the location of a variant reading, but this fifth element which contains the wording of the text used as collation basis ; the second sorting criterion will be the wording of the variant reading (element c) ; the manuscript code (element d) which will serve as the reference in the resulting list must be used as the third criterion[21].

[21] For efficiently performing those sorts, TUSTEP adds further fields containing those elements in the form of sort keys ; they need not be discussed here. For more details see OTT, 1979 (cf. above footnote 15). Some details have been changed in the meantime ; they are described completely in the current edition of the TUSTEP reference manual (TUSTEP. Tübinger System von Textverarbeitungsprogrammen. Tübingen, Zentrum für Datenverarbeitung 1989) p. 441-457 and, for additional sort keys, p. 417-430.

D. Integrating the tasks

Recording differences in the way just described makes sense only when other programs are available which can handle the information recorded in these elements. This means that the collation program must be integrated into a system of related programs. Therefore, when developing TUSTEP, the « TUebingen System of TExt processing Programs », we have tried to build such a system, covering every step from the first data entry over all stages of textual analysis until the final typesetting[22].

The flexibility of this system is achieved by two features :

a) for every module, parameters can be provided by the user in order to adapt it to the requirements of his project (e.g. by defining a sorting sequence for greek texts), and

b) the output of any one program may serve as input for any other module.

It is this integration which makes consistent computer application to the scholarly tasks of critical editing easy (and therefore feasible). Though this is valid also for all the other programs, I will for the rest of this paper discuss the role of this integration for the collation program only.

Before the collation of the different versions of a text can be started in order to collect variant readings, the texts themselves must be entered and carefully corrected. Else, the transcription errors would be recorded among the variants. As indicated biefly above, automatic collation can be an economic replacement of conventional proofreading and interactively correcting a text[23]. When applied for this purpose only, two

[22] Under the title *A Text Processing System for the Preparation of Critical Editions*, I described the basic ideas of this system in a paper given at the 4th ALLC Conference (Oxford 1976 ; published in CHum 13 (1979) p. 29-35, cf. above footnote 7). In the meantime, the system has been streamlined and the single programs have got different names.

[23] We had our first experiences with this method of double transcription and automatic collation for correcting a text when the text of the Vulgate was transcribed for the preparation of the *Novae Concordantiae Bibliorum Sacrorum iuxta Vulgatam Versionem Critice Editam* which B. FISCHER has published in 5 volumes (Stuttgart,

transcriptions of the same text are collated by program, the differences are both listed and recorded in a file. These records are later interpreted as correcting instructions. This is possible because the elements a, b and c of these records described above (C 3), which allow the complete description of the variants contained in the second one of the two compared texts, are also the elements required for an instruction to perform appropriate changes to a text : at the location given (element a), delete, add, or replace (element b) one or more words contained in the text file by the word(s) contained in this record (element c). With these results of the collation, one needs only to check the printed list of differences. In every instance, where the second transcription is wrong, the relevant entry from the « variant file » is deleted, using an editor ; the remaining « variants » are those where the first transcription was wrong and the second one has the right wording. Then the batch correcting program KAUSFUEHRE is run, taking as input both the file containing the first transcription and the file containing the remaining (corrected) variants of the second transcription ; the program interprets the variants as correcting instructions to the text of the first transcription and writes the corrected text to an output file.

As soon as the texts to be collated have been corrected, automatic collation can be used to collect the variants contained in the texts.

It has often been argued that for many purposes this procedure — computer collation from machine-readable versions of a text — is too expensive in time and money, because the texts must be transcribed in full length (or scanned by optical character recognition) and corrected before they can be collated.

For cases where this is not feasible, the modular approach described above makes possible the mixing of automatic and conventional collation, since it separates the recording of the variant readings from any evaluation of these variants. To integrate the results of conventional collation, one needs only transcribe these results in the syntax described above, which is required also for correting instructions to be carried out by the batch correcting program. It consists of the location (page-, line-

1977). See my report *Transcription and Correction of Texts on Paper Tape. Experiences in preparing the Latin Bible text for the computer* in *Revue (LASLA)*, 1970, Nr. 2, p. 51-66. On the economics of this procedure, see also OTT, 1989 (cf. footnote 15), p. 421.

and word-number) of the word where the variant is found (or where it begins and ends, in the case of variants containing more than one word), a symbol (« + », « – » or « = ») for « addition », « omission », « replacement », and (except for omission), the wording of the variant. The code of the manuscript can be added by program after the whole file of variants has been transcribed. This format allows the merging of variants collected manually with variants collected by automatic collation. As the format in which the variants are recorded is identical to the format of correcting instructions, it is even possible to reconstruct the whole text of the version containing those variants, which may then be collated to the other versions if required. This reconstruction will be as accurate as the manual collation itself and the transcription of its results have been ; it will show up everything except (if not recorded) the page and line divisions of the version collated manually.

The evaluation of the collation results is done with the help of the other programs of the system : sort keys are added before sorting the variants, and groupings are carried out after sort by the same programs which are used for preparing (hierarchically structured) indexes, including, if required, frequency counts and other operations. The selection of variants to be used for different purposes (such as the compilation of the critical apparatus) can be carried out by the same programs which allow manual or automatic selection for other purposes.

Collecting variants is only one application of automatic collation. The work on the edition text, on the apparatuses, on the preface, the comments and the indexes will generate many different versions of all these parts during the process of editing. Some of the relevant work will be carried out interactively, using an editor to insert, exchange or delete single characters or whole paragraphs. Other procedures are carried out automatically by search and replace operations or more sophisticated replacement algorithms. In every case, the editor must rely on the exact execution of the manual operations and on the adequate formulation of the replacement rules. In both cases it is advisable to carefully check the results[24].

[24] Checking by automatic collation the changes made to a text is so convenient that we recommend it to our users also for « normal » corrections. Before they correct texts interactively with the help of a text editor, or make students carry out those corrections, they are advised to always make a copy of the text before starting

Automatic collation can be an invaluable tool also for this purpose, giving prefect control over every stage of manual or automatic work on a text — presupposing that it is integrated into the whole system of programs.

The possibility of automatically checking the results of all operations may lead to otherwise unattainable reliability of scholarly work, which is especially important in the preparation of critical editions. A short but impressive report on the importance of automatic collation, of the automatic transfer of collation results into a critical text and perfect control it offers over every alteration made to the text, may be found in the afterword of Gabler's edition of James Joyce's Ulysses[25].

This reliability achieved by persistent application of automatic collation is supplemented by indexing and sorting procedures, and maintained until the final publication using computer typesetting. It depends however to a large extent on the integration of the respective tools into one system of compatible programs, in which the output from any one stage of computer work is usable for the next stage, and which covers all the steps necessary for editorial work, starting from the first data input until final typesetting.

Zentrum für Datenverarbeitung, Universität Tübingen

their corrections. So, after correction, they can run an automatic collation of the uncorrected and the corrected file and can check in the resulting printer listings if all and only those corrections have been carried out which had been intended. Of course, in additon to working automatically, without human intervention, the program must be fast enough, and, of course, not rely on identical line divisions of the two files to be collated : it must be possible to reformat a text between two collations.

[25] J. JOYCE, *Ulysses. A Critical and Synoptic Edition.* Prepared by H.W. GABLER with W. STEPPE and C. MELCHIOR, New York and London, 1984.