TRANSCRIPTION AND CORRECTION OF TEXTS ON PAPER TAPE
EXPERIENCES IN PREPARING THE LATIN BIBLE TEXT FOR THE COMPUTER

One of the great obstacles to a more common use of electronic data processing in the philological field concerns the preparation of the input data. Normally, the texts to be examined or analyzed are not yet available in machine readable format. So, first of all, they must be transferred to a mechanical or magnetic data carrier by transcribing them, letter by letter and line by line, on appropriate equipment, e.g. a key punch or a typewriter equipped with a paper tape punch or a magnetic tape drive. Therefore, in addition to the task of devising suitable computer programs for the analysis of literary texts, the preliminary difficulty of an accurate transcription must be solved if data processing is to become practicable in this field.

There are two main problems connected with this preparatory work :

1) the problem of avoiding typing errors, of detecting errors which nevertheless are made, and of correcting the detected errors;
2) the problem of reducing the time and cumbersomeness of the transcription to a minimum.

It is clear that the latter point is of importance not only for the happiness of the typist (which here may be left out of consideration), but also for the correctness and the costs of the process of transcription.

The amount of comfort and speed which may be attained in transcription depends on reducing the number of strokes necessary for copying the text to a minimum. In transcribing a text, at least one stroke is required for each character of the original data (the space between words counting as one character) and for each record of the data carrier (e.g. pressing "carriage return" on the electric typewriter).

The absolute minimum of strokes per record can be reached only when it is possible to guarantee correct sequencing of the single records in the data file without inserting by hand additional labels containing, e.g., the current number of the card. When transcribing voluminous texts on punched cards it seems to be important to keep open the possibility of a later sequence checking : there are too many possibilities in the course of transcription, verification, correction, and transport to the computer, that single records (cards) get out of order, that cards get lost, or that additional cards are inserted. On punched cards, therefore, a field of at least five of the eighty columns must be reserved for a label. This means additional strokes, and additional attention of the typist, and additional possibility of making errors. Only endless data carriers (paper tape or magnetic tape) allow getting by with the minimum of one additional stroke per record, because only with these it is impossible for single records to get out of order. Supplementary numbering by hand is superfluous, and there is the additional advantage that we are not restricted to the eighty characters per

record as on a punched card. Longer records (lines) are possible, and when a record happens to be shorter, this means no waste of the material, which is, in any case, cheaper and less voluminous than punched cards.

The minimum of one stroke per character in the text can be reached only when the number of different characters occurring in the text corresponds to the number of keys on the keyboard of the equipment used. With a key punch (card punch), the most common equipment for manual data collection, or with equipment having equivalent keyboards, this is only the case for texts which contain no more than thirty-five different (including blank) characters. This keyboard does not even allow for the transcription of the twenty-six alphabetic and the ten numeric characters by pressed simultaneously with the other keys to utilize the full sed of fourty-eight or sixty-four characters available on the keypunch.
pressed simultaneously with the other keys to utilize the full set pressing a single key. A second key, the shift key, must be of fourty-eight or sixty-four characters available on the keypunch.

Sixty-four characters are, however, not enough when upper and lower case characters are to be distinguished. Using a keypunch, this distinction can be made only by prefixing the upper case characters by a control character, i.e. by pressing two different keys, *one after the other* (possibly the shift key must be pressed in addition).

On the keyboard of a typewriter, this distinction can be made by pressing two keys *simultaneously*. This is faster and more convenient to typists than pressing two keys, one after the other. Since the keyboard of an electric typewriter, as it is used in data collecting equipment, has at least fourty-four character keys, it permits a transcription of eighty-eight different characters and the blank space by single strokes, when the act of

pressing two keys simultaneously is viewed as one stroke. Eighty-eight different characters cover the fifty-two lower and upper case letters (fifty-nine on German keyboards, including äÄöÖüÜ and ß), the ten decimal digits, and twenty-six (nineteen on German typewriters) special signs which may be used for punctuation marks, for the most common diacritics, and, if necessary, for additional shift- or font-change characters. Whether a set of eighty-eight different characters is encoded on the data carrier in as many different codes or not, is of no importance for the process of transcription, provided that all the different characters of the key-board are also discernible on the data carrier(1). Apart from the reliability of the equipment, the only thing of importance is that the process of transcribing is as simple and convenient as possible for the typist.

The change of upper and lower case characters is not always as frequent as in German texts, and therefore the advantage of the possibility to transcribe texts on an ordinary typewriter keyboard many not be as great for other languages, where only proper names and new sentences begin with upper case characters. Nevertheless, the greater character set available on a typewriter keyboard is a further reason to employ this equipment instead of a keypunch. This always reduces the number of strokes to be made, even when transcribing ordinary Latin or English texts, which contain almost no diacritics and few upper case characters.

---

(1) Usually, the different characters available on the keyboard produce not as many different codes on the data carrier. For example, in the case of 8-track paper tape, three codes are recorded when a single upper case character is typed : a locking shift-out character, generated when the shift key is pressed, a code representing the character itself, and a shift-in character ("return to lower case"), which is generated when the shift key is released.

In the preparatory works to a concordance of the Vulgate, which we are compiling at Beuron (Vetus-Latina-Institut) and Tübingen (Computing Center of the University) based on the new critical edition by R. Weber OSB (Stuttgart 1969), we had to transcribe a Latin text without punctuation. However, since the critical variants contained in the edition will also appear in the concordance, not only the occurrence of capital letters increased, but many additional characters had to be transcribed. So, in addition to the normal alphabet of the text in roman letters, an italic alphabet, several Gothic and some Greek letters, some special signs (parentheses, brackets, point, colon, semicolon, addition sign, hyphen, asterisk, obelus, tilde), and some supralineal characters had to be coded. As far as the special signs were concerned, there were enough keys free on the typewriter which we could use without additional prefixes apart from the shift key. For the change from one alphabet to another, the use of "escape" (shift) characters was indispensable. Since characters from the Gothic and the Greek alphabet occurred only as single characters, not forming whole words, we chose a non-locking shift character to distinguish them in the transcription. This control character, of course, had to be transformed by program to a locking shift-out character (font change character) and to the corresponding shift-in character after the letter, depending on the automatic typesetting system for which the final computer output is to be prepared. For the italic alphabet, we chose a locking shift-out character, since whole words may appear in italics; but here, too, the pertinent shift-in character was not typed; it was generated by the program according to such rules as : when a right parenthesis occurs, and a preceeding shift-out character for italics was not yet negated, the corresponding shift-in character is inserted before the parenthesis.

Of course, conventions like these are independent of the data carrier and of the type of equipment employed for transcription; they can be realized on any keyboard.

Another advantage of the typewriter is important in this regard. The transcription of an apparatus criticus, where strings of upper and lower case characters of different fonts and special signs are to be typed, results in many additional control characters, sometimes more than one for one character of the text. This demands high concentration of the typist, who will welcome therefore any possibility to verify and, if necessary, to correct what he has just typed. This verification is possible on the typewriter print-out at once : every stroke is registered not only as a code on the data carrier, but also on the sheet of paper in the typewriter, so that the typist may conveniently check every stroke he makes. On a printing card punch, this would be possible only several strokes later and requires more attention, not only because the characters printed on the upper edge of the punched card are not very well legible when the card is still in the punching station, but also because here the shift characters for upper and lower case letters appear as additional characters, whereas on the typewriter printout these characters do not appear, but result in a different grapheme of the subsequent character. This is less confusing and allows for faster verification.

Typing errors may be divided into three groups :

1) errors which are noticed immediately when pressing the wrong key;
2) errors which may be detected after a glance at the typewriter printout, e.g. after typing a complicated character string. These errors

may be detected only some words or even whole lines after typing them;

3) errors which are not detected during the transcription.

For the errors of group 1, the possibility of revising at once what has been typed may be advantageous; for detecting the errors of group 2, this possibility is indispensable. Everyone who has had some experience in typing knows that often, after making a typing error, one is aware of having made the error, but does not yet know what actually is wrong. On the other hand, the problem of getting error-free copies is so serious that no chance of detecting an error should be overlooked. Without the possibility of fast and simple revision of a suspicious line, many such chances would escape. Further, errors detected during the typing stage can be corrected at once; the earlier errors are corrected, the more reliable all subsequent stages of revision and processing will turn out. Finally, the possibility of detecting and correcting errors at once frees the typist from the anxiety of making mistakes. This pressure not only reduces the productive speed, but also, paradoxically, sometimes seems even to increase the rate of errors.

Naturally, the process of correcting detected errors immediately must not be too difficult. An error detected when transcribing a text onto punched cards is usually corrected by first releasing the card from the punching station and moving it to the reading station, and then duplicating the card up to the erroneous column, from where normal typing proceeds. The wrong card must then be taken out of the file.

On a paper tape equipment, errors which are detected immediately, or not

later than about ten strokes after the erroneous stroke, can be corrected by the over-punching facility of this equipment. This correction is made by pressing the backspace key of the typewriter until the erroneous character is reached. By this operation, the paper tape is backspaced an equal number of frames. By now pressing the delete key, the backspaced characters are deleted by being overpunched on the paper tape and converted into erase characters, which consist of code holes in all code positions. These erase characters are ignored when the paper tape is read by the computer.

For errors which are not detected so soon, but which perhaps occurred half a line or more before, this "hardware" facility is not practicable. It would take too long to backspace over more than about ten characters, and additional care would be required when shift characters are among the codes to be deleted, since they do not appear as separate characters on the printout. We therefore introduced an additional "software" convention for the elimination of errors. A special sign which did not occur in the text was defined as an ignore character, which could be typed at the end of a line. When the program by which the paper tape later was read detected this ignore code, the whole line was ignored; when the ignore character was typed as the first character of a line, the preceeding line was ignored. When errors were detected in earlier lines, one could indicate, by typing a number after the ignore character, how many lines (up to three) were to be ignored by the program. These lines then had to be retyped.

These two possibilities to correct errors detected during typing fully compensate for the duplicating facility of the keypunch, as far as it may

be used for error correcting. But what about the correction of errors which are detected only after revision or verification ? And what about the process of verification itself ?

It is commonly known that verification by retranscription on a verifier affords less time and less attention and gives better results than (even repeated) proofreading. But verifiers do not exist for paper tape. Therefore, we devised a verification program which, at least for our project, more than fully compensated for the lack of a paper tape verifier. After the first transcription, the text was transcribed a second time on the same equipment. The two paper tapes which resulted from the two transcriptions were then (after conversion to magnetic tape by a background program) separately updated and checked for convention errors by a program. In this phase, lines marked with the ignore code were deleted. Omitted shift-in-characters were inserted (see above). The full reference for each stich was compiled from the heading of every chapter (e.g. Gn 1,1), which was typed only once for each chapter, and from the number of the verse, which was typed where it stands in the text (this may also be in the middle of a line); this reference was added to each stich of the text. Moreover, in this phase a number of preliminary corrections and checks were made. Double spaces between words were eliminated; when the two shift characters for upper and lower case followed each other with no intervening character, they were eliminated. The checks which were made affected, above all, the most complicated parts of the transcription, namely the *apparatus criticus*, which was inserted into each line after the word to which it refers. The computer compared the symbols representing the manuscripts with a list containing, in the prescribed order, all the

symbols of manuscripts which may occur in the single books of the Bible. Thus, most errors in the *apparatus criticus*, especially wrong or missing shift codes, were detected (in some cases, it was even possible to correct the edition from which the text was transcribed : the edition had not yet gone to press for final printing when the transcription was made). For all these errors a detailed error message was provided on the printer output, containing the text and reference of the line, and a message indicating which rule had been violated by which character. On the magnetic tape, the respective lines were marked by an error code. The output of this program was the preliminary list of errors in each of the copies of the text and two magnetic tapes containing the updated text of the two transcriptions.

In the next phase, the verification was made by comparing these two magnetic tapes. A single copy of the transcription was recorded on a third magnetic tape which then was printed out. On this tape, the stichs and lines (sometimes more than one line belongs to a single stich) were numbered, in order to facilitate later references, e.g. in the correction phase. Where both input tapes coincided, letter for letter, the text of the single lines was recorded only once (including on the right margin of the printer output the reference and the number of the line). When differences between the input tapes were detected, the text of both tapes was recorded on the output tape, one after the other. When printed, the two lines stood one above the other. Between these lines, the first and the second difference was marked by a sign; the number of the line was also printed on the left margin, next to the second line. This made it possible to detect the differences at a glance. According to some rules, the computer decided which one of the two different

forms of the two lines was to be recorded and printed first, in order to increase the probability that the correct form stood first on the output tape and on the printer output. These rules included, for instance, that all the lines which had been marked in the updating phase by an error code were to be placed second. The opposite was the case when a line contained exactly one character more than the line on the other tape but did not differ from it otherwise; this took into account omission errors which might have occurred when typing very fast.

By means of this printer output, a very fast and very reliable revision was possible. Since all differences were marked by prefixing the second of two different lines by the number of the line, only the left margin had to be skimmed to detect those lines. Within the line, the wrong character was underlined (one of the characters above or below the sign had to be wrong); and when there was no second sign in this line, one was sure that the first difference was the only one. What remained, was to decide which one of both lines was the correct one, or if perhaps both lines were wrong.

The correction was made by preparing on the same equipment used for the transcription a paper tape, which contained the instructions for the correction program. These instructions were extremely short and simple. When the first of two diverging lines was correct, nothing had to be indicated; the computer would automatically reject the second line when it did not receive other instructions. When the line which was printed in the second position (always marked by the number of the line in the left margin) happened to be correct, only the number of this line followed by an equal sign (=) had to be typed, e.g., 8900=. The computer would then

reject the first of the two lines. If, however, the text was correct in one of these lines, but the reference was still incorrect, the correct reference could be typed after typing the number of the line followed by a slash, if the first of the pair of lines was the correct one, or by an equal sign *and* a slash, if the second of the pair was correct, e.g. 9110/Gn 15,20; 992760=/ 4Esr 10,10.

The text of the whole line had to be retyped only when the text of both lines was wrong, or if a new line had to be inserted. An addition sign (+) typed between the number of the line and the text indicated that this text was to be inserted after the line with the corresponding number on the magnetic tape, instead of replacing it. In a similar way, a line could be deleted by typing the line number and a minus sign (–), e.g. 106450.

This method of verification and correction seems to be more effective and more convenient than the verification and correction of punched cards using a card verifier. The number of strokes necessary for verification is the same in both cases, but the correction of punched cards always requires the removal of the wrong card from the verifier, feeding it into the keypunch, correcting it by duplicating the correct part of the card and by retyping the wrong characters (if omissions are to be corrected, this would require perhaps retyping more than half a card and adding a continuation card), and finally reinserting the corrected card into the file. And with this method, the text of wrong cards must not only be typed three times (transcription, retranscription on the verifier, and correction), but, in addition, even "correction" can introduce new errors. Not even the duplication of a card on a keypunch is always reliable. Of course, the

correction of paper tape by duplicating the tape and retyping the wrong characters would be subject to the same errors and would be much more complicated and wearisome than correcting punched cards : the whole tape must be duplicated, not just single records (cards) of the file. This way of correcting paper tape may be suitable for the preparation of short texts for teletype transmission, but nobody would even consider employing this method for the preparation of large data files for the computer.

The effectiveness of our method may be demonstrated by some facts which have been compiled by a short computer program. The text (including the apparatus criticus) which was transcribed contains 116,632 lines with an average of 44.5 strokes per line. This is a total of about 5.185 million strokes for one copy, or 10.37 million strokes for transcription and verification. Only 4,791 lines did not correspond completely in both copies; this means that only one stroke of every 2,164 strokes, or about 0.047%, were wrong and not yet corrected either by overpunching or by adding ignore codes when the tapes were checked for the first time by the computer. This is an extremely low rate which could only be attained by the very convenient correcting possibilities during transcription. These possibilities were used very extensively. In the first transcription (the respective numbers for the retranscription are added in parentheses), 13,211 (13,279) times one or more characters were backspaced and overpunched; 9,993 (10,280) of these corrections were corrections of only one character (a single character was overpunched), 1,880 (1,740) times two characters were overpunched, 649 (619) times three characters were overpunched, in 689 (640) cases, four or more characters were

backspaced and overpunched. The average number of characters overpunched with one correction was 1.55 (1.52).

The correction by ignore codes at the end of a line was used 1331 (1145) times; 48 (49) times two subsequent lines were deleted by adding to the ignore code the number 2, 14 (13) times three lines were deleted.

For the revision and correction, 4,791 out of the 116,632 lines of the text (4.2%) had to be checked in the described way. The following numbers refer only to the second part of the text, containing 54,712 lines(2). In this part, 1,440 lines = 2.6% had to be checked. For the correction of this part, a corrective code had to be typed for only 568 of the 1,440 wrong lines (=39.4%), since, in order to increase the probability that the correct version of a line was printed first, the computer was programmed to make a preselection according to the rules described above. Thus, in 60.6% of the cases, the computer could simply reject the incorrect second copy. In fact, 30 superfluous correction codes were typed (598 instead of 569 or 41.6% instead of 39.4%). From these 598 corrective lines, 500 lines (83.6%) contained nothing but the number of the line (indicating that the computer had to reject the first copy); 16 lines (2.7%) contained, in addition, the reference; 75 lines (12.5%) contained the number of the line and

(2) The division is made after the book of Sapientia. The first part thus contains 61,920 lines from Genesis to Sapientia, the second part 54,712 lines from the book of Sirach to the end of the Old Testament, the New Testament, and the Apocrypha contained in the appendix of the edition. I made this division, because I wanted to check which of the correction instructions were necessary and which were not. To reduce the time necessary for this work, I chose the last third of the corrections, covering nearly half of the text.

the text (in those cases, both copies were wrong, but the reference was right); 7 lines (1.2%) had to be retyped completely, including the text and the reference. This means that the text had to be retyped only for 5.7% of the wrong lines (= fifteen out of every ten thousand lines of the text).

The preparation for the transcription and the double transcription of the Vulgate text took one man about seven months; the revision and corrections were made in about four weeks.

Zentrum für Datenverarbeitung der
Universität

Dr. Wilhelm OTT
74 Tübingen, Köllestrasse 1

Explanation of the signs on the printer output

↑ shift to upper case
↓ shift to lower case
⌐? shift-out character for italics
≠ corresponding shift-in character
≡ non-locking shift character for Greek letters (when the letter is upper case) and Gothic letters (lower case only)
< end of the text part of a line
¬ end of line
$ transcription for
% transcription for

The first number after the reference (e.g. Ex 38,15) is the number of the input tape (0 or 1) from which the text was taken. Then follows the number of the stich (e.g. 65630; the stichs are numbered in steps of 10), and the number of the line within single stichs, starting from 0.

	INDUXITQUE IN CIRCULOS QUI IN ALTARIS LATERIBUS ($ L. A. ΞC.)<		0	65450	0
	EMINEBANT<	⍲↑E↓X ↑38,7↓⍯¬	0	65450	1
	IPSUM AUTEM ALTARE NON ERAT SOLIDUM<	⍲↑E↓X ↑38,7↓⍯¬	0	65460	0
	SED CAVUM EX TABULIS (TABULATIS ⍲↑GCTM⍯ΞF↓) ET INTUS VACUUM<	⍲↑E↓X ↑38,7↓⍯¬	0	65470	0
	FECIT ET LABRUM AENEUM CUM BASE (¬SI ↑ΞSΞL⍲T↓⍯ΞCΞR) SUA<	⍲↑E↓X ↑38,8↓⍯¬	1	65480	0
65480	FECIT ET LABRUM AENEUM CUM BASE (¬I ↑ΞSΞL⍲T↓⍯ΞCΞR) SUA<	⍲↑E↓X ↑38,8↓⍯¬	0	65480	0
	DE SPECULIS MULIERUM QUAE EXCUBABANT (¬CUBANT ⍲↑ATM⍯ΞF↓) IN OSTIO<		0	65490	0
	TABERNACULI<	⍲↑E↓X ↑38,8↓⍯¬	0	65490	1
	(↓ FECIT ΞC) ET ATRIUM IN CUIUS AUSTRALI PLAGA ERANT TENTORIA DE<		0	65500	0
65500	(↓ FECIT ΞC) ET ATRIUM IN CUIUS AUTRALI PLAGA ERANT TENTORIA DE<		1	65500	0
	BYSSO RETORTA CUBITORUM CENTUM<	⍲↑E↓X ↑38,9↓⍯¬	0	65500	1
	COLUMNAE AENEAE VIGINTI CUM BASIBUS SUIS<	⍲↑E↓X ↑38,10↓⍯¬	0	65510	0
	CAPITA COLUMNARUM ET TOTA OPERIS CELATURA ARGENTEA<	⍲↑E↓X ↑38,10↓⍯¬	0	65520	0
	AEQUE AD (A ⍲↑A↓⍯.) SEPTENTRIONALIS (¬I ⍲↑A↓⍯.⍲I⍯ ¬EM ΞC⍲I⍯<		0	65530	0
65530	AEQUE AD (A ⍲↑A↓⍯.) SEPTENTRIONALIS (¬LI ⍲↑A↓⍯.⍲I⍯ ¬LEM ΞC⍲I⍯<		1	65530	0
	SEPTENTRIONIS ⍲↑G↓⍯ΞR) PLAGAM (¬A ⍲↑G↓⍯.)<	↑E↓X ↑38,11↓⍯¬	0	65530	1
	TENTORIA COLUMNAE BASESQUE ET CAPITA COLUMNARUM<	⍲↑E↓X ↑38,11↓⍯¬	0	65540	0
	EIUSDEM ET (% ⍲↑A⍯ΞS⍲TM⍯ΞF↓ΞC) MENSURAE ET OPERIS AC (ET ⍲↑A⍯ΞS↓.)<		0	65550	0
	METALLI ERANT<	⍲↑E↓X ↑38,11↓⍯¬	0	65550	1
	IN EA VERO PLAGA QUAE (↓ AD ΞC) OCCIDENTEM RESPICIT<	⍲↑E↓X ↑38,12↓⍯¬	0	65560	0
	FUERE (¬ERUNT ⍲↑T⍯ΞF↓ΞC) TENTORIA CUBITORUM QUINQUAGINTA<	⍲↑E↓X ↑38,12↓⍯¬	0	65570	0
	COLUMNAE DECEM CUM BASIBUS SUIS AENEAE<	⍲↑E↓X ↑38,12↓⍯¬	0	65580	0
	ET CAPITA COLUMNARUM CELATA (CELATURA ⍲↑GH↓⍯I⍯ CUM CAELATURA ΞR.⍲I⍯<		0	65590	0
	ET TOTA OPERIS CELATURA ⍲↑A↓⍯ΞC) ARGENTEA<	⍲↑E↓X ↑38,12↓⍯¬	0	65590	1
	PORRO CONTRA ORIENTEM QUINQUAGINTA CUBITORUM PARAVIT TENTORIA<	⍲↑E↓X ↑38,13↓⍯¬	0	65600	0
	E QUIBUS QUINDECIM CUBITOS (¬ORUM ⍲↑A↓⍯) COLUMNARUM TRIUM CUM<		0	65610	0
	BASIBUS SUIS UNUM TENEBAT LATUS<	⍲↑E↓X ↑38,14↓⍯¬	0	65610	1
	ET IN PARTE ALTERA<	⍲↑E↓X ↑38,15↓⍯¬	0	65620	0
	QUIA (QUI ⍲↑T⍯ΞF↓⍲I⍯ ↓ AB ⍲↑H↓⍯I⍯ ↓ INTER ΞC) UTRAQUE INTROITUM<		0	65630	0
	TABERNACULI FACIT (FECIT ⍲↑OC↓⍯ΞC⍲I⍯ FACIAT ⍲↑T⍯ΞF↓)<	⍲↑E↓X ↑38,15↓⍯¬	0	65630	1
65630 1	TABERNACULI FACIT (FECIT ⍲↑OC↓⍯ΞC⍲I⍯ ¬CIAT ⍲↑T⍯ΞF↓)↓<	⍲↑E↓X ↑38,15↓⍯¬	1	65630	1
	QUINDECIM AEQUE CUBITORUM ERANT TENTORIA<	⍲↑E↓X ↑38,15↓⍯¬	0	65640	0
	COLUMNAE (¬AEQUE ΞC) TRES ET BASES TOTIDEM<	⍲↑E↓X ↑38,15↓⍯¬	0	65650	0
	CUNCTA ATRII TENTORIA BYSSUS TORTA (RETORTA ⍲↑ATM⍯ΞF↓ΞC) TEXUERAT<	⍲↑E↓X ↑38,16↓⍯¬	0	65660	0
	BASES COLUMNARUM FUERE AENEAE<	⍲↑E↓X ↑38,17↓⍯¬	0	65670	0
	CAPITA AUTEM EARUM CUM (↓ CUNCTIS ΞC) CELATURIS SUIS ARGENTEA<	⍲↑E↓X ↑38,17↓⍯¬	0	65680	0